Régression isotonique itérée

Nicolas Jégou

THESE / Université Rennes 2

Sous le sceau de l'Université Européenne de Bretagne

pour obtenir le diplôme de :

DOCTEUR EN MATHEMATIQUES APPLIQUEES

Spécialité : Statistique

Ecole Doctorale : MATISSE

présentée par

Nicolas JEGOU

Régression isotonique itérée

Soutenue le **23 Novembre 2012** devant le jury composé de :

Christophe Abraham	SupAgro Montpellier	Examinateur
Gérard Biau	Université Pierre et Marie Curie, Paris	Examinateur
Cécile Durot	Université Paris X	Rapporteur
Arnaud Guyader	Université Rennes II	Directeur de thèse
Eric Matzner-Løber	Université Rennes II	Directeur de thèse
Sylvain Sardy	Université de Genève	Rapporteur

REMERCIEMENTS

Je tiens tout d'abord à remercier mes encadrants. Après avoir grandemment contribué à me faire venir enseigner à l'Université, Eric Matzner-Løber m'a fait confiance pour ce travail de recherche ; je tiens à lui témoigner toute ma reconnaissance et mes remerciements à son égard dépassent très largement le cadre professionnel. Arnaud Guyader a accepté sans hésitation de m'aider au moment même où les difficultés à venir s'annonçaient grandes. Je le remercie infiniment pour son écoute et sa très grande disponibilité ; ce manuscrit lui doit beaucoup. Enfin, rendons à César ce qui appartient à César : l'idée de la méthode présentée ici est tout droit sortie de l'esprit prodigieux de Nicolas Hengartner. En m'invitant deux fois à Los Alamos ces dernières années, il m'a donné le privilège de travailler à ses côtés. Je suis désormais au moins sûr d'une chose : le génie et l'enthousiasme sont liés !

Je tiens à remercier chaleureusement Cécile Durot et Sylvain Sardy pour l'intérêt qu'ils ont porté à mon travail en acceptant de rapporter sur cette thèse. Je suis également très reconnaissant à Christophe Abraham et Gérard Biau d'avoir bien voulu prendre sur leur temps précieux pour faire partie du jury.

Je voudrais aussi remercier Marie de Tayrac pour m'avoir fourni les données médicales permettant de développer la partie appliquée de cette étude. Je n'oublie pas Mathieu Emily, Yuna Blum et à nouveau Marie car sans leur aide, j'en serais sans doute encore à me demander ce qu'est un gène !

Un très grand merci à Dominique Dehay, Directeur du Laboratoire de Statistique à l'Université Rennes 2, qui a défendu ma cause auprès des conseils centraux pour que ma charge d'enseignement soit réduite ces dernières années. Je veux aussi saluer l'ensemble de mes collègues des départements MASS et Géographie à l'Université avec qui j'ai le plaisir de travailler au quotidien. J'adresse une pensée particulière à Pierre-André Cornillon pour ses conseils en informatique ainsi que pour avoir facilité mon second déplacement aux USA. Mes remerciements vont également à Laurent Rouvière, Bruno Pelletier, Alain Mom et Jacques Benasseni. Leur point de vue sur mon travail, leurs suggestions bibliographiques ainsi que leurs avis sur les aspects administratifs m'ont été d'un grand secours.

Je tiens également à remercier François Le Gland. En acceptant de m'accueillir à L'INRIA l'été dernier, il m'a permis de découvrir les joies de la vie monastique (enfin !) et de bénéficier des conseils avisés de Frédéric Cérou pour achever dans les meilleures conditions mon manuscrit.

Sur un plan plus personnel, je voudrais saluer quelques amis pour la qualité des moments passés en leur compagnie ces dernières années ; je pense en particulier aux honorables membres du Binchbury Group [1] ainsi qu'à Hang, Louisa, Johan, Mikrolax, Alex et bien sûr à Yuna, Paco et Olivier avec qui j'ai tant de plaisir à faire de la musique.

Pour finir, j'ai une pensée émue pour mes parents, mon frère Pierre-Yves et sa petite famille. Je pense aussi très fort à mon fils Arthur qui a accepté, avec le courage d'un grand garçon, que son père reste si souvent à sa table de travail au lieu de s'occuper de lui. Je remercie enfin Yuna pour tout son soutien et son Amour.

1. Formule empruntée à Guyader (2011).

Résumé

Résumé Ce travail se situe dans le cadre de la régression non paramétrique univariée. Supposant la fonction de régression à variation bornée et partant du résultat selon lequel une telle fonction se décompose en la somme d'une fonction croissante et d'une fonction décroissante, nous proposons de construire et d'étudier un nouvel estimateur combinant les techniques d'estimation des modèles additifs et celles d'estimation sous contraintes de monotonie. Plus précisément, notre méthode consiste à itérer la régression isotonique selon l'algorithme backfitting. On dispose ainsi à chaque itération d'un estimateur de la fonction de régression résultant de la somme d'une partie croissante et d'une partie décroissante.
Le premier chapitre propose un tour d'horizon des références relatives aux outils cités à l'instant.
Le chapitre suivant est dédié à l'étude théorique de la régression isotonique itérée. Dans un premier temps, on montre que, la taille d'échantillon étant fixée, augmenter le nombre d'itérations conduit à l'interpolation des données. On réussit à identifier les limites des termes individuels de la somme en montrant l'égalité de notre algorithme avec celui consistant à itérer la régression isotonique selon un algorithme de type réduction itérée du biais. Nous établissons enfin la consistance de l'estimateur.
Le troisième chapitre est consacré à l'étude pratique de l'estimateur. Comme augmenter le nombre d'itérations conduit au sur-ajustement, il n'est pas souhaitable d'itérer la méthode jusqu'à la convergence. Nous examinons des règles d'arrêt basées sur des adaptations de critères usuellement employés dans le cadre des méthodes linéaires de lissage (AIC, BIC,...) ainsi que des critères supposant une connaissance a priori sur le nombre de modes de la fonction de régression. Il en ressort un comportement intéressant de la méthode lorsque la fonction de régression possède des points de rupture. Nous appliquons ensuite l'algorithme à des données réelles de type puces CGH où la détection de ruptures est d'un intérêt crucial. Enfin, une application à l'estimation des fonctions unimodales et à la détection de mode(s) est proposée.

Mots clés Régression non paramétrique, Régression isotonique, Modèle additif, Algorithme Backfitting, Régression unimodale.

abstract This thesis is part of non parametric univariate regression. Assume that the regression function is of bounded variation then the Jordan's decomposition ensures that it can be written as the sum of an increasing function and a decreasing function. We propose and analyse a novel estimator which combines the isotonic regression related to the estimation of monotone functions and the backfitting algorithm devoted to the estimation of additive models.

The first chapter provides an overview of the references related to isotonic regression and addi-

tive models. The next chapter is devoted to the theoretical study of iterative isotonic regression. As a first step we show that increasing the number of iterations tends to reproduce the data. Moreover, we manage to identify the individual limits by making a connexion with the general property of isotonicity of projection onto convex cones and deriving another equivalent algorithm based on iterative bias reduction. Finally, we establish the consistency of the estimator.

The third chapter is devoted to the practical study of the estimator. As increasing the number of iterations leads to overfitting, it is not desirable to iterate the procedure until convergence. We examine stopping criteria based on adaptations of criteria usually used in the context of linear smoothing methods (AIC, BIC, ...) as well as criteria assuming the knowledge of the number of modes of the regression function. As it is observed an interesting behavior of the method when the regression function has breakpoints, we apply the algorithm to CGH-array data where breakopoints detections are of crucial interest. Finally, an application to the estimation of unimodal functions is proposed.

Keywords Nonparametric Regression, Isotonic regression, Additive model, Backfitting algorithm, Unimodal regression.

Table des matières

Introduction

Le cadre de ce travail est celui de la régression non paramétrique univariée. Nous considérons un couple (X, Y) de variables aléatoires réelles liées par le modèle général

$$Y = r(X) + \varepsilon \tag{1}$$

où ε est une variable aléatoire réelle telle que $\mathbb{E}[\varepsilon|X] = 0$. On sait que

$$r(X) = \mathbb{E}[Y|X] = \underset{f:\mathbb{R}\to\mathbb{R}}{\operatorname{argmin}}\, \mathbb{E}[(Y - f(X))^2].$$

Disposant de réplications $(X_i, Y_i)_{i=1,\dots,n}$ i.i.d. de (X, Y), on cherche à estimer r en construisant un estimateur $\hat{r}$ qui minimise

$$\mathbb{E}\left[(\hat{r}(X) - r(X))^2\right].$$

Les deux approches principales pour estimer r sont l'approche paramétrique et l'approche non paramétrique. Dans le premier cas, on suppose que r appartient à une classe $\mathcal{F}$ de fonctions indexables par un nombre fini de paramètres, l'ensemble des fonctions linéaires par exemple. La fonction r est donc supposée appartenir à une famille connue et l'estimer consiste alors à déterminer l'ensemble des paramètres la caractérisant. Dans un cadre non paramétrique, la classe $\mathcal{F}$ n'est pas supposée indexable par un nombre fini de paramètres. Les hypothèses faites sur la fonction r sont en ce sens beaucoup moins restrictives que dans le cas précédent puisqu'elles proviennent en général de simples conditions de régularité : $\mathcal{F}$ est par exemple l'ensemble des fonctions continues ou celui des fonctions k fois différentiables.

Dans ce travail, nous nous plaçons dans un cadre non paramétrique et prenons pour $\mathcal{F}$ l'ensemble des fonctions à variation bornée. Avant de justifier ce choix, donnons quelques éléments généraux sur cette classe de fonctions. Le lecteur souhaitant plus de détails pourra se référer à Rudin (1975).

La notion de fonction à variation bornée a été introduite par Jordan en 1881 pour étendre le théorème de Dirichlet sur la convergence des séries de Fourier. La définition est la suivante :

Définition 0.1 (Fonction à variation bornée)
Une fonction f définie sur un intervalle compact I est dite à variation bornée si sa variation totale

$$V(f) = \sup_{t_1<\dots<t_n\in I} \sum_i |f(t_{i+1}) - f(t_i)|$$

est finie.

Pour simplifier, nous prenons $I = [0, 1]$ et considérons donc $r \in \mathcal{F}$, avec $\mathcal{F}$ définie par

$$\mathcal{F} = \left\{ f : I = [0,1] \mapsto \mathbb{R} \qquad \sup_{0 \leq t_1 < \ldots < t_n \leq 1} \sum_i |f(t_{i+1}) - f(t_i)| < \infty \right\}. \tag{2}$$

$\mathcal{F}$ est un ensemble vaste puisqu'il contient par exemple la classe des fonctions lipschitziennes, celle des fonctions monotones, celle des fonctions dérivables à dérivée bornée ou encore celle des fonctions ayant un nombre fini d'extremums locaux. Un contre-exemple typique est la fonction f définie par $f(0) = 0$ et $f(x) = \sin(1/x)$ pour $x \in]0, 1]$ qui oscille une infinité de fois entre -1 et 1 quand x se rapproche de 0.

Ce n'est pas tant la définition 0.1 qui motive notre idée qu'une caractérisation de $\mathcal{F}$ assurant qu'une fonction est à variation bornée si et seulement si elle est la somme d'une fonction croissante et d'une fonction décroissante. En notant $\mathcal{C}^+$ (resp. $\mathcal{C}^-$) le cône des fonctions croissantes (resp. décroissantes) sur $[0, 1]$, nous avons donc

$$f \in \mathcal{F} \Leftrightarrow \exists u \in \mathcal{C}^+, \exists b \in \mathcal{C}^- : f = u + b. \tag{3}$$

On peut voir cette écriture comme un modèle additif mettant en jeu la partie croissante et la partie décroissante de la fonction de régression et c'est ce point de vue qui est à l'origine du travail présenté dans ce document.

D'un côté, la régression sous contrainte de monotonie, communément appelée régression isotonique, intéresse les statisticiens depuis le milieu des années cinquante et les écrits de Ayer *et al.* (1955). Ses développements se poursuivent encore actuellement avec les travaux de Durot (2007) ou Tibshirani *et al.* (2011) par exemple. De l'autre, les modèles additifs ont été introduits par Friedman & Stuetzle (1981) au début des années quatre-vingt en statistique multivariée comme un moyen de s'accommoder du fléau de la dimension. Les travaux de Buja *et al.* (1989), relayés par ceux de Hastie & Tibshirani (1990), concernent l'algorithme backfitting destiné à les estimer, et ont été à l'origine de nombreux prolongements théoriques. Ils ont également posé les fondations de leur mise en œuvre pratique, contribuant grandement à en faire une méthode très utilisée dans la communauté statistique.

Nous proposons ici de nous appuyer sur la caractérisation (3) pour faire le lien entre ces deux champs de la statistique. La méthode que nous présentons consiste ainsi à itérer la régression isotonique selon l'algorithme backfitting pour estimer une fonction de régression à variation bornée. Dans cette introduction, nous décrivons brièvement l'essentiel des idées et résultats qui sont développés dans la suite du manuscrit.

Nous supposons que la fonction de régression r du modèle (1) est définie sur $I = [0, 1]$ et à variation bornée. Nous savons donc qu'il existe u croissante et b décroissante telles que $r = u + b$. Une telle décomposition n'est en général pas unique. Par exemple, soit $r = u + b$ une décomposition quelconque, alors pour toute fonction w croissante, une autre décomposition est

$$r = u + b = (u + w) + (b - w).$$

L'unicité suppose de fixer des contraintes sur la décomposition. Il est facile de vérifier par exemple que si l'on considère des fonctions dérivables, la condition

$$u' \times b' \equiv 0$$

aboutit (à une constante près) à une décomposition unique. Une contrainte plus générale fait intervenir la mesure de Stieltjes (cf. Revuz & Yor (2005), chapitre 0). Elle assure que si r,

en plus d'être à variation bornée, est continue à droite en tout point, alors il existe (toujours à une constante près) une unique décomposition dont les termes ont des mesures de Stieltjes mutuellement singulières (ou étrangères). Sous une forme générale, ce résultat s'énonce de la manière suivante :

Théorème 0.1

Soit f une fonction définie sur $[0,1]$, à variation bornée et continue à droite en tout point de $[0,1]$ avec $f(0_-)=0$. Il existe une unique décomposition $f=F^+-F^-$ avec F^+ et F^- croissantes sur $[0,1]$ pour lesquelles les mesures de Stieltjes sont mutuellement singulières.

La preuve, qui n'est pas détaillée dans la référence mentionnée à l'instant, figure en annexe A page 105. De façon pratique, le résultat signifie qu'à une constante additive près, il y a une seule décomposition $r=u+b$ telle que localement, les fonctions u et b ne varient pas simultanément : lorsque u est croissante, b est constante et inversement. En ce sens cette décomposition peut être vue comme la décomposition à variation minimale de r puisque localement, la variation de r n'est portée que par l'un des deux termes mais jamais par les deux. Nous la notons

$$r = u^\star + b^\star. \tag{4}$$

La situation est représentée sur un exemple, à gauche en figure 1. Le fait que l'unicité n'ait lieu qu'à une constante additive près signifie que les courbes de $u^\star$ et de $b^\star$ peuvent subir des translations opposées.

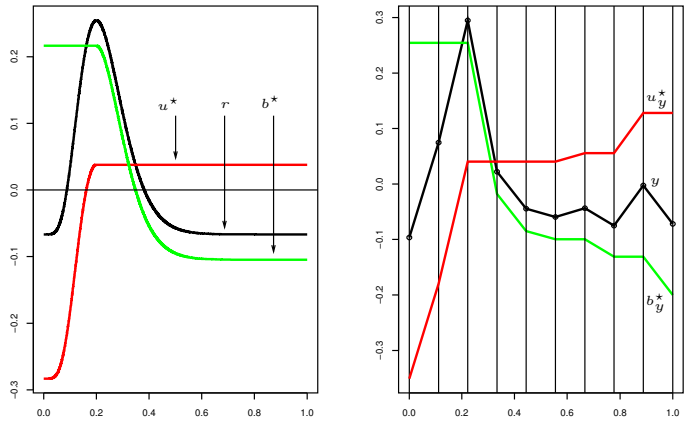


Fig. 1 – Décomposition à variation minimale d'une fonction (à gauche) et d'un vecteur (à droite).

Pour construire notre estimateur, nous considérons n réplications (X_i,Y_i) i.i.d. du modèle (1). L'algorithme PAV ou PAVA (pour Pool Adjacent Violator Algorithm), connu depuis Ayer *et al.* (1955), est l'algorithme le plus couramment utilisé pour la régression isotonique. Il suppose d'utiliser les observations réordonnées selon l'ordre des X_i. Nous les notons $(X_{(i)},Y_{(i)})$. Nous considérons la fonction de répartition de X continue et avons donc $X_{(1)}<\ldots<X_{(n)}$. Pour $i=1,\ldots,n$, nous notons x_i (resp. y_i) les observations de $X_{(i)}$ (resp. $Y_{(i)}$). Nous disposons ainsi des observations (x_i,y_i) avec $x_1<\ldots<x_n$ et $y_1,\ldots,y_n$ les observations correspondantes.

L'algorithme PAVA permet de calculer le meilleur ajustement croissant (ou décroissant) au sens quadratique de $y=(y_1,\ldots,y_n)'$. Nous notons

$$\mathrm{iso}(y)=(\hat{y}_1,\ldots,\hat{y}_n)'=\operatorname*{argmin}_{u\in\mathcal{C}_n^+}\frac{1}{n}\sum_{i=1}^{n}(y_i-u_i)^2=\operatorname*{argmin}_{u\in\mathcal{C}_n^+}\|y-u\|_n^2 \tag{5}$$

avec $\mathcal{C}_n^+ = \{u \in \mathbb{R}^n : u_1 \leq \ldots \leq u_n\}$ et $\|.\|_n$ la norme quadratique empirique.
De façon analogue, l'ajustement décroissant ou antitonique est noté

$$\text{anti}(y) = (\hat{y}_1, \ldots, \hat{y}_n)' = \underset{b \in \mathcal{C}_n^-}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - b_i)^2 = \underset{b \in \mathcal{C}_n^-}{\text{argmin}} \|y - b\|_n^2 \tag{6}$$

avec $\mathcal{C}_n^- = \{b \in \mathbb{R}^n : b_1 \geq \ldots \geq b_n\}$.
$\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ sont des cônes fermés dans $\mathbb{R}^n$ et les estimateurs (5) et (6) correspondent aux projections sur ces cônes pour le produit scalaire associé à la norme empirique. Notons que ces opérateurs de projection n'étant pas linéaires, les estimateurs que nous construisons ne le sont par conséquent pas non plus.

Le principe du backfitting pour estimer les composantes d'une somme consiste, en partant d'estimateurs initiaux, à actualiser tour à tour chacun des termes de la somme en ajustant les résidus partiels. En partant d'ajustements initiaux fixés à $\hat{u}^{(0)} = \hat{b}^{(0)} = \hat{r}^{(0)} = 0$, et en commençant chaque cycle par un lissage isotonique, cela donne pour une étape $k \geq 1$:

$$\hat{u}^{(k)} = \text{iso}\left(y - \hat{b}^{(k-1)}\right) \qquad \hat{b}^{(k)} = \text{anti}\left(y - \hat{u}^{(k)}\right) \qquad \hat{r}^{(k)} = \hat{u}^{(k)} + \hat{b}^{(k)}. \tag{7}$$

D'ordinaire, on arrête l'algorithme backfitting lorsque les estimations individuelles des termes de la somme ne varient plus sensiblement d'une étape à l'autre. Nous montrons que dans notre cas, augmenter le nombre d'itérations conduit à interpoler les données, ce qui n'est pas souhaitable et rend donc ce critère d'arrêt inopérant. Nous avons dès lors proposé plusieurs critères d'arrêt dans la mise en œuvre pratique de la méthode. Cependant, l'obtention de ce résultat d'interpolation et plus généralement l'analyse de l'estimateur à n fixé, permet de mettre en lumière un certain nombre de propriétés intéressantes.

Il existe une façon directe de montrer que $\lim_{k\to\infty} \|y - \hat{r}^{(k)}\|_n = 0$ en voyant la méthode comme une version particulière de l'algorithme de Von Neumann (cf. Von Neumann (1950), théorème 13.7 page 55) puis en utilisant les résultats de Bauschke & Borwein (1994). A notre connaissance, ceux-ci ne permettent pas de préciser la convergence des termes individuels $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$. En revanche, par notre approche, nous établissons qu'en augmentant le nombre d'itérations, $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$ convergent respectivement vers les termes de ce que nous appelons la décomposition à variation minimale $y = u_y^\star + b_y^\star$ du vecteur y. Comme pour une fonction, $u_y^\star$ capture les variations croissantes de y pendant que $b_y^\star$ reste constant et vice-versa. On peut se représenter la situation en s'appuyant sur la courbe reliant les points (x_i, y_i) par des segments comme en figure 1 page 3 à droite.

Certaines étapes de la démonstration s'appuient sur le fait remarquable que le backfitting coïncide dans notre cas avec l'algorithme de réduction itérée du biais consistant, à partir d'un ajustement initial, à faire un lissage croissant des résidus courants suivi d'un lissage des résidus partiels puis à actualiser parties croissante et décroissante. Nous montrons que l'égalité des deux algorithmes est due au fait que $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ sont des cônes "minces" de $\mathbb{R}^n$.

Nous montrons ensuite la consistance de l'estimateur, mais sans pouvoir nous appuyer réellement sur les résultats connus pour la régression isotonique. En effet, ceux-ci étudient le comportement asymptotique de la régression isotonique lorsque la fonction de régression est elle-même croissante, or notre algorithme commence par une régression isotonique sur des points distribués autour de la courbe d'une fonction qui ne l'est pas en général. Notre travail s'articule de la manière suivante. Dans un premier temps, nous montrons ainsi que $\hat{u}^{(1)}$ converge vers $u^{(1)}$, la fonction isotonique la plus proche de r définie par :

$$u^{(1)} = \underset{u \in \mathcal{C}^+}{\text{argmin}}\, \mathbb{E}\left[(r(X) - u(X))^2\right] = \underset{u \in \mathcal{C}^+}{\text{argmin}} \|r - u\|^2.$$

Plus précisément, moyennant l'hypothèse que le bruit ε est borné, et grâce à des outils classiques comme les inégalités de concentration et les "covering numbers", nous montrons

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}^{(1)} - u^{(1)}\|^2\right] = 0.$$

Ceci est illustré dans la partie supérieure de la figure 2 où $\hat{u}^{(1)}$ et $u^{(1)}$ sont respectivement représentées en rouge et bleu.

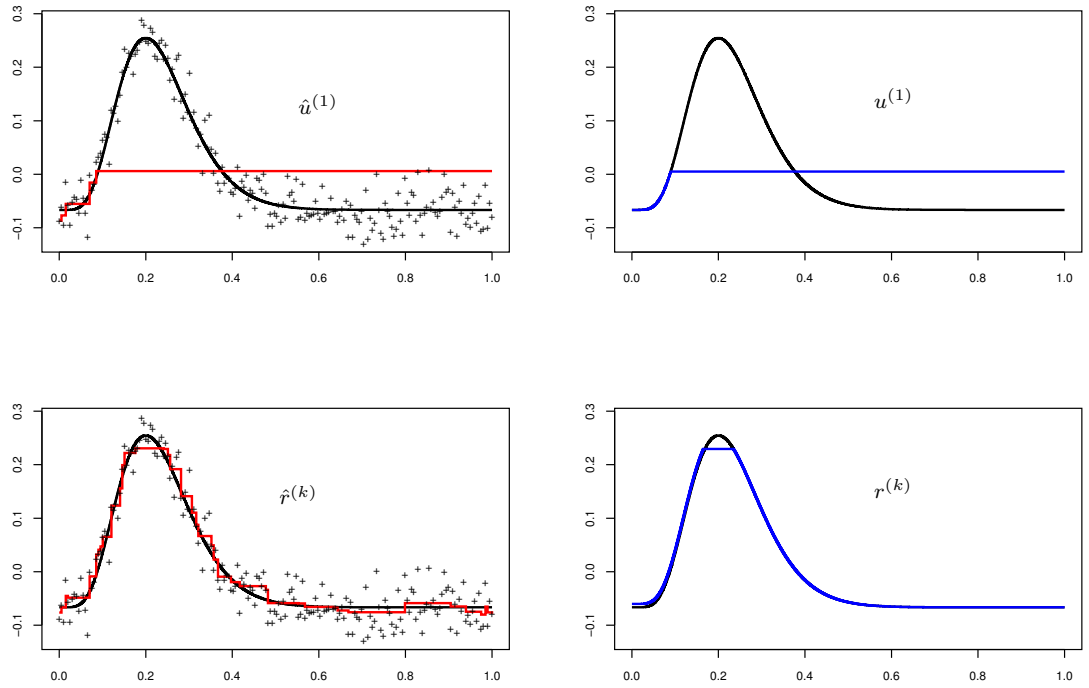


Fig. 2 – Illustration de la consistance.

Ensuite, le principe consiste à itérer ce résultat pour montrer que, pour toute itération k, $\hat{r}^{(k)}$ converge vers la fonction $r^{(k)}$ résultant de l'application de k itérations sur la vraie fonction r. On a ainsi pour tout $k \geq 1$,

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{r}^{(k)} - r^{(k)}\|^2\right] = 0,$$

résultat illustré en bas de la figure 2. On observe également que k augmentant, l'erreur d'approximation, c'est-à-dire l'écart $\|r - r^{(k)}\|$ entre $r^{(k)}$ et la fonction de régression diminue. La preuve proposée s'appuie, comme dans le cas où n est fixé, sur les travaux de Bauschke & Borwein (1994). Finalement, par un jeu de compromis entre le nombre d'itérations k et le nombre de points n, on justifie l'existence d'une suite (k_n) croissante avec n telle que

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{r}^{(k_n)} - r\|^2\right] = 0.$$

La dernière partie de ce document est consacrée aux applications. Nous commençons par étudier le comportement de notre méthode sur des exemples simulés. Plusieurs critères d'arrêt sont envisagés : des critères pénalisés classiques (de type AIC, BIC, etc.) qui sont d'ordinaire utilisés

dans le cadre de lisseurs linéaires et que nous transposons à notre cas ; des critères fondés sur des contraintes de formes comme le nombre de maximums locaux souhaités. En analysant la décomposition biais-variance des résultats, nous observons le bon comportement de la régression isotonique itérée pour des fonctions présentant des discontinuités. Aussi l'appliquons-nous à des données réelles où la détection de ruptures est un objectif. Les données en question proviennent de puces dites d'Hybridation Génomique Comparative ou puces CGH. Elles fournissent des niveaux d'expression de gènes et sont utilisées pour la détection des anomalies du nombre de copies de l'ADN. Enfin, nous proposons d'appliquer la méthode à la régression unimodale qui consiste à estimer le mode d'une fonction de régression lorsque l'on sait par avance que celle-ci est croissante puis décroissante. La régression unimodale est un des prolongements naturels de la régression isotonique et a fait l'objet de plusieurs études (cf. Turner & Wollan (1997)). Nous montrons que notre méthode fournit en quelques étapes un intervalle qui contient le mode. Il suffit d'appliquer ensuite la régression unimodale habituelle au sein de cet intervalle, l'ensemble de l'opération étant avantageux du point de vue calculatoire.

Chapitre 1

Les outils : régression isotonique, modèle additif

1.1 Régression isotonique

Il semble que le terme de "monotone regression" apparaisse pour la première fois dans Lombard & Brunk (1963). Cependant, les premiers résultats d'importance sur la régression isotonique sont publiés au milieu des années cinquante avec les travaux de Ayer *et al.* (1955). Dans cet article fondateur, les auteurs donnent un algorithme permettant de déterminer l'estimateur du maximum de vraisemblance pour le paramètre de probabilité d'une loi binomiale supposé décroître avec la variable explicative. Cet algorithme, qui sera ensuite popularisé sous le nom de PAVA pour "Pool Adjacent Violators Algorithm", servira de base à de nombreux travaux destinés à généraliser le résultat initial. Nous le présentons en section 1.1.2. Dans cet article, les auteurs donnent également une formulation explicite de l'estimateur qui sera elle aussi abondamment reprise pour prolonger son étude théorique : nous présentons les "min-max formulas" en section 1.1.3. Auparavant, nous donnons quelques définitions et généralités dans le cadre plus spécifique de notre étude.

1.1.1 Généralités

Dans quels cas pratiques est-il intéressant d'avoir recours à la régression isotonique ? Selon les phénomènes étudiés, il arrive que l'on ait une connaissance a priori sur la monotonie de la fonction r. Si l'on s'intéresse par exemple à l'influence que peut avoir la dose d'un engrais (variable X) sur le rendement d'une variété de blé (variable Y), il est naturel de supposer r croissante. A l'inverse, on peut penser que le risque de développer une maladie décroît avec la distance à la source de pollution (cf. Diggle *et al.* (1999)). Il semble alors pertinent de supposer la monotonie de la fonction de régression dans l'écriture du modèle et donc de contraindre les estimateurs à respecter cette monotonie. Une idée peut bien sûr consister à forcer la monotonie d'un estimateur paramétrique en imposant un signe particulier à une droite de régression par exemple. Cependant, on ne s'affranchit pas dans ce cas des inconvénients de la régression paramétrique liés à une spécification trop rigide du modèle. D'un autre point de vue, les connaissances actuelles en régression non paramétrique permettent aussi de rendre la plupart des lisseurs usuels monotones. C'est ce que proposent de faire Mammen & Thomas-Agnan (1999) pour des splines de lissage ou Hall & Huang (2001) pour des estimateurs à noyaux par exemple.

Cependant, la régression isotonique au sens où on l'entend usuellement désigne un ensemble de méthodes non paramétriques répondant spécifiquement au problème de l'ajustement d'une fonction monotone aux données, et qui ne s'appuient pas sur des estimateurs conçus pour un contexte habituel de régression. Considérons des observations $(x_i, y_i)_{i=1\ldots n}$, les x_i appartenant à un ensemble $\mathcal{X}$, étant classés selon un ordre partiel $\preceq$ et les y_i appartenant à $\mathbb{R}$ par exemple. Dans ce cadre, une fonction $f : \mathcal{X} \to \mathbb{R}$ est dite isotonique si elle préserve cet ordre c'est-à-dire si pour tous $x, x' \in \mathcal{X}$,

$$x \preceq x' \Rightarrow f(x) \leq f(x').$$

A l'inverse, elle est dite antitonique lorsque

$$x \preceq x' \Rightarrow f(x) \geq f(x').$$

La régression isotonique (resp. antitonique) consiste à trouver une fonction isotonique (resp. antitonique) qui ajuste au mieux les observations. Selon la fonction de coût choisie, on est ainsi amené à minimiser une quantité de la forme

$$L_p(f) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \omega_i |y_i - f(x_i)|^p & 1 \leq p < \infty \\ \max_{1\leq i\leq n} \omega_i |y_i - f(x_i)| & p = \infty \end{cases} \tag{1.1}$$

sur l'ensemble des fonctions isotoniques sur $\mathcal{X}$, les ω_i étant des poids positifs donnés aux observations.

Dans ce chapitre, nous proposons de faire un tour d'horizon des travaux ayant trait à la régression isotonique. Commençons par dire, et nous venons de faire l'abus de langage dans la phrase précédente, que par régression isotonique, nous entendons régression sous contrainte de monotonie c'est-à-dire régression isotonique ou régression antitonique, le contexte permettant de lever l'ambiguïté. Les estimateurs et leurs propriétés diffèrent évidemment selon les problématiques, les ensembles sur lesquels on travaille et les fonctions de coût envisagés. Nous choisissons de centrer cette présentation sur le cadre qui nous intéresse et qui a été énoncé en introduction. Ainsi, rappelons que nous considérons le modèle de régression

$$Y = r(X) + \varepsilon. \tag{1.2}$$

Nous supposons que X prend ses valeurs sur $I = [0, 1]$; cet ensemble est bien sûr muni de l'ordre naturel strict (ou total). Disposant d'un échantillon i.i.d. $(X_i, Y_i)_{i=1\ldots n}$ de (X, Y) avec la fonction de répartition de X continue, on réordonne les données selon l'ordre des X_i ce que l'on note $(X_{(i)}, Y_{(i)})_{i=1\ldots n}$. On note x_i (resp. y_i) les observations de $X_{(i)}$ (resp. $Y_{(i)}$) et l'on a ainsi

$$0 < x_1 < x_2 < \ldots < x_n < 1. \tag{1.3}$$

Une fonction isotonique sur I est simplement une fonction non décroissante sur cet intervalle :

Définition 1.1 (Fonction isotonique)
Une fonction f est dite isotonique sur $I = [0, 1]$ si

$$\forall x, x' \in I : \qquad x \leq x' \Rightarrow f(x) \leq f(x').$$

On note $\mathcal{C}^+$ l'ensemble des fonctions isotoniques sur I.

Une fonction est antitonique sur I si elle est non croissante sur I :

Définition 1.2 (Fonction antitonique)
Une fonction f est dite antitonique sur $I = [0, 1]$ si

$$\forall x, x' \in I : \qquad x \leq x' \Rightarrow f(x) \geq f(x').$$

On note $\mathcal{C}^-$ l'ensemble des fonctions antitoniques sur I.

Par ailleurs, l'essentiel des résultats que nous présentons concerne le cas le plus fréquemment étudié dans la littérature à savoir celui où les observations ont toutes un poids identique et où la fonction de coût est quadratique ($\omega_i = 1$ et $p = 2$ dans l'équation (1.1)). Nous considérons ainsi sur $\mathbb{R}^n$ la norme quadratique empirique $\|.\|_n$ et $\langle ., . \rangle_n$ le produit scalaire associé :

$$\langle y, y' \rangle_n = \frac{1}{n} \sum_{i=1}^{n} y_i y'_i \qquad \|y\|_n = \left(\frac{1}{n} \sum_{i=1}^{n} y_i^2 \right)^{1/2}.$$

La régression isotonique est alors définie comme le meilleur ajustement des données au sens quadratique :

Définition 1.3 (Régression isotonique)
La régression isotonique de $y = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ est la solution au problème de minimisation

$$\operatorname*{argmin}_{u \in \mathcal{C}_n^+} \frac{1}{n} \sum_{i=1}^{n} (y_i - u_i)^2 = \operatorname*{argmin}_{u \in \mathcal{C}_n^+} \|y - u\|_n^2$$

où $\mathcal{C}_n^+ = \{u \in \mathbb{R}^n : u_1 \leq \ldots \leq u_n\}$. On note $\mathrm{iso}(y)$ le vecteur des valeurs ajustées.

Définition 1.4 (Régression antitonique)
La régression antitonique de $y = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ est la solution au problème de minimisation

$$\operatorname*{argmin}_{b \in \mathcal{C}_n^-} \frac{1}{n} \sum_{i=1}^{n} (y_i - b_i)^2 = \operatorname*{argmin}_{b \in \mathcal{C}_n^-} \|y - b\|_n^2$$

où $\mathcal{C}_n^- = \{b \in \mathbb{R}^n : b_1 \geq \ldots \geq b_n\}$. On note $\mathrm{anti}(y)$ le vecteur des valeurs ajustées.

Ces définitions sous-entendent que les problèmes de minimisation considérés ont une solution et que celle-ci est unique. L'existence et l'unicité se justifient facilement en remarquant que $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ sont des cônes convexes fermés de $\mathbb{R}^n$ et que $\mathrm{iso}(y)$ et $\mathrm{anti}(y)$ correspondent aux projetés de y sur ces cônes pour la norme $\|.\|_n$ introduite ci-dessus. Cet argument de projection sur des cônes permet de donner une caractérisation géométrique de $\mathrm{iso}(y)$ et $\mathrm{anti}(y)$:

Proposition 1.1 (Caractérisation de la régression isotonique)
Soit $y \in \mathbb{R}^n$. $\hat{y} \in \mathcal{C}_n^+$ (resp. $\mathcal{C}_n^-$) est la régression isotonique (resp. antitonique) de y si et seulement si

$$\langle y - \hat{y}, \hat{y} \rangle_n = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) \hat{y}_i = 0 \tag{1.4}$$

et

$$\forall u \in \mathcal{C}_n^+ (\textit{resp. } \mathcal{C}_n^-) : \langle y - \hat{y}, u \rangle_n = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) u_i \leq 0. \tag{1.5}$$

On illustre graphiquement cette Proposition en figure 1.1. L'équation (1.4) traduit l'orthogonalité entre $y - \hat{y}$ et $\hat{y}$ et l'équation (1.5) le fait que $y - \hat{y}$ forme un angle obtus avec tout vecteur de $\mathcal{C}_n^+$.

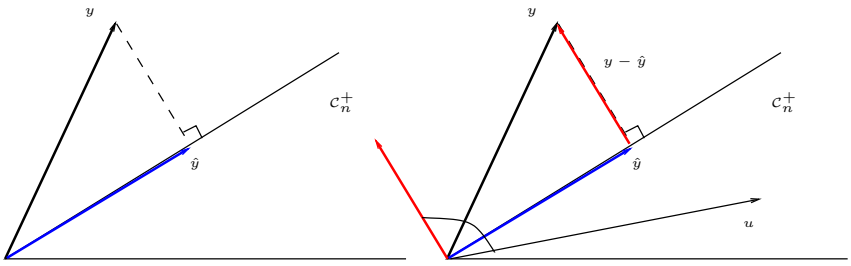


Fig. 1.1 – Caractérisation de la régression isotonique.

Une conséquence immédiate de la Proposition 1.1 est aussi que les valeurs ajustées ont même moyenne que les y_i. En effet, il est clair que toute suite constante de valeurs est à la fois isotonique et antitonique. Ainsi par exemple, $\mathbb{1}_n = (1, \ldots, 1)' \in \mathcal{C}_n^+ \cap \mathcal{C}_n^-$ et $-\mathbb{1}_n \in \mathcal{C}_n^+ \cap \mathcal{C}_n^-$. L'équation (1.5) appliquée avec $u = \mathbb{1}_n$ puis $u = -\mathbb{1}_n$ donne alors :

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{iso}(y)_i = \frac{1}{n}\sum_{i=1}^{n} \mathrm{anti}(y)_i = \overline{y}. \tag{1.6}$$

Il est important de noter que les projections que nous évoquons ne possèdent pas toutes les propriétés habituelles des projections sur les sous-espaces vectoriel. Outre le fait que dans un tel cas, l'appartenance au sous-espace et l'orthogonalité avec le résidu (l'équivalent de l'équation (1.4)) suffisent à caractériser le projeté, il faut préciser que les projections ici ne sont pas linéaires. Ainsi, sauf cas particuliers

$$\mathrm{iso}(y + y') \neq \mathrm{iso}(y) + \mathrm{iso}(y') \qquad \mathrm{anti}(y + y') \neq \mathrm{anti}(y) + \mathrm{anti}(y'). \tag{1.7}$$

Comme propriété de la projection isotonique et des projections orthogonales classiques, nous avons cependant celle de réduction des distances. Cette proposition ainsi que la Proposition 1.1 sont démontrées dans le cadre plus général des espaces de Hilbert par Zarantonello (1971) par exemple.

Proposition 1.2 (Réduction des distances)
Les applications iso(.) *et* anti(.) *sont 1-lipschitziennes :*

$$\forall y, y' \in \mathbb{R}^n : \qquad \|\mathrm{iso}(y) - \mathrm{iso}(y')\|_n \leq \|y - y'\|_n \qquad \|\mathrm{anti}(y) - \mathrm{anti}(y')\|_n \leq \|y - y'\|_n. \tag{1.8}$$

Les quelques généralités que nous venons de présenter découlent directement de la définition du cadre de notre étude et de l'aspect géométrique naturel associé. Nous en venons maintenant à l'algorithme PAVA qui permet le calcul pratique de l'estimateur.

1.1.2 L'algorithme PAVA

Dans la littérature, la paternité de l'algorithme PAVA est le plus souvent attribuée à Ayer *et al.* (1955). A la même époque cependant, Bartholomew (1959) construit un test d'hypothèse d'égalité de moyennes contre une hypothèse alternative de moyennes décroissantes et décrit sensiblement

la même procédure. Il n'y est cependant pas fait mention de l'article de Ayer *et al.* (1955), pas plus d'ailleurs que dans les travaux de Van Eeden (1957) qui concernent un cas plus général d'ordre et où l'algorithme est lui aussi implicite.

Ayer *et al.* (1955) s'appuient sur des questions de titrages biologiques comme celles soulevées dans Goulden (1952) et considèrent le problème consistant à estimer une séquence monotone de probabilités $p_i, i = 1, \ldots, n$, chaque probabilité p_i mesurant la chance de succès d'une variable réponse au niveau de stimulation i. Il s'agit ainsi d'estimer une suite de paramètres pour une loi binomiale, ces paramètres étant supposés varier de façon monotone avec le niveau i de stimulation. Pour cela, ils disposent, pour chaque niveau i, de N_i essais indépendants dont a_i ont abouti à un succès et $b_i = N_i - a_i$ à un échec. Ils considèrent donc une séquence de valeurs $p_i^* = a_i/N_i$ et cherchent à en déduire la séquence monotone $\hat{p}_i$ du maximum de vraisemblance. Bien que dans leur article le terme ne soit pas employé, cet algorithme peut être vu comme la version initiale du PAVA.

Pour une suite isotonique, il se décline de la manière suivante :
- Si $p_1^* \leq p_2^* \leq \ldots \leq p_n^*$ alors, pour $i = 1, \ldots n$, prendre
$$\hat{p}_i = p_i^*.$$
- Si pour un indice $k \in \{1, 2, \ldots, n-1\}$, on a $p_k^* > p_{k+1}^*$, remplacer dans la séquence des p_i^*, ces deux valeurs par le seul ratio $(a_k + a_{k+1})/(N_k + N_{k+1})$ formant ainsi une suite de $n-1$ valeurs.
- Itérer, à partir de la suite obtenue, les deux points précédents jusqu'à obtenir une suite isotonique.
- A l'issue de cette procédure, pour chaque $i \in \{1, \ldots, n\}$, $\hat{p}_i$ est égal au ratio final dont il a participé au calcul.

La figure 1.2 illustre la démarche sur un exemple.

i	1	2	3	4	5	6
p_i^*	1/3	0/3	1/3	1/3	0/3	2/3
	1/6		1/3	1/3	0/3	2/3
	1/6		1/3	1/6		2/3
	1/6		2/9			2/3
$\hat{p}_i$	1/6	1/6	2/9	2/9	2/9	2/3

Fig. 1.2 – Illustration de l'algorithme PAVA selon Ayer *et al.* (1955).

Pour comprendre comment l'algorithme s'adapte aux données $y_1, \ldots, y_n$, nous détaillons deux étapes successives de l'algorithme. Dans notre cas, les ratios a_i/N_i sont remplacés par les valeurs y_i que l'on peut considérer comme associées aux ratios $y_i/1$. Lorsque l'on rencontre deux valeurs adjacentes y_k et y_{k+1} qui violent la monotonie ($y_k > y_{k+1}$), on les regroupe en les remplaçant par la seule valeur
$$\tilde{y}_k = \frac{y_k + y_{k+1}}{1+1} = \frac{y_k + y_{k+1}}{2}$$

dans une suite ayant un élément de moins. Si cette valeur n'est pas inférieure ou égale à la suivante y_{k+2}, on les regroupe en les remplaçant par

$$\frac{(y_k + y_{k+1}) + y_{k+2}}{2+1} = \frac{y_k + y_{k+1} + y_{k+2}}{3} = \frac{2\tilde{y}_k + y_{k+2}}{3}.$$

A l'issue de ces deux étapes, les 3 valeurs y_k, y_{k+1} et y_{k+2} sont remplacées par une seule et même valeur correspondant à la moyenne des trois. D'un point de vue pratique, la seconde partie de l'égalité précédente montre que la valeur $\tilde{y}_k$ prise en compte est affectée d'un poids correspondant au nombre de valeurs ayant contribué à son calcul. Ce point de vue facilite le déroulement de l'algorithme ainsi que sa présentation formelle comme nous le voyons ensuite. On illustre auparavant la démarche sur un exemple en figure 1.3. On porte au fur et à mesure les poids associés aux valeurs calculées entre parenthèses.

i	1	2	3	4	5	6
y_i	1	0	1	1	0	2
	0.5(2)		1	1	0	2
	0.5(2)		1		0.5(2)	2
	0.5(2)			2/3(3)		2
$\hat{y}_i$	0.5	0.5	2/3	2/3	2/3	2

Fig. 1.3 – Illustration de l'algorithme PAVA.

C'est souvent cette présentation que l'on retrouve dans la littérature. Nous décrivons ci-dessous l'algorithme en nous inspirant de la version donnée dans Barlow *et al.* (1972). La situation présentée correspond à la résolution du problème plus général de minimisation

$$\underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \omega_i \left(y_i - u_i\right)^2$$

où les observations y_i ont des poids positifs ω_i (dans notre cas, $\omega_i = 1$ pour tout i) :

Algorithme 1 Algorithme PAVA

(1) Si $y_1 \leq \ldots \leq y_n$, alors cette séquence initiale constitue aussi la séquence finale et pour $i = 1, \ldots, n$,

$$\hat{y}_i = y_i.$$

(2) Sinon,

– considérer n'importe quelle paire de valeurs successives violant l'ordre souhaité ("adjacent violators") i.e. sélectionner un indice k tel que $y_k > y_{k+1}$,

– regrouper ("pool"), dans la séquence précédente, ces deux valeurs en un seul bloc c'est-à-dire substituer au couple de valeurs (y_k, y_{k+1}) la seule valeur

$$\frac{w_k y_k + w_{k+1} y_{k+1}}{w_k + w_{k+1}}$$

associée au poids $w_k + w_{k+1}$.

(3) Itérer les deux points précédents en considérant la séquence actualisée jusqu'à obtenir une séquence isotonique.

(4) A l'issue de la procédure, pour $i = 1, \ldots, n$, l'estimation $\hat{y}_i$ est égale à la valeur finale associée au bloc dont elle fait partie.

Nous illustrons en figure 1.4 l'application de la méthode sur des relevés de températures moyennes annuelles en Argentine entre 1850 et 2010 (ces données ont été prises comme exemple dans Wu *et al.* (2001) et Tibshirani *et al.* (2011)).

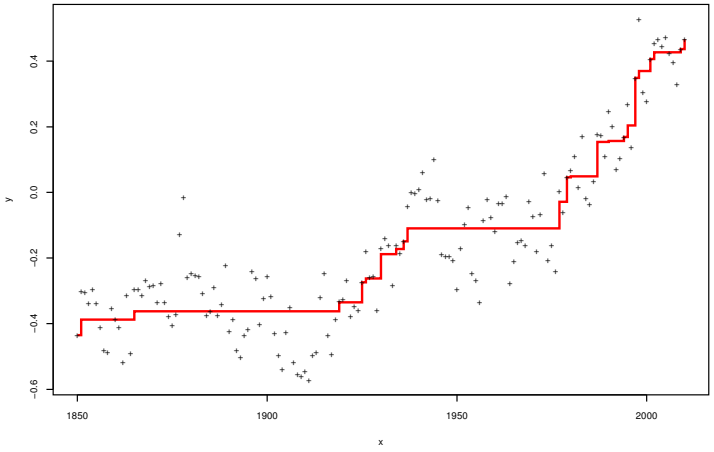


Fig. 1.4 – Régression isotonique sur des données de température.

L'allure de la courbe ajustée correspond à celle d'une fonction constante par morceaux : les observations sont regroupées en un certain nombre de blocs au sein desquels la valeur ajustée est la moyenne des observations. Nous nous autoriserons à employer le terme de "lisseur" pour qualifier la régression isotonique bien que l'ajustement n'ait pas une allure particulièrement lisse. Cependant, les valeurs ajustées étant moins variables que les données elles-mêmes, l'emploi de ce qualificatif ne contredit pas la définition d'un lisseur donnée par Hastie & Tibshirani (1990).

Pour la plupart des estimateurs non paramétriques, on peut contrôler la régularité de l'ajustement

via le réglage d'un paramètre de lissage (la largeur de la fenêtre pour un noyau par exemple). Le manque de régularité de la régression isotonique est souvent présenté comme un inconvénient de la méthode et certains travaux ont ainsi visé à régulariser la régression isotonique : Mukerjee (1988), par exemple, présente un estimateur résultant d'une régression isotonique suivie de l'application d'un noyau destinée à la lisser. Ne supposant le réglage d'aucun paramètre de lissage, l'ajustement par régression isotonique est en revanche adaptatif dans le sens où il est totalement fondé sur les données ("data-driven" en anglais) ce qui est un avantage de la méthode.

Une autre remarque importante est à faire : dans le calcul des estimations, les x_i n'interviennent que par leur rang. La mise en œuvre de l'algorithme suppose donc de ranger préalablement les observations suivant l'ordre des valeurs prises par la variable explicative. Ainsi, une façon naturelle d'obtenir la régression antitonique est de la déduire de la régression isotonique sur les valeurs prises dans l'ordre inverse de l'ordre initial. La démarche est illustrée en figure 1.5 sur un exemple simulé de $n = 100$ points. Partant des données $(x_i, y_i)_{i=1,\dots,n}$ avec $x_1 < \dots < x_n$, on considère les points $(i, y_i)_{i=1,\dots,n}$ représentés à gauche. On relabellise ensuite les y_i dans l'ordre inverse en posant $\tilde{y}_1 := y_n, \dots, \tilde{y}_n := y_1$. Cela donne pour les points $(i, \tilde{y}_i)_{i=1,\dots,n}$, une représentation symétrique de la première (au centre). On calcule la régression isotonique sur les $\tilde{y}_i$ avant de reprendre l'indexation initiale ce qui annule la symétrie précédente et donne l'ajustement antitonique (à droite).

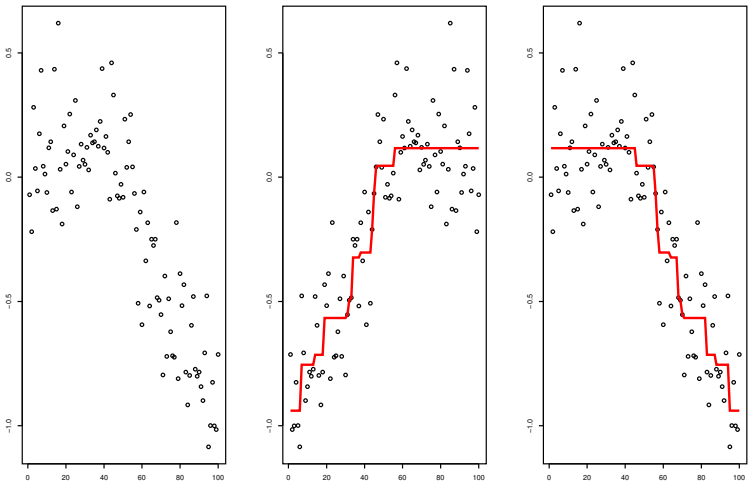


Fig. 1.5 – Régression antitonique.

L'algorithme PAVA est à l'origine de nombreuses améliorations et variantes destinées à des situations plus générales. Tel qu'il est présenté par Ayer *et al.* (1955), il s'applique au cas d'un ordre strict. Cependant, dès les premiers travaux sur la régression isotonique apparaissent des questions où une relation d'ordre moins stricte doit être envisagée. Ainsi, Brunk (1955) considère une application où l'on chercherait à estimer la probabilité $\theta(t) = \theta(t_1, t_2)$ qu'un observateur perçoive un objet ayant la forme d'un segment, t_1 mesurant la distance entre l'observateur et l'objet et t_2 l'angle d'observation. Il pose ainsi le problème d'estimation de paramètres de lois binomiales sous contraintes de monotonie mais, les couples (t_1, t_2) n'étant pas tous comparables, le problème doit être envisagé selon un ordre partiel. Outre certaines évolutions de l'algorithme pour un ordre total (cf. Kruskal (1964) et l'algorithme "Up-and-Down Blocks" par exemple), les recherches portent principalement sur sa transposition à l'ordre partiel, les améliorations visant

à réduire la complexité qui augmente dans ces situations plus générales.

On peut penser que le premier algorithme général est celui suggéré par Brunk (1955) qui sera repris sous le nom de "Minimum Lower Set Algorithm" dans Barlow *et al.* (1972). Dans le même temps, Van Eeden (1957) donne un algorithme qui traite le cas où, en plus du respect de l'ordre, on impose que les estimations restent dans un intervalle donné. Notons aussi les travaux de Thompson (1962) qui avec le "Minimum Violator Algorithm" (resp. "Maximum Violator Algorithm") traite le cas un peu plus général que l'ordre strict où chaque élément ne peut avoir qu'un seul prédécesseur (resp. successeur). Ces algorithmes où le principe initial consistant à "amalgamer" les blocs violant la monotonie est conservé, sont décrits en détail dans Barlow *et al.* (1972), section 2.3, à partir de la page 72.

L'essentiel des recherches qui suivent visent à améliorer les méthodes existantes pour les rendre applicables à des jeux de données toujours plus volumineux. Citons ainsi les travaux de Gebhardt (1970), de Dykstra (1981), de Dykstra & Robertson (1982), de Lee (1983) ou encore ceux de Pardalos & Xue (1999) qui concernent tous des ordres partiels. Une bonne revue des problèmes de complexité est donnée par Best & Chakravarti (1990).

L'approche choisie par ces derniers est fondée sur la connaissance de méthodes développées en optimisation numérique ("active sets methods", cf. Best (1984)). Elle permet de considérer le problème de minimisation associé à la Définition (1.3) comme une forme particulière de problèmes analogues posés sur des arbres et de situer le PAVA, le "Minimum Lower Set Algorithm" et l'algorithme de Van Eeden dans ce cadre plus général. Ils proposent ainsi un algorithme résolvant la régression isotonique avec une complexité $\mathcal{O}(n)$ et se penchent sur la complexité d'algorithmes existants. Ils justifient le fait que l'algorithme "Minimum Lower Set" a une complexité de l'ordre de $\mathcal{O}(n^2)$, retrouvent par un moyen direct la complexité en $\mathcal{O}(n)$ déjà établie dans Grotzinger & Witzgall (1984) de l'algorithme PAVA et montrent que la méthode proposée par Van Eeden a elle une complexité exponentielle.

Il est également intéressant de noter qu'une déclinaison directe de l'algorithme PAVA telle qu'il vient d'être décrit permet de résoudre le problème de régression isotonique L_1

$$\underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} |y_i - u_i|. \tag{1.9}$$

La procédure est la même à ceci près qu'au lieu de prendre, dans le déroulement de l'algorithme, la moyenne des valeurs lors des regroupements, c'est la médiane du groupe qui est choisie. Si l'on se souvient que la médiane M d'une série de valeurs $c_1, \ldots, c_r$ est définie comme un réel c minimisant $\sum_{j=1}^{r} |c - c_j|$, il faut quand même convenir par exemple de

$$M = \frac{c_{\left(\frac{r}{2}\right)} + c_{\left(\frac{r}{2}\right)+1}}{2}$$

pour assurer l'unicité. Le problème, motivé par un exemple comparable à celui de Ayer *et al.* (1955), est proposé initialement par Robertson & Waltman (1968), Robertson & Wright (1980) rectifiant ensuite une erreur présente dans l'algorithme proposé initialement (cf. Chakravarti (1989)).

Intuitivement, étant donnée la complexité en $\mathcal{O}(n)$ pour l'algorithme PAVA, l'appliquer à la résolution du problème (1.9) de régression isotonique L_1 aboutit à une complexité $\mathcal{O}(n^2)$, la détermination de la médiane sur un ensemble de r points nécessitant elle-même r opérations (cf. Chakravarti (1989)). Des améliorations de l'algorithme sont proposées par Menendez & Salvador

(1987), Chakravarti (1989), Pardalos *et al.* (1995) ou bien Stout (2008), améliorations qui dans le meilleur des cas, aboutissent à une complexité de l'ordre de $n \log n$.

Certaines des méthodes qui viennent d'être mentionnées s'étendent au cas de l'ordre partiel, comme celle de Robertson & Wright (1980) ou Stout (2008) par exemple. Notons que certains travaux visent à traiter les gros jeux de données provenant de la génomique en particulier, comme Angelov *et al.* (2006) ou Luss *et al.* (2011). Ces derniers développent une idée de partitionnement récursif de l'espace des variables explicatives (proche de la méthode CART de Breiman *et al.* (1984)) qui fournit une suite de modèles isotoniques tendant vers l'ajustement des moindres carrés. Ajoutons pour finir que les applications et les prolongements de la régression isotonique dépassent finalement assez largement le seul cadre statistique. Les prolongements d'ordre algorithmique sont bien souvent l'œuvre d'informaticiens travaillant notamment dans le domaine de la recherche opérationnelle ou du traitement d'image. On peut penser que, outre le fait que les applications y sont nombreuses (cf. Maxwell & Muckstadt (1985), Roundy (1986), Restrepo Palacios & Bovik (1994) par exemple), la motivation vient de ce que le problème a un intérêt théorique. Il fait en effet partie des rares problèmes quadratiques dont les solutions sont des algorithmes fortement polynomiaux, i.e. polynomiaux en le nombre de données et en la dimension des objets en entrée.

Si l'algorithme PAVA est utilisé en pratique pour déterminer les valeurs ajustées par la régression isotonique, des formulations explicites de l'estimateur sont utiles pour en établir les propriétés théoriques. On les présente dans la section suivante.

1.1.3 "Min-max formulas" et "Greatest Convex Minorant"

Rappelons que Ayer *et al.* (1955) présentent l'algorithme PAVA comme une façon commode d'obtenir l'estimateur du maximum de vraisemblance des paramètres $p_1, \ldots, p_n$ de lois binomiales, ces paramètres étant supposés varier de façon monotone. Dans cet article, les auteurs explicitent par ailleurs l'estimateur qui répond au problème de maximisation de la vraisemblance sous cette contrainte. Dans la situation que nous venons de rappeler et avec les notations que nous avons introduites en début de section précédente page 10, la formulation est la suivante : pour $1 \leq i \leq n$,

$$\begin{array}{rl} \hat{p}_i & = \max_{1\leq u\leq i} \min_{i\leq v\leq n} Av(u,v) \\ & = \min_{i\leq v\leq n} \max_{1\leq u\leq i} Av(u,v) \\ & = \max_{1\leq u\leq i} \min_{u\leq v\leq n} Av(u,v) \\ & = \min_{i\leq v\leq n} \max_{1\leq u\leq v} Av(u,v) \end{array} \tag{1.10}$$

où, pour des entiers u et v tels que $1 \leq u \leq v \leq n$, $Av(u,v)$ est la proportion de succès sur l'ensemble des dosages effectués entre le niveau u et le niveau v.

Tout comme l'algorithme PAVA, les "min-max formulas" vont être généralisées à un grand nombre de cas. Brunk (1955) situe l'estimateur initialement proposé par Ayer *et al.* (1955) dans un cadre plus général. S'appuyant sur des travaux qui ne paraîtront que deux ans plus tard (cf. Brunk *et al.* (1957)), il définit des formules généralisant (1.10) et qui permettent de caractériser les estimateurs du maximum de vraisemblance de paramètres sous contrainte de monotonie, la loi de Y sachant X appartenant plus généralement cette fois à une famille exponentielle de distributions. Dès lors, le cas classique où l'on suppose que Y sachant $X = x_i$ suit une loi normale d'espérance $r(x_i)$ et de variance constante σ^2 apparaît comme un cas particulier de leur étude. Si l'on considère les x_i tous différents comme dans notre cas (cf. équation (1.3) page 8), les formules deviennent ainsi :

$$\begin{array}{ll} \hat{y}_i & = \max_{u\leq i} \min_{v\geq i} Av(u,v) \\ & = \min_{v\geq i} \max_{u\leq i} Av(u,v) \\ & = \max_{u\leq i} \min_{v\geq u} Av(u,v) \\ & = \min_{v\geq i} \max_{u\leq v} Av(u,v) \end{array} \tag{1.11}$$

où $Av(u,v)$ est la moyenne des valeurs $y_u, \ldots, y_v$.

Remarque
Il est intéressant de préciser qu'une interprétation semblable permet de justifier l'emploi de formules analogues pour la régression isotonique L_1. En effet, si l'on considère des variables aléatoires Y_i de densité $f(y,\theta_i) = \frac{1}{2}e^{-|y-\theta_i|}$ (distribution de Laplace) et que l'on cherche à estimer les paramètres θ_i en respectant une contrainte de monotonie, la maximisation de la vraisemblance conduit à chercher le maximum de

$$\sum_{i=1}^{n} |y_i - \hat{\theta}_i|$$

les $\hat{\theta}_i$ respectant la contrainte en question. La solution de ce problème conduit à des estimateurs qui s'explicitent sous la même forme que (1.11), la moyenne étant remplacée par la médiane (cf. Robertson & Waltman (1968) et Cryer *et al.* (1972)).

Une interprétation graphique fort éclairante permet de faire le lien entre l'algorithme PAVA et les "min-max formulas" pour la régression isotonique L_2. Si l'on considère une fonction croissante et intégrable $f : [0,1] \to \mathbb{R}$, on sait que sa fonction cumulative

$$\left\{ \begin{array}{ccc} [0,1] & \to & \mathbb{R} \\ t & \mapsto & F(t) = \int_0^t f(\nu)d\nu \end{array} \right.$$

est convexe. Pour une fonction f non croissante, F n'est pas convexe, mais on se doute qu'une fonction cumulative $\tilde{F}$ "proche" de F sera associée à une fonction croissante $\tilde{f}$ "proche" de f.

Une illustration est donnée en figure 1.6. On représente en noir une fonction f à gauche et la courbe de sa fonction cumulative F de la même couleur à droite. On adopte la terminologie anglo-saxonne en appelant la courbe de F le CSD de f pour "Cumulative Sum Diagram".

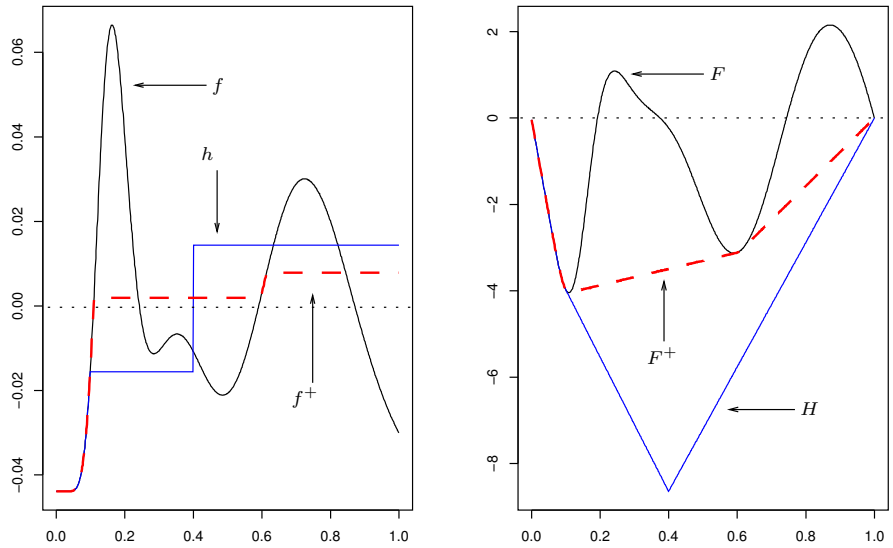


Fig. 1.6 – Graphe des fonctions (à gauche) et des fonctions cumulatives (à droite).

Sont par ailleurs représentées à droite en figure 1.6 les courbes cumulatives F^+ et H de deux fonctions croissantes f^+ et h dont les courbes figurent à gauche. On observe que la courbe de F^+ joue un rôle particulier, puisqu'elle pourrait se déduire de celle de F en tendant une corde le long de la partie inférieure de cette dernière. En ce sens elle représente la fonction convexe maximale dominée par F. La courbe de F^+ est ainsi appelée le "Greatest Convex Minorant" de F, que nous notons en abrégé GCM.

Formellement F^+ est la fonction définie point par point par

$$F^+ := \operatorname{argmax} \{H \text{ convexe et } \forall t \in [0,1], H(t) \leq F(t)\}. \tag{1.12}$$

On sait (cf. Hörmander (2007), théorème 1.1.9) que F^+ est dérivable partout à droite et à gauche et que la dérivée à gauche est donnée par

$$\left(F^+\right)'(t) = \max_{u<t} \min_{v \geq t} \frac{F(v) - F(u)}{v - u}.$$

Dans notre cas, cela donne :

$$\left(F^+\right)'(t) = \max_{u<t} \min_{v \geq t} \frac{1}{v-u} \int_u^v f(\nu) d\nu. \tag{1.13}$$

Le lien avec la régression isotonique est donné par le théorème suivant tiré de Anevski & Soulier (2011) :

Théorème 1.1 (Anevski & Soulier (2011))
Soit f intégrable sur $[0,1]$ et $F : t \mapsto F(t) = \int_0^t f(\nu)d\nu$ alors

$$\left(F^+\right)' = \operatorname*{argmin}_{h \in \mathcal{C}^+} \int_0^1 \left(f(\nu) - h(\nu)\right)^2 d\nu.$$

En résumé, la régression isotonique de f s'obtient par dérivation à gauche de la fonction F^+, fonction dont la courbe est le GCM de F. Par ailleurs, une forme explicite de cette fonction notée f^+ est donnée par la formule (1.13) :

$$f^+(t) = \left(F^+\right)'(t) = \max_{u<t} \min_{v \geq t} \frac{1}{v-u} \int_u^v f(\nu) d\nu. \tag{1.14}$$

La quantité $\frac{1}{v-u} \int_u^v f(\nu) d\nu$ correspondant à la valeur moyenne de la fonction f entre u et v, l'analogie entre ce cas fonctionnel et les formules (1.11) données par Ayer *et al.* (1955) pour le cas discret est directe.

Nous avons choisi de citer les travaux récents d'Anevski & Soulier pour présenter les choses car en quelques lignes et de façon complètement rigoureuse, leur article donne à la fois l'équation (1.13) et établit le Théorème 1.1 permettant de faire le lien direct avec la régression isotonique. Il est en effet assez difficile d'établir une chronologie claire des résultats en lien avec le GCM qui émergent dans les années cinquante. La justification de l'emploi des "min-max formulas" n'est pas clairement donnée dans Ayer *et al.* (1955), les auteurs renvoyant la preuve à un papier ultérieur sans donner de référence. Cependant, en recoupant Brunk (1956) et Brunk *et al.* (1957), on obtient un résultat analogue exprimé sous forme un peu plus générale que celui de Anevski & Soulier (2011). Avec X une distribution continue sur $[0, 1]$, l'analogue de (1.14) est

$$f^+(t) = \max_{u<t} \min_{v \geq t} \frac{\int_u^v f(\nu) dX(\nu)}{\int_u^v dX(\nu)}$$

et il est dit (sans justification) que cette fonction minimise

$$\int_0^1 \left(f(\nu) - h(\nu)\right)^2 dX(\nu)$$

parmi les fonctions h non décroissantes. La formulation de Anevski & Soulier (2011) apparaît alors comme le cas particulier d'une distribution uniforme. Notons aussi que cette interprétation graphique est introduite de manière indépendante de Brunk et ses co-auteurs mais à la même époque par Grenander (1956) dans le cadre de l'estimation de densité.

L'algorithme PAVA trouve ainsi une interprétation graphique naturelle. Nous reprenons le petit exemple $y = (1, 0, 1, 1, 0, 2)'$ vu en section précédente et sur lequel nous avons illustré l'application de l'algorithme PAVA (cf. figure 1.3 page 12).

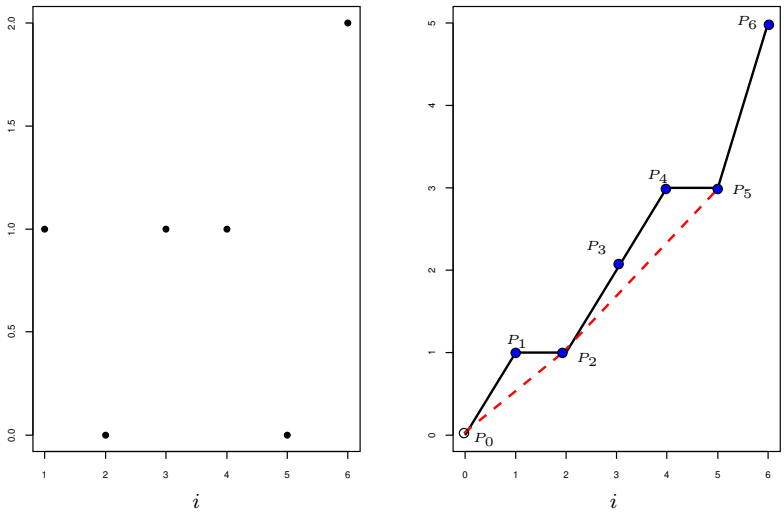


Fig. 1.7 – Les points (i, y_i) à gauche et le CSD à droite.

Le CSD est la courbe joignant par des segments les points (P_i, P_{i+1}) où $P_0 = (0, 0)$ et $P_i = (i, \sum_{j=1}^{i} y_j)$ pour $i \geq 1$. Sur ce graphe, la pente du segment reliant deux points consécutifs P_i et P_{i+1} correspond à y_{i+1}. Ainsi, si la pente de $[P_i, P_{i+1}]$ est inférieure à celle de $[P_{i+1}, P_{i+2}]$, cela signifie que $y_{i+2} > y_{i+1}$. Le graphe offre aussi une interprétation commode des valeurs moyennes. Par exemple, la pente de $[P_0, P_2]$ correspond à $(y_1 + y_2)/2$ soit à la moyenne de y_1 et y_2, et celle de $[P_2, P_5]$ à $(y_3 + y_4 + y_5)/3$ soit à la moyenne de y_3, y_4 et y_5. Voyons en figure 1.8 les différentes étapes de l'algorithme sur l'exemple.

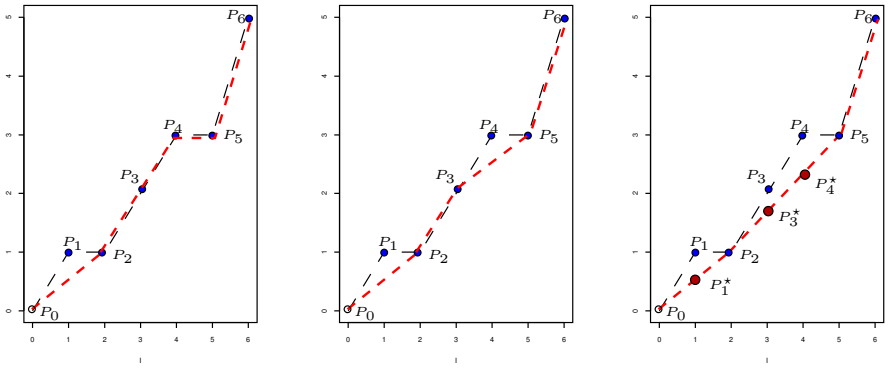


Fig. 1.8 – Tracé du GCM.

La pente de (P_0P_1) étant supérieure à celle de (P_1P_2), c'est que $y_2 < y_1$. On remplace la ligne brisée reliant les points P_0 et P_2 via P_1 par celle reliant directement P_0 à P_2 : cette étape correspond au regroupement ("pool") de y_1 et de de y_2 en la seule valeur $(y_1 + y_2)/2$ associée à la pente de (P_0P_2). Il faut de même regrouper y_4 et y_5 qui violent la croissance en la seule valeur $(y_4 + y_5)/2$ (que l'on schématise par le tracé de (P_3P_5)) puis encore cette valeur avec y_3 en la seule valeur $(y_3 + y_4 + y_5)/3$ (regroupement schématisé par le tracé de (P_2P_5)) pour obtenir

le graphe d'une fonction convexe. Ce graphe est celui de la fonction convexe maximale dominée par la fonction associée au CSD, c'est-à-dire le GCM. Les valeurs ajustées $\hat{y}_i$ correspondent aux dérivées à gauche du GCM aux points $P_i^\star$ d'abscisses i, c'est-à-dire à la pente à gauche en chacun des points $P_i^\star$. Nous retrouvons ainsi les valeurs ajustées obtenues en figure 1.3 :

$$\begin{aligned}
\hat{y_1} = \hat{y_2} &= \frac{y_1 + y_2}{2} = 1/2 \\
\hat{y_3} = \hat{y_4} = \hat{y_5} &= \frac{y_3 + y_4 + y_5}{3} = 2/3 \\
\hat{y_6} &= y_6 = 2
\end{aligned}$$

Remarque
Lorsque les observations y_i sont pondérées par des poids ω_i, le principe est le même à ceci près que les points du CSD ont pour abscisses $\sum_{j=1}^{i} \omega_j$ et que la pente de la $i^{\text{ème}}$ valeur ajustée est

$$\hat{y}_i = \frac{z_{P_i^\star} - z_{P_{i-1}^\star}}{\omega_i - \omega_{i-1}},$$

z désignant ici l'ordonnée du point sur le graphe (cf. Barlow *et al.* (1972) page 11 pour un exemple).

Poursuivons l'exemple en vérifiant que l'on retrouve ces valeurs de pentes en s'aidant des "min-max formulas". On reprend par exemple la $3^{\text{ème}}$ formulation dans l'équation (1.11) :

$$\hat{y}_i = \max_{u \leq i} \min_{v \geq u} Av(u, v).$$

Considérons que l'on cherche à calculer $\hat{y}_i$ pour $i = 4$. A $u \leq 4$ fixé, on envisage les pentes des segments reliant P_u à chacun des points P_v, $v \geq u$. On retient la plus petite de ces pentes : $\min_{v \geq u} Av(u, v)$ (représentée en rouge pour $u = 3$ sur la figure 1.9). On répète cette opération en considérant successivement tous les $u \leq 4$ et la plus grande de ces pentes minimales donne $\hat{y}_i$ (représentée en vert).

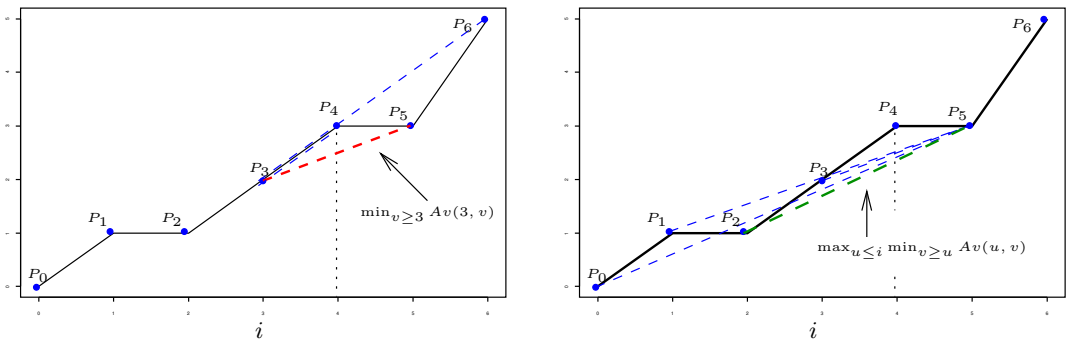


Fig. 1.9 – Interprétation des "min-max formulas" sur le GCM.

Nous reprenons le Théorème 1.1 page 12 de Barlow *et al.* (1972) qui justifie, dans le cas fini, l'emploi du GCM pour obtenir la régression isotonique :

Théorème 1.2
Soit $\hat{y}_i$ la pente gauche au point d'abscisse i du GCM. On a, pour toute fonction f isotonique sur $\mathcal{X} = \{x_1 < x_2 < \ldots < x_n\}$,

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 \geq \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 + \sum_{i=1}^{n} [\hat{y}_i - f(x_i)]^2 . \tag{1.15}$$

Ainsi, les valeurs $(\hat{y}_i)_{i=1\ldots n}$ donnent les valeurs ajustées de la régression isotonique de la suite $\{y_1, \ldots, y_n\}$.

Preuve
Il revient au même de montrer qu'en retranchant le terme de droite au terme de gauche dans (1.15), on obtient une quantité positive. On montre facilement que cette quantité vaut :

$$2 \sum_{i=1}^{n} [y_i - \hat{y}_i] [\hat{y}_i - f(x_i)] .$$

D'après le lemme d'Abel, on sait que, pour deux suites (a_i) et (b_i),

$$\sum_{i=1}^{n} a_i b_i = a_n B_n - \sum_{i=1}^{n} B_{i-1}(a_i - a_{i-1})$$

où $B_i = \sum_{j=1}^{i} b_j$ pour $i = 1, \ldots, n$ et $B_0 = 0$.
Nous posons $a_i = \hat{y}_i - f(x_i)$ et $b_i = y_i - \hat{y}_i$. Il vient alors :

$$\begin{aligned}\sum_{i=1}^{n} [y_i - \hat{y}_i] [\hat{y}_i - f(x_i)] = & [\hat{y}_n - f(x_n)] \left(\sum_{i=1}^{n} (y_i - \hat{y}_i) \right) \\ & + \sum_{i=1}^{n} \left((f(x_i) - f(x_{i-1})) \sum_{j=1}^{i-1} (y_j - \hat{y}_j) \right) \\ & - \sum_{i=1}^{n} \left((\hat{y}_i - \hat{y}_{i-1}) \sum_{j=1}^{i-1} (y_j - \hat{y}_j) \right) .\end{aligned}$$

Le dernier point du CSD coïncide avec le dernier point du GCM. On a donc $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$ et le premier terme est nul (on a déjà vu cette propriété comme une conséquence directe de la Proposition 1.1 page 9).
Considérons un élément quelconque $(\hat{y}_i - \hat{y}_{i-1})(\sum_{j=1}^{i-1}(y_j - \hat{y}_j)$ de la somme formant le $3^{\text{ème}}$ terme. Il y a deux possibilités. Soit au point d'abscisse $i-1$, le CSD coïncide avec le GCM, soit il est strictement au-dessus. Dans ce second cas, le couple (y_{i-1}, y_i) viole l'ordre souhaité et les valeurs sont regroupées pour donner $\hat{y}_{i-1} = \hat{y}_i$ de sorte que, sur le GCM, les pentes gauches en $i-1$ et i sont les mêmes. Dans les deux cas, cet élément est donc nul. Ainsi, chacun des éléments de la somme formant le $3^{\text{ème}}$ terme est nul.
Enfin, comme le CSD est au-dessus du GCM et que la fonction f est supposée isotonique, le second terme est positif ou nul. L'inégalité (1.15) est donc vraie. □

Pour la régression antitonique, on a l'analogue des "min-max formulas" ainsi qu'une interprétation sur la courbe cumulative des valeurs. Il suffit d'inverser min et max dans les formules (1.11) pour

avoir l'équivalent antitonique :

$$\begin{aligned} \hat{y}_i &= \min_{u\leq i} \max_{v\geq i} Av(u,v) \\ &= \max_{v\geq i} \min_{u\leq i} Av(u,v) \\ &= \min_{u\leq i} \max_{v\geq u} Av(u,v) \\ &= \max_{v\geq i} \min_{u\leq v} Av(u,v) \end{aligned} \tag{1.16}$$

où $Av(u,v)$ est la moyenne des valeurs $y_u, \ldots, y_v$.

En traçant le CSD associé aux observations puis, selon le principe précédent, la courbe concave qui en est la plus proche, on obtient le LCM (pour "Least Concave Majorant"). Les valeurs ajustées par régression antitonique s'interprètent à nouveau comme les pentes à gauche aux points d'abscisse i de cette courbe (cf. figure 1.10).

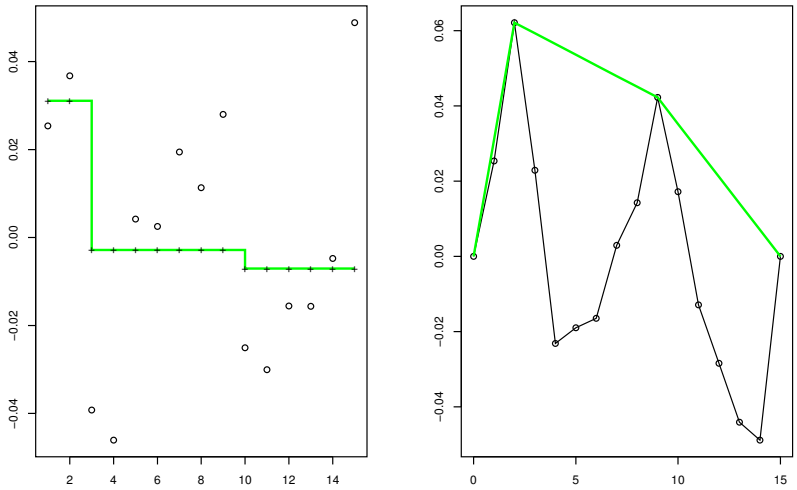


Fig. 1.10 – Régression antitonique (à gauche) et LCM (à droite).

Les "min-max formulas" et l'approche par "Greatest Convex Minorant" ont permis d'obtenir la plupart des résultats de consistance puis de vitesse de convergence de l'estimateur de régression isotonique. Rappelons tout d'abord que Ayer *et al.* (1955) montrent que les "min-max formulas" donnent les estimateurs du maximum de vraisemblance pour les paramètres de probabilités d'un ensemble de lois binomiales, ces probabilités étant soumises à une contrainte de monotonie. Le cas est étendu par Brunk (1955) à des familles exponentielles de distributions et pour un ordre partiel. Ces deux articles étudient également la consistance des estimateurs. La formulation la plus générale, qui est énoncée dans le second article, est un résultat de consistance faible : pour le cas d'un ordre strict et pour ces familles de distributions, il assure qu'en tout point de continuité de la fonction de régression, l'estimateur sera proche de la vraie valeur avec probabilité 1 pourvu que l'on dispose de part et d'autre de ce point de suffisamment d'observations. Un peu plus tard, Brunk, s'appuyant sur des problèmes de minimisations de fonctions convexes sur l'intersection d'ensembles convexes fermés (et rejoignant en cela une partie des travaux de Van Eeden (1956) et Van Eeden (1957)), obtient, moyennant l'hypothèse de densité des x_i, un résultat de convergence uniforme presque sûre (sur un intervalle fermé inclus dans $]0,1[$) sans référence à une famille particulière de distributions.
Il complète ces travaux en développant la théorie mathématique de la régression isotonique dans un cadre abstrait de théorie de la mesure (cf. Brunk (1965)). Il étend ainsi la définition d'espérance

conditionnelle usuelle par rapport à une tribu à une espérance conditionnelle par rapport à un σ-lattice, laquelle confère à la régression isotonique certaines propriétés de projection. Certes cette projection n'est pas linéaire mais on récupère de nombreuses propriétés communes à celles des martingales, ce qui permet de retrouver certains théorèmes classiques de convergence. Cela lui permet ainsi d'exprimer les solutions aux problèmes déjà mentionnés d'estimation du maximum de vraisemblance de paramètres de distributions de familles exponentielles en tant qu'espérance conditionnelle par rapport à un σ-lattice.
Pour conclure sur les travaux de Brunk, citons l'article "Estimation of Isotonic Regression" qu'il publie en 1970. Ce papier corrige tout d'abord une erreur présente dans Brunk (1958) où l'hypothèse de densité, qui en réalité ne suffit pas, est renforcée pour obtenir la consistance uniforme. L'auteur énonce également un théorème sur la distribution asymptotique de l'estimateur en un point, résultat qui sera ensuite abondamment repris comme référence. Avec en particulier l'hypothèse que dans un voisinage d'un point particulier x_0, la dérivée r' de la fonction de régression est continue (et non nulle au point d'intérêt), il établit que

$$c \times n^{\frac{1}{3}}[\hat{r}(x_0) - r(x_0)]$$

avec c une constante fonction de $r'(x_0)$, converge en distribution vers la pente à gauche en 0 du GCM de $W(t) + t^2$ où $W(.)$ est un mouvement brownien symétrique standard ($\hat{r}(x_0)$ désigne l'estimateur habituel pris au point x_0 : on l'obtient en prolongeant l'estimateur entre les points du design soit par des constantes soit par interpolation linéaire). Il obtient donc une vitesse de convergence de l'estimateur de l'ordre de $n^{-1/3}$ (cf. Brunk (1970)).

Hanson *et al.* (1973) reprennent les résultats de consistance de Brunk un peu plus tard. Ils apportent quelques améliorations en relâchant les conditions sur les moments imposées sur les erreurs et obtiennent en plus un résultat dans le cas bi-dimensionnel. Wright poursuit ces travaux en s'intéressant à l'estimation de fonctions strictement monotones. Arguant du fait que l'estimateur de régression isotonique étant constant par morceaux, il peut sembler inadapté lorsque la fonction de régression est strictement monotone, il propose un estimateur combinant régressions isotonique et linéaire dont il étudie les performances sur des exemples simulés (cf. Wright (1978)). Notons que la question avait déjà été envisagée par Barlow et ses co-auteurs sous un angle différent, puisque ceux-ci supposaient connus des réels γ_i tels que $r(x_{i+1}) \geq r(x_i) + \gamma_i$ et s'appuyaient sur la seule régression isotonique (cf. Barlow *et al.* (1972) page 58). Wright affine également le résultat de Brunk (1970) sur la vitesse de convergence de l'estimateur. Plus précisément, il montre que si la fonction de régression a ses $\alpha - 1$ premières dérivées nulles en x_0 et que la dérivée d'ordre α est strictement positive en ce point, la vitesse devient $n^{-\alpha/(2\alpha+1)}$, la distribution asymptotique étant toujours associée à un mouvement brownien standard (cf. Wright (1981)).

Citons également les travaux récents de Durot (2007) qui, dans un cadre unifié, complète l'étude asymptotique. Sous la condition que la fonction de régression soit differentiable et de dérivée strictement positive, elle établit entre autres que

$$\mathbb{E}\left[\int_0^1 |\hat{r}_n(t) - r(t)|^p dt\right] = \mathcal{O}(n^{-p/3})$$

donnant ainsi une mesure de l'erreur L_p associée à l'estimateur. Elle démontre de plus un théorème central limite pour la distribution asymptotique de cette erreur.

Enfin, nous avons déjà mentionné le parallèle existant entre régressions isotoniques L_1 et L_2, aussi bien du point de vue de l'algorithme PAVA (cf. page 15) que des "min-max formulas" (cf. page 17). La comparaison se poursuit avec l'étude asymptotique. Des résultats de consistance

sont donnés dans Robertson & Waltman (1968) puis Cryer *et al.* (1972). Wang & Huang (2002) établissent par exemple une convergence en $n^{-1/3}$ de l'estimateur vers la pente à gauche en 0 du GCM d'un processus Brownien.

1.2 Modèle additif / Algorithme backfitting

L'estimation non paramétrique a connu de nombreux développements ces quarante dernières années. Les méthodes usuelles sont cependant difficilement applicables en pratique dès que la dimension devient grande ; nous expliquons pourquoi en section 1.2.1. Une façon de contourner cette inconvénient consiste à faire des hypothèses structurelles sur la fonction de régression. Le modèle additif en est un exemple classique ; nous le présentons en section 1.2.2. Nous voyons en section 1.2.3 l'algorithme backfitting, généralement utilisé pour estimer ces modèles. Nous concluons cette section par quelques résultats complémentaires présents dans la littérature.

Pour présenter les résultats qui suivent, nous nous plaçons dans un cadre de régression multivariée. Nous considérons ainsi le modèle

$$Y = r(X) + \varepsilon$$

avec $r : \mathbb{R}^d \to \mathbb{R}$, $d > 1$.

1.2.1 Le fléau de la dimension

Le terme de "fléau de la dimension" (curse of dimensionality) a été évoqué pour la première fois par Bellman (1961). Dans un cadre statistique, il traduit le fait que lorsque la dimension d augmente, il est de plus en plus difficile d'observer des points dans un voisinage de taille raisonnable autour d'un point particulier de l'espace. On illustre en figure 1.11 le fait que pour couvrir 10% de la superficie d'un carré de côté 1, il est nécessaire de choisir un carré de côté $l = \sqrt{0,1} \approx 0,31$. En dimension 3, l est tel que $l^3 = 0,1$ soit $l = (0,1)^{1/3} \approx 0,46$ et en dimension d, $l = (0,1)^{1/d}$.

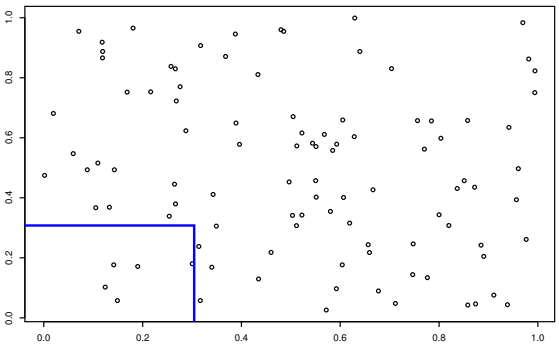


Fig. 1.11 – Fléau de la dimension.

De manière plus générale, couvrir la proportion p d'un hypercube de côté 1 supposera de considérer un hypercube de côté $l = p^{1/d}$. La figure 1.12, qui représente $p \mapsto p^{1/d}$ pour différentes valeurs de d, montre la difficulté à rencontrer des points dans un voisinage qui reste "local" quand d augmente.

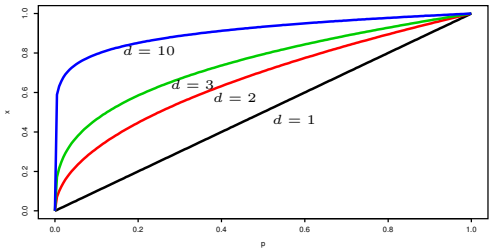


Fig. 1.12 – Fléau de la dimension.

Ainsi, pour les estimateurs à noyaux de type Nadaraya-Watson (Nadaraya (1964), Watson (1964)) ou ceux de régression polynomiale locale (Fan & Gijbels (1996)) qui sont basés sur des moyennes locales, il est clair que, d augmentant, il faudra considérer des voisinages de plus en plus grands autour d'un point x particulier de l'espace pour espérer capturer des observations et estimer $\hat{r}(x)$. Les valeurs des prédicteurs pour les points présents dans l'échantillon sont ainsi souvent très éloignées de la région d'intérêt. On est donc contraint d'utiliser de grandes fenêtres pour capturer des observations, ce qui a pour conséquence de produire des estimateurs en général trop lisses et donc bien souvent biaisés.

Pour ce qui est des méthodes de régression pénalisée comme les splines plaque mince (la généralisation multivariée des splines de lissage), leur simple mise en œuvre est même bien souvent impossible pour de grandes valeurs de d. Pour des compléments sur les splines de lissage et les splines plaque mince, on peut se référer à Green & Silverman (1994), ainsi qu'à Wahba (1990) ou Gu (2002) pour l'approche par noyaux reproduisants. Pour ce type de lisseurs, on considère le problème de minimisation

$$\underset{f:\mathbb{R}^d\to\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda J[f] \tag{1.17}$$

avec ici $f : \mathbb{R}^d \to \mathbb{R}$ et $J[f]$ une quantité mesurant les variations de f et définie par

$$J[f] = \sum_{\alpha_1+\ldots+\alpha_d=m} \frac{m!}{\alpha_1!\ldots\alpha_d!} \times \int_{\mathbb{R}^d} \left(\frac{\partial^m f}{\partial t_1^{\alpha_1}\ldots\partial t_d^{\alpha_d}}\right)^2 dt_1\ldots dt_d. \tag{1.18}$$

Dans (1.17), le premier terme mesure la fidélité aux données alors que le second pénalise les grandes variations de f. Le principe est donc de trouver un compromis entre qualité de l'ajustement et pénalisation des trop grandes variations de la fonction ajustée, le paramètre de lissage λ venant contrôler l'importance du terme de pénalisation. Pour avoir une spline continue, il faut $m > d/2$ et à λ fixé, la solution de (1.17) s'exprime dans une base de $M = \begin{pmatrix} m+d-1 \\ d \end{pmatrix}$ polynômes. Or ce nombre minore le nombre de paramètres à estimer pour construire la spline. Le tableau suivant, qui donne quelques valeurs de M pour des couples (m, d) admissibles, montre que M devient vite très grand quand d augmente, ce qui suppose donc de disposer d'un nombre au moins aussi grand de données pour la simple mise en œuvre de la méthode :

d	m	M
2	2	3
5	3	21
8	5	495
10	6	1001

L'aspect asymptotique donne un point de vue complémentaire à cette question du fléau de la dimension. Considérons par exemple le cas d'un estimateur à noyau. On sait (cf. Györfi *et al.* (2002), théorème 5.2 page 77) qu'avec un noyau uniforme et pour une fonction lipschitzienne, le terme de biais s'écrit

$$\int_{\mathbb{R}^d} (\text{biais}(\hat{r}(x)))^2 \mu(dx) = c_1 h^2 + o\left(h^2\right)$$

et celui de variance

$$\int_{\mathbb{R}^d} \text{var}(\hat{r}(x)) \mu(dx) = c_2 \frac{1}{nh^d} + o\left(\frac{1}{nh^d}\right).$$

Le choix d'une fenêtre optimale conduit alors à la convergence en moyenne quadratique de l'estimateur vers la fonction de régression à une vitesse de l'ordre $\mathcal{O}(n^{-2/(d+2)})$. Il est clair que cette vitesse diminue avec la dimension. Ainsi, lorsque la dimension augmente, il faut un plus grand nombre de points qu'en dimension 1 si l'on veut retrouver la même qualité d'estimation.

1.2.2 Le modèle additif

Les modèles additifs et l'algorithme "backfitting" destiné à les estimer ont été introduits par Friedman & Stuetzle (1981) dans le cadre de la méthode "projection pursuit". Ils ont fait ensuite l'objet de nombreux travaux. On peut citer en particulier Stone (1985), Buja *et al.* (1989) et Hastie & Tibshirani (1990). Postuler un modèle additif consiste à supposer que la fonction r du modèle $Y = r(X) + \varepsilon$ s'écrit comme la somme de fonctions univariées :

$$r(X) = r(X_1, \ldots, X_d) = \alpha + r_1(X_1) + \ldots + r_d(X_d).$$

Ainsi, le modèle s'écrit :

$$Y = \alpha + \sum_{j=1}^{d} r_j(X_j) + \varepsilon. \tag{1.19}$$

En plus des hypothèses habituelles $\mathbb{E}[\varepsilon | X_j] = 0$, on a les hypothèses implicites :

$$\forall j, \qquad \mathbb{E}[r_j(X_j)] = 0. \tag{1.20}$$

Ces dernières signifient que les fonctions sont privées de tout terme constant (ces éventuels termes étant portés par α) et assurent l'identification du modèle.

Un tel modèle peut être vu comme une généralisation du modèle de régression multiple puisqu'on y retrouve une forme additive mais sans imposer la contrainte de linéarité sur les fonctions individuelles. En ce sens, il représente un compromis entre le modèle linéaire et un lisseur multidimensionnel. Le fait de ne pas donner de forme paramétrique particulière aux fonctions r_j accorde au modèle une certaine flexibilité et l'aspect additif autorise une interprétation de l'effet individuel de chaque variable. Détaillons ce point ; une fonction additive r à deux variables est représentée en figure 1.13. Nous observons que les coupes de sa surface faites en fixant l'une des variables donnent des fonctions de l'autre représentées par des courbes parallèles. Il suffit ainsi d'avoir l'allure d'un seul représentant de cette famille de courbes pour pouvoir mesurer l'effet d'un prédicteur en particulier. On décompose ainsi l'effet global des prédicteurs en une somme d'effets individuels.

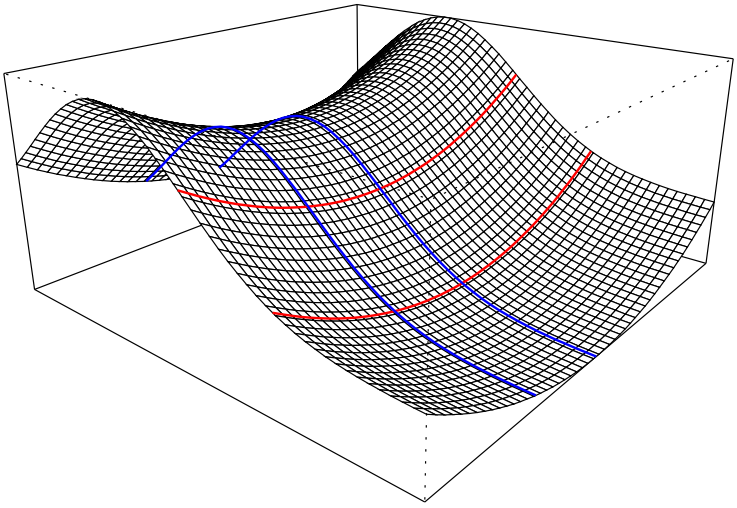


Fig. 1.13 – Un exemple de fonction additive.

En quoi le modèle additif permet-il de contourner le fléau de la dimension ? L'élément essentiel est que l'estimation de r en un point particulier de l'espace se fait en estimant individuellement chacune des composantes. Si l'on considère par exemple un estimateur à noyau, on pourra estimer la composante r_j en une valeur $x^{(j)}$ particulière, en prenant en compte toutes les observations de X dont les $j^{\text{èmes}}$ coordonnées tombent dans le voisinage de $x^{(j)}$ et ce indépendamment des autres dimensions. On illustre ceci en figure 1.14 pour deux dimensions.

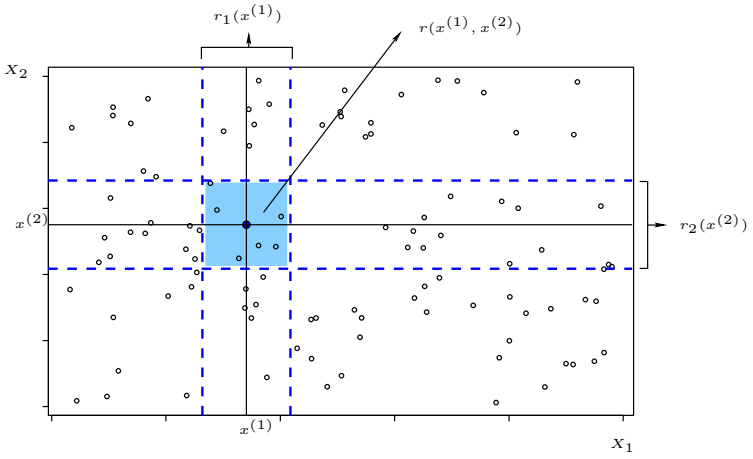


Fig. 1.14 – Modèle additif et fléau de la dimension.

Pour un lisseur multivarié classique, on n'utilise pour calculer $\hat{r}(x)$, que les observations appartenant au voisinage du point d'intérêt $x = \left(x^{(1)}, x^{(2)}\right)$, à savoir les points présents dans la zone représentée en bleu. En revanche avec un modèle additif, tous les points de la bande verticale seront utilisés pour estimer la première composante $r_1(x^{(1)})$ et tous les points de la bande horizontale utilisés pour estimer $r_2(x^{(2)})$. Ce faisant, on utilise finalement le même nombre d'observations

qu'en régression sur une variable.

D'un point de vue asymptotique, la conséquence naturelle est que, construisant d estimateurs univariés, on peut retrouver pour chaque direction et donc pour l'estimateur global, une vitesse optimale qui correspond à l'optimal univarié (vitesse de l'ordre $\mathcal{O}(n^{-2/(3)})$ pour le cas envisagé plus haut par exemple). Stone (1985) établit ce résultat et construit un estimateur basé sur des splines qui atteint cette vitesse.

Notons enfin que le modèle peut être adapté en fonction des connaissances que l'on pourrait avoir sur les variables. Ainsi, on peut imaginer faire du lissage paramétrique dans certaines directions particulières, ajouter des interactions dans le modèle ou bien regrouper certaines variables entre elles et appliquer un lisseur multidimensionnel au bloc ainsi constitué.

Nous présentons maintenant l'algorithme le plus courant d'estimation de ces modèles : l'algorithme backfitting.

1.2.3 Estimation : l'algorithme backfitting

L'algorithme backfitting a été introduit par Breiman & Friedman (1985) mais nous nous inspirons essentiellement de Buja *et al.* (1989) et Hastie & Tibshirani (1990) pour le présenter ici.

Postuler le modèle additif (1.19) revient à considérer que $r(X) = \mathbb{E}[Y|X]$, qui est la projection de Y sur $L^2(X)$, coïncide avec la projection de Y sur le sous-espace $L^2(X_1) + \ldots + L^2(X_d)$:

$$r(X) = \underset{g(X)\in L^2(X_1)+\ldots+L^2(X_d)}{\operatorname{argmin}} \mathbb{E}\left[(Y - g(X))^2\right]. \tag{1.21}$$

$r(X)$ est ainsi caractérisée par $Y - r(X) \perp L^2(X_1) + \ldots + L^2(X_d)$ ce qui équivaut à $Y - r(X) \perp L^2(X_j)$ pour tout j. Si l'on interprète l'espérance conditionnelle $\mathbb{E}[.|X_j]$ comme la projection orthogonale sur $L_2(X_j)$, on a ainsi

$$\forall j = 1 \ldots d \qquad P_j\left(Y - r(X)\right) = 0$$

soit (pour simplifier les écritures, on considère $\alpha = 0$)

$$r_j(X_j) = P_j\left(Y - \sum_{k\neq j} r_k(X_k)\right) = \mathbb{E}\left[Y - \sum_{k\neq j} r_k(X_k)\middle| X_j\right].$$

Formellement, $r(X) = \sum_{j=1}^{d} r_j(X_j)$ solution de (1.19) équivaut à $(r_1(X_1), \ldots, r_d(X_d))$ solutions du système d'équations dites normales :

$$\begin{pmatrix} I & P_1 & P_1 & \ldots & P_1 \\ P_2 & I & P_2 & \ldots & P_2 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ P_d & P_d & P_d & \ldots & I \end{pmatrix} \begin{pmatrix} r_1(X_1) \\ r_2(X_2) \\ . \\ . \\ . \\ r_d(X_d) \end{pmatrix} = \begin{pmatrix} P_1(Y) \\ P_2(Y) \\ . \\ . \\ . \\ P_d(Y) \end{pmatrix}. \tag{1.22}$$

Buja *et al.* (1989) présentent l'algorithme backfitting comme une procédure de résolution de type Gauss-Seidel (cf. Hageman & Young (1981)) de la transposition sur les données du système (1.22). Cette transposition s'écrit de la manière suivante :

$$\begin{pmatrix} I & S_1 & S_1 & \dots & S_1 \\ S_2 & I & S_2 & \dots & S_2 \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ S_d & S_d & S_d & \dots & I \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ . \\ . \\ . \\ r_d \end{pmatrix} = \begin{pmatrix} S_1(y) \\ S_2(y) \\ . \\ . \\ . \\ S_d(y) \end{pmatrix}. \tag{1.23}$$

où les S_j sont les lisseurs unidimensionnels appliqués dans chaque direction. La méthode de Gauss-Seidel de résolution de ce système donne l'algorithme backfitting qui s'écrit de la manière suivante :

Algorithme 2 Algorithme backfitting

(1) Initialisation : $\hat{\alpha} = \bar{y}$ et pour $j = 1 \dots d$

$$\hat{r}_j = 0.$$

(2) Cycle : pour $j = 1 \dots d$,

$$\hat{r}_j(X_j) = S_j \left(Y - \hat{\alpha} - \sum_{k \neq j} \hat{r}_k(X_k) \middle| X_j \right).$$

(3) Itérer (2) jusqu'à ce que toutes les $\hat{r}_j$ varient "peu" d'une itération à la suivante.

Tel qu'est écrit l'algorithme, on commence donc par estimer la constante du modèle (1.19) en calculant la moyennes des y_i et on initialise toutes les composantes en les mettant à 0. L'idée est ensuite d'estimer tour à tour chacune des composantes en conservant les autres fixées aux dernières estimations : on actualise ainsi dans la $j^{\text{ème}}$ étape du cycle, l'estimation de $\hat{r}_j$ en lissant les résidus courants $Y - \sum_{k \neq j} \hat{r}_k(X_k)$.

Précisons qu'en pratique, la façon d'initialiser la procédure dépend de la connaissance que l'on peut avoir sur les fonctions r_j. Sans connaissance a priori, il est d'usage de fixer $\hat{r}_j = 0$ comme on vient de l'écrire ou de déterminer les premières estimations en effectuant une régression linéaire multiple. Il est également possible, suivant les données que l'on étudie, d'utiliser des lisseurs propres aux différentes directions de l'espace. Il est clair par ailleurs que l'ordre dans lequel les lissages sont effectués peut avoir une importance sur les estimations individuelles voire sur leur somme.

Il est naturel de se demander pourquoi on ne cherche pas à inverser le système (1.23) plutôt que de chercher une solution de manière itérative. Il faut remarquer que chacun des $S_j(y)$ correspond au vecteur des valeurs ajustées par lissage de $y = (y_1, \dots, y_n)'$ dans la direction j. Il s'agirait donc d'inverser un système $nd \times nd$ ce qui, à moins de travailler avec des petits jeux de données, est d'un coût opératoire prohibitif.

L'existence et l'unicité de la solution du système (1.23) sont liées au spectre des lisseurs utilisés. Pour la régression isotonique itérée, détaillons le cas $d = 2$ qui nous intéresse avant tout. Le système s'écrit alors

$$\begin{pmatrix} I & S_1 \\ S_2 & I \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} S_1(y) \\ S_2(y) \end{pmatrix}.$$

Il équivaut à

$$\left\{ \begin{array}{rcl} r_1 & = & S_1(y - r_2) \\ r_2 & = & S_2(y - r_1) \end{array} \right. \tag{1.24}$$

et sous réserve d'existence des matrices inverses, sa solution est

$$\left\{ \begin{array}{rcl} r_1 & = & (I - S_1S_2)^{-1}S_1(I - S_2)y \\ r_2 & = & (I - S_2S_1)^{-1}S_2(I - S_1)y \end{array} \right. . \tag{1.25}$$

Une synthèse des résultats pour ce cas est donnée dans Hastie & Tibshirani (1990). On peut retenir que si $\|S_1S_2\| < 1$ et $\|S_2S_1\| < 1$ alors l'algorithme backfitting converge (lorsque le nombre d'itérations augmente) en restituant à la limite les bonnes composantes et ce indépendamment de l'ordre dans lequel on effectue les lissages. Il faut noter que cette condition est assez restrictive car, si une propriété de base d'un lisseur est que les valeurs ajustées sont moins variables que les valeurs initiales (soit pour $y \in \mathbb{R}^n$, $\|S(y)\| \leq \|y\|$ et donc $\|S\| \leq 1$), l'inégalité n'est pas toujours stricte. Elle ne peut pas s'appliquer au splines cubiques par exemple. En effet, si l'on sait que les valeurs propres d'une spline cubique sont dans $[0, 1]$, deux d'entre elles valent 1 puisque ces lisseurs conservent les fontions constantes et les fonctions linéaires. On peut malgré tout montrer que si S_1 et S_2 sont des lisseurs symétriques, de valeurs propres appartenant à l'intervalle $]-1, 1]$ (c'est le cas pour les splines cubiques), le système (1.24) a au moins une solution. L'algorithme backfitting converge alors vers l'une d'elles et le vecteur des valeurs ajustées obtenues à la limite, $\hat{r}_1^{(\infty)} + \hat{r}_2^{(\infty)}$, est indépendant des fonctions prises au départ. Dans le cas où $||S_1S_2|| = 1$, les limites $\hat{r}_1^{(\infty)}$ et $\hat{r}_2^{(\infty)}$ dépendront par contre des fonctions choisies initialement.

On peut citer pour terminer cette partie quelques travaux complémentaires. Ansley & Kohn (1994) étendent au cas $d \geq 2$ l'application de l'algorithme backfitting lorsque les lisseurs sont symétriques avec des valeurs propres comprises dans $[0, 1]$ (donc pour des splines de lissage par exemple). Ils donnent une preuve géométrique en s'appuyant sur une généralisation de l'algorithme de Von Neumann proposée par Halperin (1962). Härdle & Hall (1993) étudient l'algorithme backfitting lorsque les lisseurs employés correspondent à des projections (bin smoothers, splines de régression avec des nœuds fixes par exemple). Ils montrent que dans ce cas, la limite obtenue peut s'écrire comme l'application aux données d'un opérateur global de projection. Ils vérifient que de manière générale, l'estimation limite globale $\sum_j \hat{r}_j^{\infty}$ ne dépend pas de l'ordre dans lequel sont appliqués les lisseurs et précisent des conditions pour avoir invariance des estimations limites individuelles des termes de la somme. Les auteurs explicitent la variance de leur estimateur et montrent ainsi sa consistance.

Opsomer & Ruppert (1997) explorent, dans le cas bivarié, l'application de régressions polynomiales locales. Les auteurs partent des solutions explicites aux équations normales (1.25). Ils spécifient ensuite les conditions sur les noyaux et les tailles de fenêtres pour lesquelles les matrices inverses existent et donc pour lesquelles de telles solutions peuvent s'écrire ainsi. Ils explicitent ensuite les termes de biais et de variance de l'estimateur et retrouvent les vitesses de convergence univariée sous une condition supplémentaire portant sur la régularité des fonctions r_1 et r_2. Leur travail est complété par Opsomer (2000) pour $d \geq 2$ quelques années plus tard. De leur côté, Mammen *et al.* (1999) considèrent un estimateur construit en appliquant préalablement un lisseur multivarié à noyau de type Nadaraya-Watson ou de régression polynomiale locale. Ils s'appuient sur les travaux de Mammen *et al.* (2001) (publiés plus tard mais déjà écrits à l'époque) pour donner une interprétation de ce type de lisseurs en termes de projections dans un espace de Hilbert muni d'une norme adaptée. L'idée consiste ensuite à estimer la projection de cet estimateur sur l'espace des fonctions additives par une procédure de type backfitting. Ils montrent la convergence de la méthode et étudient les propriétés asymptotiques de l'estimateur.

Ce faisant, ils retrouvent le biais et la variance des estimateurs oracles. Ces travaux sont ensuite généralisés à d'autres types de lisseurs par Horowitz *et al.* (2006) .

L'algorithme backfitting est également utilisé par Mammen & Yu (2007) pour estimer les composantes d'un modèle additif isotonique de régression c'est-à-dire un modèle où la fonction de régression s'écrit

$$r(X) = \sum_{j=1}^{d} r_j(X_j)$$

avec les r_j toutes croissantes. Le backfitting est ici utilisé de manière classique avec, pour l'actualisation de $\hat{r}_j$, l'application de la régression isotonique sur les résidus

$$Y - \sum_{k \neq j} \hat{r}_k.$$

Ils montrent qu'en augmentant le nombre d'itérations, l'algorithme converge vers la solution de

$$\underset{u \in \mathcal{C}_n^+ + \ldots + \mathcal{C}_n^+}{\operatorname{argmin}} \sum_{j=1}^{n} (Y_j - u_j)^2$$

avec $\mathcal{C}_n^+$ le cône habituel des séquences isotoniques de $\mathbb{R}^n$. Pour obtenir ce résultat, ils montrent que dans ce cas, l'algorithme backfitting peut être vu comme une adaptation de l'algorithme donné par Dykstra (1983) qui résoud le problème dual consistant à déterminer le projeté d'un vecteur de $\mathbb{R}^n$ sur l'intersection d'un ensemble de cônes convexes fermés (nous revenons sur ce point en remarque page 41). Dans cet article, les auteurs établissent également un résultat asymptotique analogue à ceux qui ont été montrés pour des lisseurs classiques sous un modèle additif usuel. Ils retrouvent en effet en chaque composante la distribution asymptotique univariée :

$$\forall j = 1, \ldots, d \qquad \forall x_j \in]0,1[\qquad n^{1/3} C \left[\hat{r}_j(x_j) - r_j(x_j)\right]$$

converge en distribution vers la pente du GCM de $W(t)+t^2$ où $W(t)$ est un mouvement brownien symétrique. Notons que l'obtention de ce résultat suppose entre autres hypothèses qu'aucune des fonctions r_j ne comporte de partie constante.

Pour conclure, disons que pour contourner le problème du fléau de la dimension, le fait de postuler un modèle additif présente le double avantage d'offrir une interprétation facilitée de l'influence individuelle des prédicteurs et d'obtenir des vitesses de convergence satisfaisantes. Il faut cependant garder présent à l'esprit que postulant ce modèle, on suppose que la fonction de régression a en effet une structure additive, ce qu'on ignore en général. Les méthodes dont on vient de parler visent ainsi à s'approcher au mieux de la fonction additive la plus proche de la fonction liant X à Y mais rien ne garantit que cette fonction ait en effet cette structure. En d'autres termes, le contrôle du terme de biais dans le calcul de l'erreur quadratique globale

$$\begin{array}{lll} \mathbb{E}[(\hat{r}(X) - r(X))^2] & = \mathbb{E}[(\hat{r}(X) - \mathbb{E}[\hat{r}(X)])^2] & + (\mathbb{E}[\hat{r}(X) - r(X)])^2 \\ & = \operatorname{var}(\hat{r(X)}) & + \operatorname{biais}^2(\hat{r}(X)). \end{array}$$

est obtenu sous cette hypothèse mais le rendre faible, dans ce cas, ne donne cependant aucune garantie quant à la bonne spécification du modèle. La méthode proposée par Cornillon *et al.* (2011) ne fait pas d'hypothèse structurelle sur la fonction de régression. Elle est basée sur la réduction itérée du biais d'où son nom : IBR pour "Iterative Bias Reduction". Le principe, proche du "boosting" (cf. Bühlmann & Yu (2003) ou Di Marzio & Taylor (2008) par exemple) consiste

à appliquer dans un premier temps un estimateur multivarié classique en donnant une grande valeur au paramètre de lissage. On s'accommode ainsi du fléau de la dimension mais on produit un estimateur biaisé. L'idée est ensuite d'estimer le biais, toujours au moyen d'un lisseur multivarié avec un grand paramètre de lissage, de corriger l'estimateur précédent puis d'itérer la procédure.

Un article publié dans la revue *Statistics and Computing* figure en annexe page 147. Il présente plus en détail la méthode IBR et compare ses performances avec des concurrents multivariés. Ce principe de réduction itérée du biais est également à l'origine de l'idée d'un autre algorithme que le backfitting pour itérer la régression isotonique et ainsi estimer une fonction de régression dans un cadre univarié. Nous voyons cela dans le chapitre qui suit.

Chapitre 2

Régression isotonique itérée

2.1 Introduction - Estimateurs envisagés

Commençons par rappeler l'idée qui motive la méthode. On considère le modèle de régression

$$Y = r(X) + \varepsilon \tag{2.1}$$

avec X à valeurs dans $[0,1]$, $\mathbb{E}\left[Y^2\right] < \infty$ et $\mathbb{E}\left[\varepsilon | X\right] = 0$. On suppose que r appartient à l'ensemble $\mathcal{F}$ des fonctions à variation bornée sur $[0,1]$. La caractérisation donnée par Jordan de telles fonctions comme somme d'une fonction croissante et d'une fonction décroissante permet de donner au modèle (2.1) la forme additive

$$Y = u(X) + b(X) + \varepsilon$$

où u appartient au cône $\mathcal{C}^+$ des fonctions croissantes et b au cône $\mathcal{C}^-$ des fonctions décroissantes. Cette décomposition n'est pas unique en général ce qui pose un problème d'identifiabilité dans la décomposition $r = u + b$. Cependant, pour des fonctions continues à droite, le Théorème 0.1 donné en introduction page 3, assure l'unicité de la décomposition pour u et b ayant des mesures de Stieltjes mutuellement singulières. Concrètement, cela signifie qu'à une constante additive près, il y a une unique décomposition $r = u^\star + b^\star$ telle que localement, les fonctions $u^\star$ et $b^\star$ ne varient pas simultanément. Ainsi lorsque $u^\star$ est croissante, $b^\star$ est constante et inversement. Cette décomposition peut être vue comme la décomposition à variation minimale de r dans la mesure où les variations locales de r sont portées par $u^\star$ lorsque r est croissante, par $b^\star$ lorsque r est décroissante, mais jamais par les deux termes.

Nous considérons finalement le modèle

$$Y = r(X) + \varepsilon = u^\star(X) + b^\star(X) + \varepsilon \tag{2.2}$$

avec en plus des hypothèses précédentes, r continue à droite en tout point et par exemple, les conditions

$$\int_0^1 u^\star(x)\mu(dx) = \int_0^1 r(x)\mu(dx) \qquad \text{et} \qquad \int_0^1 b^\star(x)\mu(dx) = 0 \tag{2.3}$$

qui assurent l'identifiabilité du modèle. Un exemple est représenté en figure 2.1.

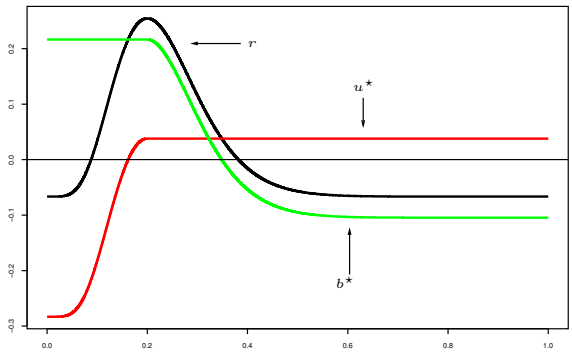


Fig. 2.1 – Décomposition de Stieltjes de la fonction r.

On peut voir l'équation (2.2) comme un modèle additif mettant en jeu partie croissante et partie décroissante. C'est ce point de vue qui donne l'idée de la méthode car il suggère de combiner les outils d'estimation des modèles additifs et les outils de régression sous contraintes de forme pour estimer les termes de la somme. Nous proposons ainsi d'appliquer la régression isotonique selon l'algorithme backfitting pour estimer $u^\star$ et $b^\star$ et en déduire r. Avant de préciser l'algorithme, nous rappelons certaines définitions et introduisons quelques notations utiles.

Nous considérons $(X_i, Y_i)_{i=1,\dots,n}$ des réplications i.i.d. du modèle (2.2). La régression isotonique suppose d'utiliser les observations réordonnées selon l'ordre des X_i. Nous les notons $\left(X_{(i)}, Y_{(i)}\right)$ avec $X_{(1)} < \ldots < X_{(n)}$ car la fonction de répartition de X est supposée continue. Pour $i = 1, \dots, n$, nous notons x_i (resp. y_i) les observations des variables aléatoires $X_{(i)}$ (resp. $Y_{(i)}$). Nous supposons ainsi que nous disposons des observations (x_i, y_i) avec $x_1 < \ldots < x_n$ et $y_1, \dots, y_n$ les observations correspondantes.

Nous allons appliquer alternativement régressions isotonique et antitonique. Rappelons que nous notons $\mathrm{iso}(y)$ la régression isotonique de $y = (y_1, \dots, y_n)'$ définie par

$$\mathrm{iso}(y) = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|y - u\|_n^2 = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - u_i)^2$$

avec $\mathcal{C}_n^+$ le cône constitué des vecteurs de $\mathbb{R}^n$ dont les coordonnées forment une séquence croissante. De même, la régression antitonique de y est

$$\mathrm{anti}(y) = \underset{b \in \mathcal{C}_n^-}{\operatorname{argmin}} \|y - b\|_n^2 = \underset{b \in \mathcal{C}_n^-}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - b_i)^2$$

avec $\mathcal{C}_n^-$ le cône de $\mathbb{R}^n$ dont les coordonnées forment une séquence décroissante.

Par ailleurs pour $z = (z_1, \dots, z_n)'$ un vecteur de $\mathbb{R}^n$, on pose $\Delta(z)$ le vecteur de $\mathbb{R}^{n-1}$ défini par

$$\Delta(z) = (z_2 - z_1, \dots, z_n - z_{n-1})'. \tag{2.4}$$

Pour deux vecteurs z et $\tilde{z}$, on pose $\Delta(z) \circ \Delta(\tilde{z})$ le vecteur de $\mathbb{R}^{n-1}$ résultat du produit terme à terme de $\Delta(z)$ et $\Delta(\tilde{z})$:

$$\Delta(z) \circ \Delta(\tilde{z}) = ((z_2 - z_1) \times (\tilde{z}_2 - \tilde{z}_1), \dots, (z_n - z_{n-1}) \times (\tilde{z}_n - \tilde{z}_{n-1}))'. \tag{2.5}$$

Ainsi, si pour deux vecteurs z et $\tilde{z}$, on a $\Delta(z) \circ \Delta(\tilde{z}) = 0$, cela signifie que, si entre deux composantes consécutives z varie, c'est que $\tilde{z}$ ne varie pas et vice-versa.

Nous proposons d'adapter l'algorithme backfitting présenté en page 30 à notre situation de la manière suivante. Nous donnons à la méthode le nom de Régression Isotonique Itérée, I.I.R. en abrégé, pour "Iterative Isotonic Regression" :

Algorithme 3 Iterative Isotonic Regression

(1) Initialisation :

$$\begin{aligned}\hat{u}^{(0)} &= \left(\hat{u}_1^{(0)}, \ldots, \hat{u}_n^{(0)}\right)' = 0\\ \hat{b}^{(0)} &= \left(\hat{b}_1^{(0)}, \ldots, \hat{b}_n^{(0)}\right)' = 0\\ \hat{y}^{(0)} &= \hat{u}^{(0)} + \hat{b}^{(0)} = 0\end{aligned}$$

(2) Cycle : pour $k \geq 1$

$$\begin{aligned}\hat{u}^{(k)} &= \text{iso}\left(y - \hat{b}^{(k-1)}\right)\\ \hat{b}^{(k)} &= \text{anti}\left(y - \hat{u}^{(k)}\right)\\ \hat{y}^{(k)} &= \hat{u}^{(k)} + \hat{b}^{(k)}.\end{aligned} \tag{2.6}$$

(3) Itérer (2) jusqu'à une condition d'arrêt à définir.

Remarques

- La notation vectorielle est utilisée dans cet algorithme. Ainsi, après k itérations, le vecteur des valeurs ajustées est
$$\hat{y}^{(k)} = \left(\hat{y}_1^{(k)}, \ldots, \hat{y}_n^{(k)}\right)'.$$
En prolongeant par des segments horizontaux par exemple, on définit l'estimateur sur $[0, 1]$.
- On verifiera en Annexe C que l'on a pour tout $k \geq 1$
$$\begin{aligned}&\Delta\left(\hat{u}^{(k)}\right) \circ \Delta\left(\hat{b}^{(k)}\right) = 0\\ &\bar{\hat{u}}^{(k)} = \bar{y}\\ &\bar{\hat{b}}^{(k)} = 0,\end{aligned} \tag{2.7}$$
où $\bar{y}$ est la moyenne des composantes de y. Ces équations garantissent l'identifiabilité de la décomposition $\hat{y}^{(k)} = \hat{u}^{(k)} + \hat{b}^{(k)}$. Elles sont en effet la transposition au cas discret des conditions (2.3) vues au-dessus pour le cas fonctionnel. Ainsi, disposant de $\hat{y}^{(k)}$, on déduit de manière unique les termes $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$ (à une constante près) en faisant porter les augmentations locales de $\hat{y}^{(k)}$ à $\hat{u}^{(k)}$ (pendant que $\hat{b}^{(k)}$ reste constant) et les diminutions locales de $\hat{y}^{(k)}$ à $\hat{b}^{(k)}$ (pendant que $\hat{u}^{(k)}$ reste constant). En choisissant de donner à $\hat{u}^{(k)}$ la moyenne $\bar{y}$, l'unicité est assurée et non plus seulement à une constante près.
- Pour compléter le point précédent, ajoutons que si l'on considérait, dans l'alorithme 3, d'autres estimateurs monotones que iso(.) et anti(.), il faudrait éventuellement ajouter des conditions d'identifiabilité telles que (2.7).
- Nous avons fait le choix d'appliquer d'abord la régression isotonique puis la régression antitonique sur les résidus. Ce choix est arbitraire et nous aurions aussi bien pu prendre les opérations en sens inverse. Sauf mention contraire, nous appliquerons les étapes dans cet ordre.

- La formulation du backfitting en page 30 n'est pas tout à fait la même que celle-ci. On aurait pu les faire coïncider en commençant par calculer $\bar{y}$, en appliquant ensuite l'algorithme sur les données centrées $y - \bar{y}$ et en prenant à la fin de chaque cycle $\hat{y}^{(k)} = \bar{y} + \hat{u}^{(k)} + \hat{b}^{(k)}$. Cela ne change pas grand-chose en fait si l'on se souvient que, de manière générale, la moyenne est conservée par régression isotonique (cf. équation (1.6) page 10). En choisissant de commencer chaque cycle par la régression isotonique, la moyenne des y est ainsi portée par les $\hat{u}^{(k)}$, les $\hat{b}^{(k)}$ étant de moyenne nulle : cela démontre au passage les deux derniers points dans (2.7).

La figure 2.2 montre l'application de l'algorithme sur l'exemple de la fonction considérée en figure 2.1. Nous avons distribué aléatoirement $n = 100$ points autour de cette fonction avec ε un bruit gaussien d'espérance nulle. L'ajustement est représenté pour $k = 1, 10, 1000$ itérations.

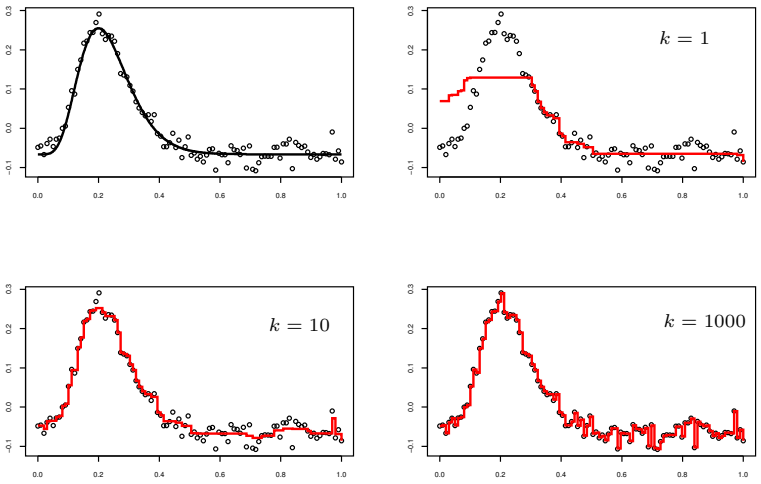


Fig. 2.2 – Application de l'algorithme I.I.R. pour $k = 1, 10, 1000$.

On remarque que l'augmentation du nombre d'itérations tend à reproduire les données. Nous allons montrer dans cette section que c'est effectivement le cas (cf. Théorème 2.1 page 39). Auparavant, nous donnons une interprétation géométrique à l'algorithme, ce qui facilite sa compréhension et rejoint la démonstration du théorème.

La régression isotonique iso(y) correspond au projeté orthogonal du vecteur y sur $\mathcal{C}_n^+$ (cf. figure 1.1 page 10). De même anti(y) est le projeté de y sur $\mathcal{C}_n^-$, le cône opposé à $\mathcal{C}_n^+$. Ainsi, on obtient $\hat{u}^{(1)}$ en projetant y sur $\mathcal{C}_n^+$. On considère ensuite le résidu $y - \hat{u}^{(1)}$, que l'on projette sur $\mathcal{C}_n^-$ pour obtenir le premier ajustement décroissant $\hat{b}^{(1)}$. La somme des deux donne l'ajustement $\hat{y}^{(1)}$ à l'issue de l'étape 1 de l'algorithme (cf. partie supérieure de la figure 2.3). On considère ensuite le résidu $y - \hat{b}^{(1)}$ après cette estimation décroissante, résidu que l'on projette sur $\mathcal{C}_n^+$ dans le but d'améliorer l'estimation de la partie croissante. On obtient ainsi $\hat{u}^{(2)}$, que l'on retranche ensuite à y, le résultat étant projeté à son tour sur $\mathcal{C}_n^-$ pour obtenir $\hat{b}^{(2)}$ (cf. partie inférieure de la figure 2.3). La somme $\hat{y}^{(2)} = \hat{u}^{(2)} + \hat{b}^{(2)}$ fournit l'estimation à la fin de l'étape 2. Le même procédé est ensuite répété.

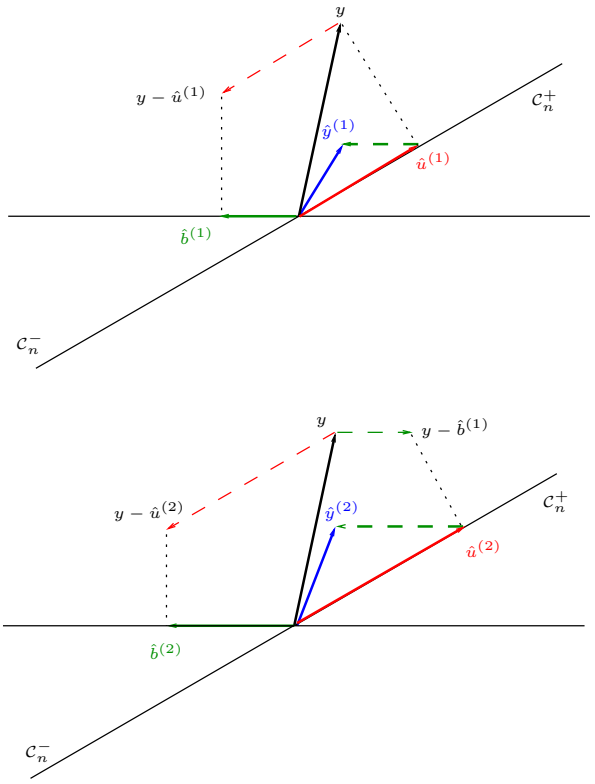


Fig. 2.3 – Interprétation géométrique de l'algorithme I.I.R.

Lorsque l'on augmente le nombre d'itérations, $\hat{y}^{(k)}$ se rapproche de y illustrant le fait que l'on tend vers l'interpolation des données. Nous en venons au théorème qui formalise ce résultat :

Théorème 2.1 (Interpolation des données)
Avec les notations précédentes, les estimations produites par l'algorithme 3 tendent vers l'interpolation des données :

$$\lim_{k\to\infty} \hat{y}^{(k)} = y.$$

La preuve que nous proposons s'appuie sur des résultats de Bauschke & Borwein (1994) qui analysent l'algorithme de Dykstra (1983) destiné initialement à trouver le projeté d'un point particulier de $\mathbb{R}^n$ sur l'intersection d'un nombre fini de cônes convexes fermés. Le travail de Bauschke & Borwein (1994) se place dans le cadre général des espaces de Hilbert et il contient également une analyse de l'algorithme de Von Neumann. Ce sont leurs résultats sur cet algorithme qui nous intéressent ici. En effet, l'algorithme I.I.R. peut être vu comme un cas particulier de l'algorithme de Von Neumann, ce que nous vérifions dans la preuve du Théorème 2.1.

Auparavant, pour situer le contexte, précisons que l'algorithme de projection alternée de Von Neumann est considéré comme la première méthode pour déterminer le projeté d'un point particulier sur l'intersection d'ensembles convexes fermés dans un espace de Hilbert. Dans sa forme originale, Von Neumann (1950) construit à partir d'un point initial x et de deux sous-espaces fermés A et B d'un espace de Hilbert $\mathcal{H}$, les séquences (a_n) et (b_n) suivantes :

$$b_0 := x \qquad a_k := P_A(b_{k-1}) \qquad b_k := P_B(a_k) \tag{2.8}$$

et montre la convergence des deux séquences vers $P_{A\cap B}(x)$. Dykstra (1983) suggère un algorithme plus général qui coïncide avec la solution précédente pour ce même problème et qui traite en plus le cas où A et B sont des cônes convexes fermés dans un espace euclidien. Boyle & Dykstra (1986) montrent ensuite que l'algorithme proposé en 1983 donne en fait une solution pour le cas de deux sous-ensembles convexes fermés dans un Hilbert. Bauschke & Borwein (1994) complètent la description de l'algorithme de Dykstra et les travaux précédents en précisant notamment la convergence. Ils en déduisent également des résultats complémentaires sur l'algorithme de Von Neumann lorsque celui-ci équivaut à celui de Dykstra.

Pour ce qui nous intéresse, on trouve en particulier dans leurs travaux le résultat suivant : si on pose $v = P_{\overline{B-A}}(0)$, projection de 0 sur la fermeture de $B - A$, alors

$$b_k - a_k \to v \qquad \text{et} \qquad b_k - a_{k+1} \to v. \tag{2.9}$$

La preuve du Théorème 2.1 repose sur le fait que l'on peut interpréter les séquences $(\hat{u}^{(k)})_{k\geq 1}$ et $(y - \hat{b}^{(k)})_{k\geq 1}$ comme des séquences de Von Neumann analogues à (2.8). Pour cela, on s'appuie sur une interprétation géométrique complémentaire de la précédente et qui s'articule sur les deux résultats qui suivent.

Lemme 2.1 (Cônes translatés)
Soit $\mathcal{H}$ un espace de Hilbert. Soit $\mathcal{M}$ un sous-ensemble convexe fermé et $h \in \mathcal{H}$. On note $P_{\mathcal{M}}(h)$ le projeté de h sur $\mathcal{M}$ i.e. le point de $\mathcal{M}$ le plus proche de h. Soit $a \in \mathcal{H}$, alors

$$P_{\mathcal{M}+a}(h) = a + P_{\mathcal{M}}(h - a). \tag{2.10}$$

où $\mathcal{M} + a = \{m + a, m \in \mathcal{M}\}$.

Lemme 2.2
Soit $y \in \mathbb{R}^n$, on a

$$\text{iso}(-y) = -\,\text{anti}(y) \qquad \textit{i.e.} \qquad P_{\mathcal{C}_n^+}(-y) = -P_{\mathcal{C}_n^-}(y). \tag{2.11}$$

Preuve (du Théorème 2.1)
On revient à notre méthode et on introduit, pour se ramener au Lemme 2.1, le cône résultant de la translation de vecteur y opérant sur le cône $\mathcal{C}_n^+$. On le note $y + \mathcal{C}_n^+$:

$$y + \mathcal{C}_n^+ = \{y + u, u \in \mathcal{C}_n^+\}.$$

On illustre cette définition en figure 2.4. Cette figure illustre également le fait que l'on peut voir les termes $(\hat{u}^{(k)})_{k\geq 1}$ et $(y - \hat{b}^{(k)})_{k\geq 1}$ comme des projections alternées sur les cônes $\mathcal{C}_n^+$ et $y + \mathcal{C}_n^+$, ce que l'on vérifie maintenant.

Par définition, on a $\hat{u}^{(1)} = P_{\mathcal{C}_n^+}(y)$. En prenant (2.10) avec $\mathcal{H} := \mathbb{R}^n$ muni de la norme $\|.\|_n$, $\mathcal{M} := \mathcal{C}_n^+$ et $a := y$ puis en utilisant (2.11), il vient :

$$P_{y+\mathcal{C}_n^+}(\hat{u}^{(1)}) = y + P_{\mathcal{C}_n^+}(\hat{u}^{(1)} - y) = y - P_{\mathcal{C}_n^-}(y - \hat{u}^{(1)}).$$

En revenant à la définition $\hat{b}^{(1)} = P_{\mathcal{C}_n^-}(y - \hat{u}^{(1)})$, on a donc

$$y - \hat{b}^{(1)} = P_{y+\mathcal{C}_n^+}(\hat{u}^{(1)}).$$

De même, par définition, on a $\hat{u}^{(2)} = P_{\mathcal{C}_n^+}(y - \hat{b}^{(1)})$ et

$$\begin{aligned} y - \hat{b}^{(2)} &= y - P_{\mathcal{C}_n^-}(y - \hat{u}^{(2)}) && \text{par définition de } \hat{b}^{(2)} \\ &= y + P_{\mathcal{C}_n^+}(\hat{u}^{(2)} - y) && \text{d'après (2.11)} \\ &= P_{y+\mathcal{C}_n^+}(\hat{u}^{(2)}) && \text{d'après (2.10).} \end{aligned}$$

De proche en proche, en posant $\hat{b}^{(0)} = 0$, on a ainsi

$$\forall k \geq 1 : \qquad \hat{u}^{(k)} = P_{\mathcal{C}_n^+}(y - \hat{b}^{(k-1)}) \qquad \text{et} \qquad y - \hat{b}^{(k)} = P_{y+\mathcal{C}_n^+}(\hat{u}^{(k)}). \tag{2.12}$$

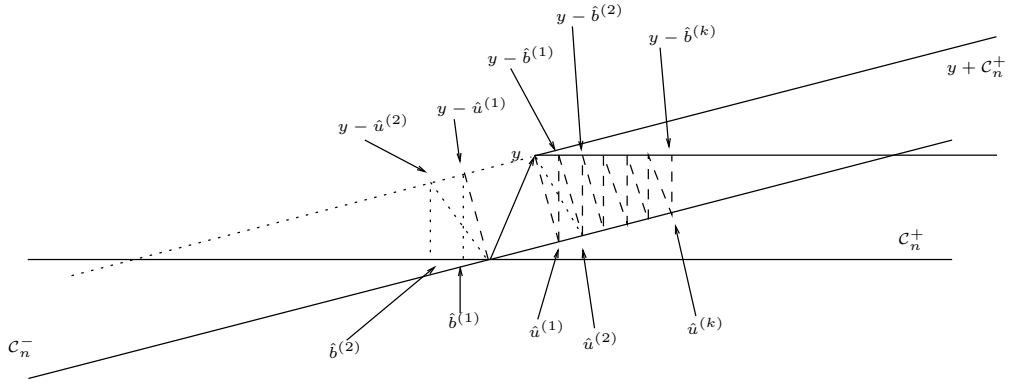


Fig. 2.4 – Interprétation de l'algorithme I.I.R. comme un algorithme de Von Neumann.

En prenant $x := 0$, $A := \mathcal{C}_n^+$ et $B := y + \mathcal{C}_n^+$ dans (2.8), il est clair que les séquences $(\hat{u}^{(k)})_{k\geq 1}$ et $(y - \hat{b}^{(k)})_{k\geq 1}$ peuvent être identifiées aux séquences de Von Neumann :

$$a_k := \hat{u}^{(k)} \qquad \text{et} \qquad b_k := y - \hat{b}^{(k)}.$$

Comme $0 \in \overline{B - A}$, $v = 0$. D'après (2.9), on a donc

$$y - \hat{b}^{(k)} - \hat{u}^{(k)} \to 0$$

ce qui se traduit par

$$\lim_{k\to\infty} \|\hat{y}^{(k)} - y\|_n = 0$$

et achève la démonstration du Théorème (2.1). □

Remarques

- Il faut noter que le Théorème 2.1 assure la convergence de la somme $\hat{u}^{(k)} + \hat{b}^{(k)}$ mais ne précise pas si les termes $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$ pris individuellement convergent. Cependant, Bauschke & Borwein (1994) donnent en corollaire du théorème que nous avons cité un résultat de Cheney & Goldstein (1959) qui permet d'avoir, dans notre cas, la convergence des termes de la somme. A notre connaissance, en revanche, on ne peut pas savoir quelles sont les limites.
- En revanche, de façon générale, rien n'est dit sur ces limites individuelles. Dans notre cas particulier, nous montrons en Théorème 2.3 que les suites $\left(u^{(k)}\right)$ et $\left(b^{(k)}\right)$ convergent vers les termes $u_y^\star$ et $b_y^\star$ de la décomposition à variation minimale de y.

- Tendre vers l'interpolation n'est pas souhaitable d'un point de vue statistique car cela revient à faire du surajustement. Nous envisagerons ainsi dans le Chapitre 3 consacré aux applications plusieurs critères d'arrêt à la méthode.
- Dans le cadre multivarié où l'on applique classiquement le backfitting, il est d'usage d'arrêter l'algorithme lorsque les estimations individuelles ne varient plus sensiblement d'une itération à l'autre. Ce critère conduirait ici à une situation proche de l'interpolation et il est naturel de s'interroger sur le fait que l'on ne puisse pas faire référence au cadre classique pour l'appliquer. On donnera une explication à ce sujet en Section 2.2.3 page 49.
- Il est possible de s'appuyer sur l'algorithme de Dykstra (cf. Dykstra (1983)) plutôt que sur celui de Von Neumann pour prouver le Théorème 2.1. Le principe est de considérer la décomposition de Moreau (cf. Moreau (1962)) dans le cas particulier d'un vecteur de $\mathbb{R}^n$ (la décomposition vaut dans les espaces de Hilbert). Tout vecteur de $\mathbb{R}^n$ peut s'écrire la somme du projeté de ce vecteur sur un cône convexe fermé et du projeté sur le cône polaire de ce cône (on donne une définition du cône polaire page 109). Dès lors, en considérant les séquences de l'algorithme de Dykstra construites par projection sur les cônes polaires de $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$, on retrouve $y - \hat{b}^{(k-1)} - \hat{u}^{(k)}$ et $y - \hat{u}^{(k)} - \hat{b}^{(k)}$. Le théorème 3.1 de Dykstra, tout comme les résultats de Bauschke & Borwein (1994) sur l'analyse de cet algorithme, permettent alors également de conclure. C'est la manière dont est utilisé l'algorithme backfitting dans Mammen & Yu (2007).

Nous proposons un second algorithme qui s'inspire cette fois de la méthode de réduction itérée du biais I.B.R. proposée par Cornillon *et al.* (2011). Le principe consiste à travailler à chaque étape, non plus sur les résidus partiels, mais sur les résidus courants $y - \hat{y}^{(k)}$, puis à actualiser l'estimateur. Plus précisément, l'algorithme proposé est le suivant :

Algorithme 4 Iterative Isotonic Bias Reduction

(1) Initialisation :

$$\hat{y}^{(0)} = \hat{u}^{(0)} + \hat{b}^{(0)} = 0$$

(2) Cycle : pour $k \geq 1$

$$\begin{array}{ll} \tilde{u}^{(k)} & = \text{iso}\left(y - \hat{y}^{(k-1)}\right) \\ \tilde{b}^{(k)} & = \text{anti}\left(y - \hat{y}^{(k-1)} - \tilde{u}^{(k)}\right) \end{array} \tag{2.13}$$

Actualisation de l'estimation :

$$\hat{y}^{(k)} = \hat{y}^{(k-1)} + \tilde{u}^{(k)} + \tilde{b}^{(k)} \tag{2.14}$$

(3) Itérer (2) jusqu'à une condition d'arrêt à définir.

Ainsi, à la $k^{\text{ème}}$ étape, on considère le vecteur $y - \hat{y}^{(k-1)}$ sur lequel on applique la régression isotonique pour obtenir $\tilde{u}^{(k)}$. On applique ensuite la régression antitonique sur le résidu $y - \hat{y}^{(k-1)} - \tilde{u}^{(k)}$ ce qui donne $\tilde{b}^{(k)}$. On actualise ensuite la partie croissante (resp. décroissante) en l'ajoutant à l'estimation croissante (resp. décroissante) que l'on avait à l'issue de l'étape $k-1$. Ainsi,

$$\hat{y}^{(k)} = \left(\sum_{j=1}^{k} \tilde{u}^{(j)}\right) + \left(\sum_{j=1}^{k} \tilde{b}^{(j)}\right).$$

A nouveau, comme on choisit de commencer chaque cycle par une régression isotonique, la moyenne de y est portée par l'estimation croissante, la partie décroissante étant de moyenne

nulle. Plus précisément, cette moyenne est portée par $\tilde{u}^{(1)} = \text{iso}(y)$, les $\tilde{u}^{(k)}$ pour $k \geq 2$ étant de moyenne nulle.

Comme pour l'algorithme précédent, on vérifiera en Annexe C l'équation

$$\Delta\left(\tilde{u}^{(k)}\right) \circ \Delta\left(\tilde{b}^{(k)}\right) = 0 \tag{2.15}$$

qui assure l'identifiabilité des termes dans (2.14).

Le fait remarquable est que les deux algorithmes coïncident. Nous le démontrons dans la Section 2.2.1. L'égalité des deux algorithmes permet en particulier de déterminer les limites individuelles des termes de la somme et ainsi de compléter le Théorème 2.1 (cf. Théorème 2.3, Section 2.2.2). La Section 2.3 est ensuite consacrée à l'étude de la consistance de l'estimateur.

2.2 Propriétés à n fixé

2.2.1 Backfitting = réduction de biais

Pour simplifier, nous supposons dans cette section les données y centrées. Pour tout k, les vecteurs $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$ obtenus par application de l'algorithme 3 ainsi que les vecteurs $\tilde{u}^{(k)}$ et $\tilde{b}^{(k)}$ obtenus par l'algorithme 4 sont donc également tous de moyenne nulle. Nous nous plaçons ainsi dans l'hyperplan de $\mathbb{R}^n$ noté $\mathcal{E}$ défini par :

$$\mathcal{E} = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 0 \right\}.$$

Nous conservons cependant les notations $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ pour désigner la restriction de ces cônes à $\mathcal{E}$.

La démonstration de l'égalité des deux méthodes repose sur le Lemme technique suivant :

Lemme 2.3
Soit $y \in \mathcal{E}$, alors

$$\forall u \in \mathcal{C}_n^+ : \text{iso}(y+u) - \text{iso}(y) \in \mathcal{C}_n^+ \tag{2.16}$$

et

$$\forall b \in \mathcal{C}_n^- : \text{anti}(y+b) - \text{anti}(y) \in \mathcal{C}_n^-. \tag{2.17}$$

Cette propriété peut être vue comme un cas particulier d'une propriété plus générale dite "d'isotonie de la projection" qui s'énonce ainsi :

Définition 2.1 (Projection isotone)
La projection $P_{\mathcal{K}}$ sur un cône convexe fermé $\mathcal{K}$ de $\mathbb{R}^n$ est dite isotone si

$$\forall x, y \in \mathbb{R}^n : y - x \in \mathcal{K} \Rightarrow P_{\mathcal{K}}(y) - P_{\mathcal{K}}(x) \in \mathcal{K}. \tag{2.18}$$

Le terme d'isotonie traduit le fait que l'ordre induit par le cône (la relation $y - x \in \mathcal{K}$ induit une relation d'ordre) est ici conservé par projection. Cette définition est donnée dans Isac & Németh (1986). Ce papier, complété par Isac & Németh (1987) qui corrige une erreur dans le document initial, montre également qu'une condition nécessaire et suffisante sur $\mathcal{K}$ pour avoir la propriété d'isotonie (2.18) est que $\mathcal{K}$ soit mince (traduction de "thin" en anglais).

De façon générale, considérons un cône convexe fermé de $\mathbb{R}^n$. Rappelons que le cône polaire est

$$\mathcal{K}^\star = \{y \in \mathbb{R}^n : \langle x, y \rangle_n \leq 0, \forall x \in \mathcal{K}\} = \{y \in \mathbb{R}^n : P_{\mathcal{K}}(y) = 0\}$$

et qu'un vecteur de $\mathcal{K}$ est sur un rayon extremal (ou arête) de $\mathcal{K}$ si on ne peut pas l'obtenir comme combinaison convexe stricte de deux vecteurs de $\mathcal{K}$. Par ailleurs $\mathcal{K}$ est dit propre si $\mathcal{K} \cap -\mathcal{K} = \{0\}$ et il est dit générateur si $\mathcal{K} - \mathcal{K} = \mathbb{R}^n$. La définition d'un cône mince est la suivante :

Définition 2.2 (Cône mince)
On dit qu'un cône $\mathcal{K}$ de $(\mathbb{R}^n, \|.\|_n)$ convexe, fermé, propre et générateur est mince si pour tous vecteurs u et $\tilde{u}$ situés sur deux rayons extrémaux quelconques de son cône polaire $\mathcal{K}^\star$, on a $\langle u, \tilde{u} \rangle_n \leq 0$.

On sait d'après Isac & Németh (1986) que la projection sur des cônes minces a la propriété d'isotonie (2.18) et il se trouve que l'on a le résultat suivant (cf. Annexe B) :

Lemme 2.4
Les cônes $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ sont des cônes minces de $\mathcal{E}$.

Ainsi, les projections sur $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ ont la propriété d'isotonie. Par conséquent le Lemme 2.3 est vérifié et nous disposons des outils pour montrer le résultat principal de cette section.

Théorème 2.2 (Egalité des algorithmes 3 et 4)
A chaque itération, les estimations fournies par les algorithmes 3 et 4 sont égales :

$$\forall k \geq 1 \qquad \hat{u}^{(k)} = \sum_{j=1}^{k} \tilde{u}^{(j)} \qquad et \qquad \hat{b}^{(k)} = \sum_{j=1}^{k} \tilde{b}^{(j)}.$$

Preuve
Montrons tout d'abord de proche en proche que pour tout $k \geq 1$,

$$\hat{u}^{(k+1)} - \hat{u}^{(k)} \in \mathcal{C}_n^+ \qquad \text{et} \qquad \hat{b}^{(k+1)} - \hat{b}^{(k)} \in \mathcal{C}_n^-. \tag{2.19}$$

On a par définition de $\hat{u}^{(1)}$ et $\hat{u}^{(2)}$:

$$\hat{u}^{(2)} - \hat{u}^{(1)} = \text{iso}(y - \hat{b}^{(1)}) - \text{iso}(y).$$

En posant $u := -\hat{b}^{(1)} \in \mathcal{C}_n^+$, cela s'écrit :

$$\hat{u}^{(2)} - \hat{u}^{(1)} = \text{iso}(y + u) - \text{iso}(y)$$

et le Lemme 2.3 donne $\hat{u}^{(2)} - \hat{u}^{(1)} \in \mathcal{C}_n^+$. De même par définition de $\hat{b}^{(2)}$ et $\hat{b}^{(1)}$,

$$\hat{b}^{(2)} - \hat{b}^{(1)} = \text{anti}(y - \hat{u}^{(2)}) - \text{anti}(y - \hat{u}^{(1)}),$$

soit

$$\hat{b}^{(2)} - \hat{b}^{(1)} = \text{anti}\left(y - \hat{u}^{(1)} - (\hat{u}^{(2)} - \hat{u}^{(1)})\right) - \text{anti}(y - \hat{u}^{(1)}),$$

ou en posant $\tilde{y} = y - \hat{u}^{(1)}$ et $b := -(\hat{u}^{(2)} - \hat{u}^{(1)})$

$$\hat{b}^{(2)} - \hat{b}^{(1)} = \text{anti}(\tilde{y} + b) - \text{anti}(\tilde{y}).$$

D'après ce que l'on vient de trouver, $b \in \mathcal{C}_n^-$ donc, par la seconde équation du Lemme 2.3, on conclut $\hat{b}^{(2)} - \hat{b}^{(1)} \in \mathcal{C}_n^-$. C'est ensuite le même principe :

$$\hat{u}^{(3)} - \hat{u}^{(2)} = \text{iso}(y - \hat{b}^{(2)}) - \text{iso}(y - \hat{b}^{(1)}) = \text{iso}\left(y - \hat{b}^{(1)} - (\hat{b}^{(2)} - \hat{b}^{(1)})\right) - \text{iso}(y - \hat{b}^{(1)}).$$

On pose $\tilde{y} := y - \hat{b}^{(1)}$ et comme on vient de voir $\hat{b}^{(2)} - \hat{b}^{(1)} \in \mathcal{C}_n^-$, le Lemme 2.3 donne $\hat{u}^{(3)} - \hat{u}^{(2)} \in \mathcal{C}_n^+$. De proche en proche on obtient (2.19).

Montrons maintenant par récurrence le résultat voulu. Il est clair qu'à la première étape ($k = 1$) des algorithmes, les estimations coïncident. Supposons maintenant qu'à une étape $k \geq 1$, nous avons

$$\hat{y}^{(k)} = \hat{u}^{(k)} + \hat{b}^{(k)} \qquad \text{avec} \qquad \hat{u}^{(k)} = \sum_{j=1}^{k} \tilde{u}^{(j)} \text{ et } \hat{b}^{(k)} = \sum_{j=1}^{k} \tilde{b}^{(j)}.$$

1. Montrons qu'alors $\hat{u}^{(k+1)} = \sum_{j=1}^{k+1} \tilde{u}^{(j)}$.
   Pour cela, montrons tout d'abord

$$\|y - \hat{b}^{(k)} - \hat{u}^{(k+1)}\|_n = \|y - \hat{y}^{(k)} - \tilde{u}^{(k+1)}\|_n. \tag{2.20}$$

   Comme $\hat{u}^{(k+1)}$ est la meilleure approximation de $y - \hat{b}^{(k)}$ parmi les séquences croissantes et que $\hat{u}^{(k)} + \tilde{u}^{(k+1)}$ est croissante comme somme de séquences croissantes, on a

$$\|y - \hat{b}^{(k)} - \hat{u}^{(k+1)}\|_n \leq \|y - \hat{b}^{(k)} - (\hat{u}^{(k)} + \tilde{u}^{(k+1)})\|_n = \|y - \hat{y}^{(k)} - \tilde{u}^{(k+1)}\|_n.$$

   $\tilde{u}^{(k+1)}$ est la meilleure approximation croissante de $y - \hat{y}^{(k)}$. Elle est en particulier meilleure que $\hat{u}^{(k+1)} - \hat{u}^{(k)}$ dont on vient de voir qu'elle est croissante (équation 2.19). Ainsi

$$\|y - \hat{y}^{(k)} - \tilde{u}^{(k+1)}\|_n \leq \|y - \hat{y}^{(k)} - (\hat{u}^{(k+1)} - \hat{u}^{(k)})\|_n = \|y - \hat{b}^{(k)} - \hat{u}^{(k+1)}\|_n.$$

   Les deux dernières inéquations donnent bien (2.20). Dès lors, on a

$$\|y - \hat{y}^{(k)} - \tilde{u}^{(k+1)}\|_n = \|y - \hat{b}^{(k)} - \hat{u}^{(k+1)}\|_n = \|y - \hat{y}^{(k)} - (\hat{u}^{(k+1)} - \hat{u}^{(k)})\|_n$$

   ce qui montre que $\tilde{u}^{(k+1)}$ et $\hat{u}^{(k+1)} - \hat{u}^{(k)}$ réalisent tous deux le minimum de distance à $y - \hat{y}^{(k)}$. On peut donc les confondre

$$\tilde{u}^{(k+1)} = \hat{u}^{(k+1)} - \hat{u}^{(k)} \Leftrightarrow \hat{u}^{(k+1)} = \hat{u}^{(k)} + \tilde{u}^{(k+1)}$$

   et l'hypothèse de récurrence donne

$$\hat{u}^{(k+1)} = \sum_{j=1}^{k+1} \tilde{u}^{(j)}.$$

2. Montrer qu'alors on a aussi $\hat{b}^{(k+1)} = \sum_{j=1}^{k+1} \tilde{b}^{(j)}$ procède de la même idée.
   On montre tout d'abord que

$$\|y - \hat{u}^{(k+1)} - \hat{b}^{(k+1)}\|_n = \|y - \hat{y}^{(k)} - \tilde{u}^{(k+1)} - \tilde{b}^{(k+1)}\|_n. \tag{2.21}$$

   $\hat{b}^{(k+1)}$ étant le projeté de $y - \hat{u}^{(k+1)}$ sur $\mathcal{C}_n^-$, il est clair que

$$\|y - \hat{u}^{(k+1)} - \hat{b}^{(k+1)}\|_n \leq \|y - \hat{u}^{(k+1)} - (\hat{b}^{(k)} + \tilde{b}^{(k+1)})\|_n.$$

   Ainsi, avec $\hat{u}^{(k+1)} = \hat{u}^{(k)} + \tilde{u}^{(k+1)}$, il vient

$$\|y - \hat{u}^{(k+1)} - \hat{b}^{(k+1)}\|_n \leq \|y - \hat{y}^{(k)} - \tilde{u}^{(k+1)} - \tilde{b}^{(k+1)}\|_n.$$

De plus, $\tilde{b}^{(k+1)}$ étant par définition la meilleure approximation décroissante de $y - \hat{y}^{(k)} - \tilde{u}^{(k+1)}$, elle est en particulier meilleure que $\hat{b}^{(k+1)} - \hat{b}^{(k)}$ qui est décroissante (cf. équation 2.19), d'où

$$\begin{aligned}\|y - \hat{y}^{(k)} - \tilde{u}^{(k+1)} - \tilde{b}^{(k+1)}\|_n &\leq \|y - \hat{y}^{(k)} - \tilde{u}^{(k+1)} - (\hat{b}^{(k+1)} - \hat{b}^{(k)})\|_n \\ &\leq \|y - \hat{u}^{(k)} - \tilde{u}^{(k+1)} - \hat{b}^{(k+1)}\|_n \\ &\leq \|y - \hat{u}^{(k+1)} - \hat{b}^{(k+1)}\|_n\end{aligned}$$

et l'égalité (2.21) est justifiée par la double inégalité. Il vient alors

$$\begin{aligned}\|y - \hat{u}^{(k+1)} - \hat{b}^{(k+1)}\|_n &= \|y - \hat{y}^{(k)} - \tilde{u}^{(k+1)} - \tilde{b}^{(k+1)}\|_n \\ &= \|y - \hat{y}^{(k)} - (\hat{u}^{(k+1)} - \hat{u}^{(k)} - \tilde{b}^{(k+1)})\|_n \\ &= \|y - \hat{b}^{(k)} - \hat{u}^{(k+1)} - \tilde{b}^{(k+1)}\|_n \\ &= \|y - \hat{u}^{(k+1)} - (\hat{b}^{(k)} + \tilde{b}^{(k+1)})\|_n\end{aligned}$$

ce qui confond $\hat{b}^{(k)} + \tilde{b}^{(k+1)}$ avec $\hat{b}^{(k+1)}$ le projeté de $y - \hat{u}^{(k+1)}$ sur $\mathcal{C}_n^-$. D'après l'hypothèse de récurrence, on a alors

$$\hat{b}^{(k+1)} = \hat{b}^{(k)} + \tilde{b}^{(k+1)} = \sum_{j=1}^{k+1} \tilde{b}^{(j)}$$

ce qui achève la preuve. □

Il est intéressant de donner une interprétation graphique en complément de celle qui a été faite en figure 2.3 page 39. La figure 2.5 illustre la construction de $\tilde{u}^{(2)}$ comme projeté de $y - \hat{y}^{(1)}$ sur $\mathcal{C}_n^+$ et l'égalité $\tilde{u}^{(2)} = \hat{u}^{(2)} - \hat{u}^{(1)}$.

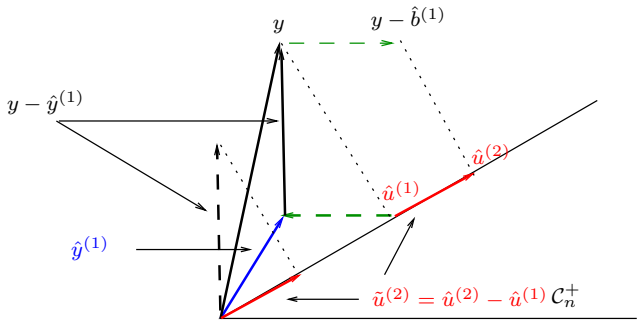


Fig. 2.5 – Illustration de l'égalité des deux méthodes.

2.2.2 Convergence des termes individuels de la somme

Nous avons montré en Théorème 2.1, Section 2.1 page 39, la convergence de $\hat{y}^{(k)}$ vers y lorsque k tend vers l'infini. La preuve s'appuie sur l'analyse de l'algorithme de Von Neumann faite par Bauschke & Borwein (1994). Comme nous l'avons mentionné en remarque page 41, un corollaire de leur résultat principal assure également l'existence d'une limite pour chacun des termes $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$ de la somme. Néanmoins, leur corollaire ne permet pas de caractériser ces limites, ce que nous allons pouvoir faire ici grâce à l'égalité des deux algorithmes.

Le Théorème 2.3 page 48 établit ainsi que $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$ convergent respectivement vers les termes de la décomposition à variation minimale du vecteur y, analogue discret de la décomposition à variation minimale de Jordan. Nous notons ces termes $u_y^\star$ et $b_y^\star$.

Avant d'énoncer le théorème, nous en donnons une illustration sur un exemple en figure 2.6. La partie gauche représente un vecteur de données y et la décomposition $y = u_y^\star + b_y^\star$. S'agissant de l'analogue pour un vecteur de la décomposition à variation minimale pour une fonction, on a

$$y_{i+1} - y_i > 0 \Rightarrow u^\star_{y,i+1} - u^\star_{y,i} = y_{i+1} - y_i \text{ et } b^\star_{y,i+1} - b^\star_{y,i} = 0$$

et inversement,

$$y_{i+1} - y_i < 0 \Rightarrow u^\star_{y,i+1} - u^\star_{y,i} = 0 \text{ et } b^\star_{y,i+1} - b^\star_{y,i} = y_{i+1} - y_i$$

la moyenne de y étant portée par $u_y^\star$ alors que $b_y^\star$ est de moyenne nulle. On a ainsi

$$\Delta(u_y^\star) \circ \Delta(b_y^\star) = 0.$$

La partie droite de la figure illustre le fait que, le nombre d'itérations k augmentant, les coordonnées des vecteurs $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$ se rapprochent de celles de $u_y^\star$ et de $b_y^\star$.

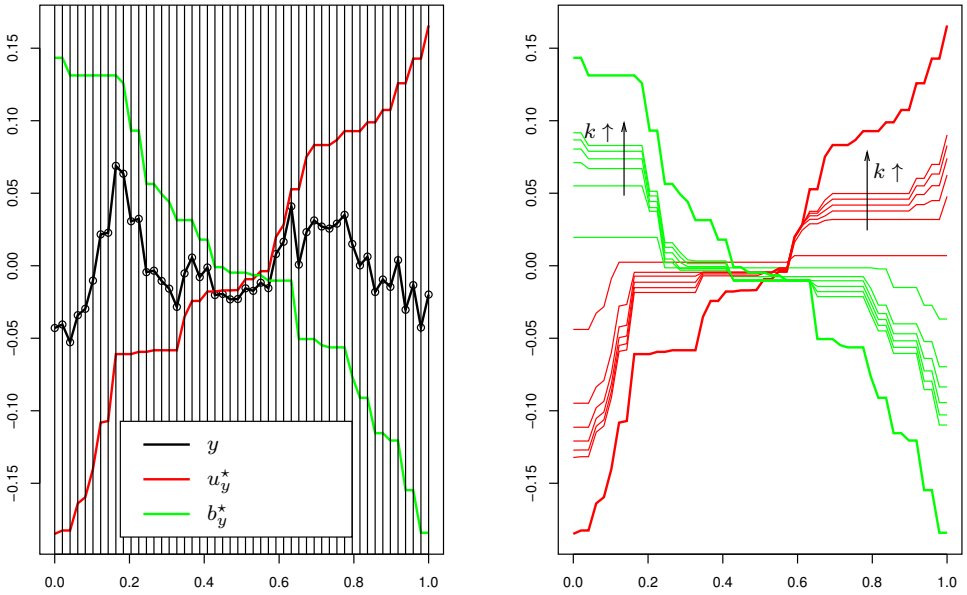


Fig. 2.6 – Convergence de $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$ vers $u_y^\star$ et $b_y^\star$.

Pour simplifier, nous considérons à nouveau les données y centrées. Les quantités produites par les algorithmes ainsi que les termes $u_y^\star$ et $b_y^\star$ sont donc tous centrés également. Cela nous permet de travailler dans $\mathcal{E}$, le sous-espace de $\mathbb{R}^n$ des vecteurs centrés, et d'introduire la norme V qui mesure les variations d'un vecteur :

$$\begin{array}{rccl} \mathrm{V}: & \mathcal{E} & \to & \mathbb{R}^+ \\ & y & \mapsto & \mathrm{V}(y) = \sum_{i=1}^{n-1} |y_{i+1} - y_i| \end{array}$$

Remarque
L'application V n'est pas une norme dans $\mathbb{R}^n$ car la propriété de séparation n'est pas vérifiée en général. Elle le devient par contre lorsqu'on se restreint à $\mathcal{E}$.

En général, pour un vecteur $z \in \mathcal{E}$ et une décomposition $z = u_z + b_z$ dans $\mathcal{E}$ de ce vecteur, on a l'inégalité usuelle pour une norme : $\mathrm{V}(z) \leq \mathrm{V}(u_z) + \mathrm{V}(b_z)$. Il est clair cependant que, si l'on a $\Delta(u_z) \circ \Delta(b_z) = 0$, alors $\mathrm{V}(z) = \mathrm{V}(u_z) + \mathrm{V}(b_z)$. La réciproque est vraie. En effet, pour tout i,

$$|z_{i+1} - z_i| \leq |u_{z,i+1} - u_{z,i}| + |b_{z,i+1} - b_{z,i}| = (u_{z,i+1} - u_{z,i}) - (b_{z,i} - b_{z,i+1})$$

L'égalité $\mathrm{V}(z) = \mathrm{V}(u_z) + \mathrm{V}(b_z)$ ne peut donc avoir lieu qu'à la condition que pour tout i, on ait

$$|z_{i+1} - z_i| = (u_{z,i+1} - u_{z,i}) - (b_{z,i} - b_{z,i+1})\,. \tag{2.22}$$

Or

$$z_{i+1} > z_i \Rightarrow |z_{i+1} - z_i| = z_{i+1} - z_i = (u_{z,i+1} - u_{z,i}) + (b_{z,i+1} - b_{z,i})$$

donc (2.22) implique $b_{z,i+1} = b_{z,i}$. De même, on montre

$$z_{i+1} < z_i \Rightarrow u_{z,i+1} = u_{z,i}$$

et

$$z_{i+1} = z_i \Rightarrow u_{z,i+1} - u_{z,i} = b_{z,i+1} - b_{z,i} = 0.$$

Finalement $\mathrm{V}(z) = \mathrm{V}(u_z) + \mathrm{V}(b_z) \Rightarrow \Delta(u_z) \circ \Delta(b_z) = 0$ et la réciproque est prouvée.

Les conditions d'identifiabilité impliquent donc que pour tout $k \geq 1$

$$\mathrm{V}(\hat{y}^{(k)}) = \mathrm{V}(\hat{u}^{(k)}) + \mathrm{V}(\hat{b}^{(k)}) \qquad \text{et} \qquad \mathrm{V}(\tilde{u}^{(k)} + \tilde{b}^{(k)}) = \mathrm{V}(\tilde{u}^{(k)}) + \mathrm{V}(\tilde{b}^{(k)}). \tag{2.23}$$

Par ailleurs, il suffit de montrer que $u^\infty = \lim_{k\to\infty} \hat{u}^{(k)}$ et $b^\infty = \lim_{k\to\infty} \hat{b}^{(k)}$ vérifient $\mathrm{V}(u^\infty) + \mathrm{V}(b^\infty) = \mathrm{V}(y)$ pour avoir $\Delta(u^\infty) \circ \Delta(b^\infty) = 0$ donc $u^\infty = u_y^\star$ et $b^\infty = b_y^\star$. C'est l'idée de la démonstration du théorème.

Théorème 2.3 (Convergence des termes individuels)
Les séquences $\left(\hat{u}^{(k)}\right)_{k\geq 1}$ *et* $\left(\hat{b}^{(k)}\right)_{k\geq 1}$ *convergent vers les termes de la décomposition à variation minimale de* y*. Autrement dit,*

$$\lim_{k\to\infty} \hat{u}^{(k)} = u_y^\star \qquad \textit{et} \qquad \lim_{k\to\infty} \hat{b}^{(k)} = b_y^\star$$

où $u_y^\star$ *et* $b_y^\star$ *sont tels que* $y = u_y^\star + b_y^\star$ *et* $\mathrm{V}(y) = \mathrm{V}(u_y^\star) + \mathrm{V}(b_y^\star)$*.*

Preuve
Les résultats de Bauschke & Borwein (1994) que nous invoquons pour justifier l'existence des limites individuelles sont donnés dans le cadre d'espaces de Hilbert pour des projections sur des convexes fermés en respect d'une certaine norme issue d'un produit scalaire. Ils sont directement transposables à notre situation pour $\mathbb{R}^n$ et la norme $\|.\|_n$. Ainsi, il existe u^∞ et b^∞ tels que $\lim_{k\to\infty} \|\hat{u}^{(k)} - u^\infty\|_n = 0$ et $\lim_{k\to\infty} \|\hat{b}^{(k)} - b^\infty\|_n = 0$. Cela implique que pour tout i,

$$\lim_{k\to\infty} \hat{u}_i^{(k)} = u_i^\infty \qquad \text{et} \qquad \lim_{k\to\infty} \hat{b}_i^{(k)} = b_i^\infty$$

donc par continuité de la norme V(.)

$$\lim_{k\to\infty} \mathrm{V}(\hat{u}^{(k)}) = \mathrm{V}(u^\infty) \qquad \text{et} \qquad \lim_{k\to\infty} \mathrm{V}(\hat{b}^{(k)}) = \mathrm{V}(b^\infty). \tag{2.24}$$

On a bien sûr l'équivalent avec $\hat{y}^{(k)} \to y$. Ainsi,

$$\begin{aligned}
\mathrm{V}(y) &= \mathrm{V}\left(\lim_{k\to\infty}\{\hat{u}^{(k)} + \hat{b}^{(k)}\}\right) \\
&= \lim_{k\to\infty} \mathrm{V}\left(\hat{u}^{(k)} + \hat{b}^{(k)}\right) \\
&= \lim_{k\to\infty} \left\{\mathrm{V}(\hat{u}^{(k)}) + \mathrm{V}(\hat{b}^{(k)})\right\} && \text{d'après (2.23)} \\
&= \lim_{k\to\infty} \mathrm{V}(\hat{u}^{(k)}) + \lim_{k\to\infty} \mathrm{V}(\hat{b}^{(k)}) && \text{d'après (2.24)} \\
&= \mathrm{V}\left(\lim_{k\to\infty} \hat{u}^{(k)}\right) + \mathrm{V}\left(\lim_{k\to\infty} \hat{b}^{(k)}\right) \\
&= \mathrm{V}(u^\infty) + \mathrm{V}(b^\infty).
\end{aligned}$$

D'après l'assertion donnée juste avant l'énoncé du théorème, on déduit que (u^∞, b^∞) est la décomposition à variation minimale $(u_y^\star, b_y^\star)$ de y, ce qui est le résultat voulu. □

Remarque
En particulier, on a convergence des séries $\sum \tilde{u}^{(k)}$ et $\sum \tilde{b}^{(k)}$ avec

$$u_y^\star = \sum_{k=1}^{\infty} \tilde{u}^{(k)} \qquad \text{et} \qquad b_y^\star = \sum_{k=1}^{\infty} \tilde{b}^{(k)}.$$

2.2.3 Conclusions

Nous terminons cette partie sur la description de notre méthode à n fixé en donnant quelques conclusions et perspectives. La figure 2.7 illustre de manière synthétique la plupart des résultats qui ont été établis jusqu'ici ainsi que des résultats complémentaires qui sont donnés en Annexe D page 117.

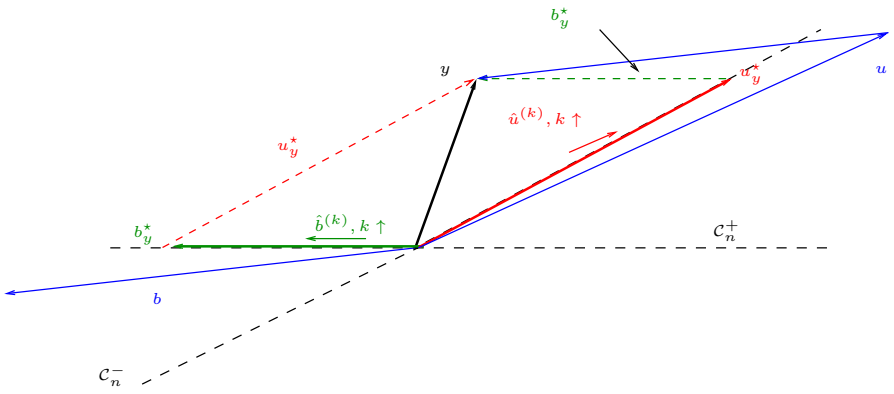


Fig. 2.7 – Illustration des propriétés.

Lorsque k augmente, les termes $\hat{u}^{(k)}$ et $\hat{b}^{(k)}$ se rapprochent de la décomposition $u^\star$ et $b^\star$ de y. Nous montrons que les écarts entre $u^\star$ et $\hat{u}^{(k)}$ ainsi qu'entre $b^\star$ et $\hat{b}^{(k)}$ tendent vers 0 en décroissant, pour la norme V(.) comme pour la norme $\|.\|_n$ (cf. Propositions D.3 et D.5). Nous avons la même chose pour la convergence de $\hat{y}^{(k)}$ vers y.

Il est intéressant de remarquer que pour toute décomposition $y = u + b$, on a

$$\left\{ \begin{array}{rcl} u & = & \mathrm{iso}(y-b) \\ b & = & \mathrm{anti}(y-u) \end{array} \right. . \tag{2.25}$$

On peut voir ce système d'équations comme l'analogue des équations dites normales présentées pour un modèle additif bivarié en Section 1.2.3 page 31

$$\left\{ \begin{array}{rcl} r_1 & = & S_1(y-r_2) \\ r_2 & = & S_2(y-r_1) \end{array} \right. \tag{2.26}$$

dont l'existence et l'unicité des solutions sont liées aux propriétés des lisseurs mis en jeu. Dans ce cadre bivarié, l'algorithme backfitting conduit à des estimateurs $\hat{r}_1^{(k)}$ et $\hat{r}_2^{(k)}$ qui, toujours selon les propriétés de S_1 et S_2, peuvent converger individuellement vers r_1 et r_2 ou bien dont la somme peut converger vers $r_1 + r_2$.

En Proposition D.6 page 123, nous montrons que pour toute décomposition $y = u+b$ avec $u \in \mathcal{C}_n^+$ et $b \in \mathcal{C}_n^-$ (représentée en bleu sur la figure 2.7), les suites

$$\left(\|u - \hat{u}^{(k)}\|_n\right)_{k\geq 1} \qquad \text{et} \qquad \left(\|b - \hat{b}^{(k)}\|_n\right)_{k\geq 1}$$

convergent également en décroissant. Ainsi, dans notre situation, tout vecteur y peut s'écrire sous la forme $y = u + b$, et n'importe quel couple (u, b) de la sorte est solution des équations normales (2.25). Toutes les suites de la forme $\left(\|\hat{u}^{(k)} - u\|_n\right)$ et $\left(\|\hat{b}^{(k)} - b\|_n\right)$ décroissent et donc convergent. En revanche seules $\left(\|\hat{u}^{(k)} - u^\star\|_n\right)$ et $\left(\|\hat{b}^{(k)} - b^\star\|_n\right)$ convergent vers 0.

Complétons la comparaison avec le modèle additif bivarié et revenons sur la remarque de la page 41 concernant l'arrêt de méthode I.I.R. Dans le cadre multivarié usuel, on arrête l'algorithme backfitting lorsque les estimations individuelles des termes de la somme ne varient plus sensiblement d'une itération à l'autre. Dans ce cadre, de vraies données y ne peuvent avoir une structure précisément additive, à moins que l'on se retrouve dans la situation de la figure 2.8 où les points sont en effet exactement distribués sur la surface représentant une fonction additive.

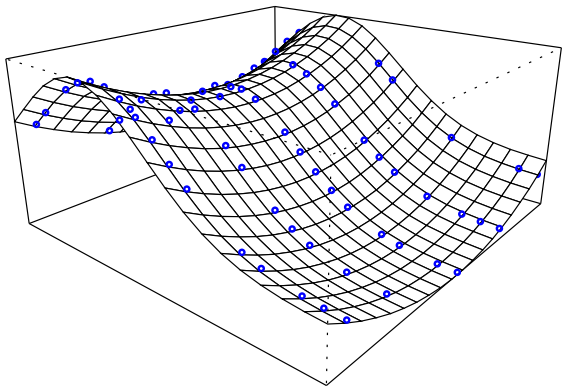


Fig. 2.8 – Données ayant une structure additive.

Dans les cas réels, on peut supposer que les observations ne sont pas trop éloignées de cette structure et les estimations produites par backfitting, quand elles auront capturé la partie additive

des données, ne varieront plus sensiblement d'une itération à l'autre. Dans notre situation au contraire, les données ont toujours exactement la structure additive $y = u + b$. Il n'est donc pas contradictoire qu'augmenter le nombre d'itérations conduise à leur interpolation.

2.3 Consistance

Dans la section précédente, nous avons décrit l'estimateur de régression isotonique itérée lorsque n est fixé. Nous étudions ici certaines de ses propriétés lorsque le nombre de points tend vers l'infini. Nous commençons par fixer les notations et décrire notre démarche.

2.3.1 Introduction

Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon i.i.d. avec

$$Y_i = r(X_i) + \varepsilon_i \tag{2.27}$$

où $r : [0,1] \to [0,1]$ est une fonction inconnue à variation bornée. On suppose $\mathbb{E}\left[Y^2\right] < \infty$ et pour tout i, $\mathbb{E}\left[\varepsilon_i | X_i\right] = 0$.

Les observations rangées dans l'ordre des X_i sont notées $\left(X_{(i)}, Y_{(i)}\right)$. On considère que la fonction de répartition de X est continue donc on a $X_{(1)} < \ldots < X_{(n)}$ et $Y_{(1)}, \ldots, Y_{(n)}$ les observations correspondantes de Y.

On applique la régression isotonique selon l'algorithme backfitting. On construit ainsi une suite d'estimateurs $(\hat{u}_n^{(k)})_{k\geq 0}$ et $(\hat{b}_n^{(k)})_{k\geq 0}$ définis itérativement par $\hat{u}_n^{(0)} = \hat{b}_n^{(0)} = 0$ et pour $k \geq 1$

$$\hat{u}_n^{(k)} = \underset{u\in\mathcal{C}_n^+}{\operatorname{argmin}} \|Y - \hat{b}_n^{(k-1)} - u\|_n^2 = \underset{u\in\mathcal{C}_n^+}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(Y_{(i)} - \hat{b}_n^{(k-1)}(X_{(i)}) - u(X_{(i)})\right)^2$$
$$\hat{b}_n^{(k)} = \underset{b\in\mathcal{C}_n^-}{\operatorname{argmin}} \|Y - \hat{u}_n^{(k)} - b\|_n^2 = \underset{b\in\mathcal{C}_n^-}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(Y_{(i)} - \hat{u}_n^{(k)}(X_{(i)}) - b(X_{(i)})\right)^2.$$

Remarque
Avant de poursuivre, il convient de préciser les notations. Pour fixer les idées, prenons $k = 1$ dans les définitions ci-dessus. Dans ce cas, $\hat{b}_n^{(0)}$ est le vecteur nul et $\hat{u}_n^{(1)}$ est un vecteur de taille n dont les coordonnées forment une séquence croissante. A partir de ce vecteur, nous pouvons définir une fonction constante par morceaux $\hat{u}_n^{(1)}$ de $[0,1]$ dans $\mathbb{R}$ comme suit : si $0 \leq x < X_{(2)}$, $\hat{u}_n^{(1)}(x)$ correspond à la première coordonnée du vecteur $\hat{u}_n^{(1)}$; si $X_{(2)} \leq x < X_{(3)}$, $\hat{u}_n^{(1)}(x)$ correspond à la deuxième coordonnée du vecteur $\hat{u}_n^{(1)}$; et ainsi de suite, avec finalement pour $X_{(n)} \leq x \leq 1$, $\hat{u}_n^{(1)}(x)$ qui correspond à la dernière coordonnée du vecteur $\hat{u}_n^{(1)}$. Ceci est illustré figure 2.9. Réciproquement, une fonction $f : [0,1] \to \mathbb{R}$ étant donnée, on lui associe le vecteur de taille n dont la coordonnée i est définie par $f_i = f(X_{(i)})$. Dans toute la suite, cette correspondance permettra les allers-retours de la norme empirique $\|.\|_n$ à la norme quadratique $\|.\|$ sans ambiguïté. Enfin, avec cette convention et puisque la somme est symétrique en ses termes, nous pouvons encore

définir $(\hat{u}_n^{(k)})_{k\geq 0}$ et $(\hat{b}_n^{(k)})_{k\geq 0}$ par

$$\hat{u}_n^{(k)} = \underset{u\in\mathcal{C}_n^+}{\operatorname{argmin}} \|Y - \hat{b}_n^{(k-1)} - u\|_n^2 = \underset{u\in\mathcal{C}_n^+}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \left(Y_i - \hat{b}_n^{(k-1)}(X_i) - u(X_i)\right)^2$$
$$\hat{b}_n^{(k)} = \underset{b\in\mathcal{C}_n^-}{\operatorname{argmin}} \|Y - \hat{u}_n^{(k)} - b\|_n^2 = \underset{b\in\mathcal{C}_n^-}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \left(Y_i - \hat{u}_n^{(k)}(X_i) - b(X_i)\right)^2,$$

ce que nous ferons dans la suite pour alléger (un peu) les écritures.

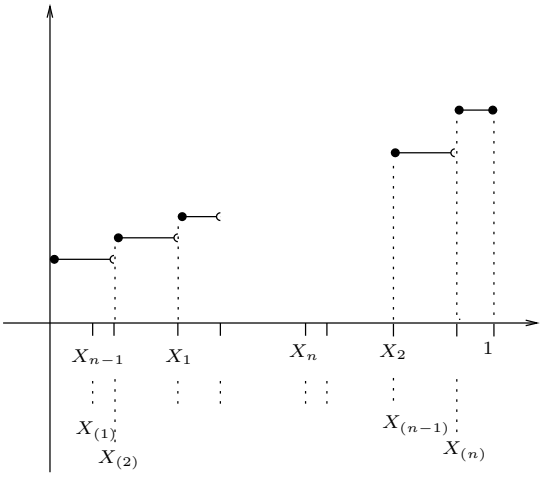


Fig. 2.9 – Prolongement des estimateurs sur $[0,1]$.

L'estimateur à la $k^{\text{ème}}$ itération est obtenu en sommant partie isotonique et partie antitonique :

$$\hat{r}_n^{(k)} = \hat{u}_n^{(k)} + \hat{b}_n^{(k)}.$$

Cette expression donne le vecteur de $\mathbb{R}^n$ des valeurs ajustées et en prolongeant par des segments horizontaux, on définit l'estimateur sur $[0,1]$.

On souhaite mesurer la proximité à la fonction de régression r. On considère pour cela la norme associée à la moyenne quadratique :

$$\|\hat{r}_n^{(k)} - r\|^2 = \int_0^1 \left(\hat{r}_n^{(k)}(x) - r(x)\right)^2 \mu(dx).$$

Cette quantité est aléatoire. Le résultat principal que l'on obtient concerne son espérance. Plus précisément, l'objet de cette partie est de montrer le théorème suivant :

Théorème 2.4 (Consistance de la régression isotonique itérée)
Sous l'hypothèse que le bruit est borné, il existe une suite d'entiers (k_n) croissante telle que

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{r}_n^{(k_n)} - r\|^2\right] = 0.$$

Pour ce faire, on s'appuie sur l'inégalité

$$\|\hat{r}_n^{(k)} - r\| \leq \|\hat{r}_n^{(k)} - r_n^{(k)}\| + \|r_n^{(k)} - r^{(k)}\| + \|r^{(k)} - r\| \tag{2.28}$$

ou

$$\|\hat{r}_n^{(k)} - r\|^2 \leq 3\left\{\|\hat{r}_n^{(k)} - r_n^{(k)}\|^2 + \|r_n^{(k)} - r^{(k)}\|^2 + \|r^{(k)} - r\|^2\right\} \tag{2.29}$$

dont on cherche à contrôler chacun des termes de droite.

Pour ce qui est des notations, précisons que $r_n^{(k)} = u_n^{(k)} + b_n^{(k)}$ est associé à l'application de k itérations de l'algorithme sur les n points $(X_i, r(X_i))$ de la courbe de régression alors que $r^{(k)} = u^{(k)} + b^{(k)}$ note le résultat de l'application de l'algorithme à la fonction de régression elle-même. Ainsi, par exemple

$$u_n^{(1)} = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|r - u\|_n^2 = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (r(X_i) - u(X_i))^2$$

et

$$b_n^{(1)} = \underset{b \in \mathcal{C}_n^-}{\operatorname{argmin}} \|r - u_n^{(1)} - b\|_n^2 = \underset{u \in \mathcal{C}_n^-}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(r(X_i) - u_n^{(1)}(X_i) - b(X_i)\right)^2$$

alors que

$$u^{(1)} = \underset{u \in \mathcal{C}^+}{\operatorname{argmin}} \|r - u\|^2 = \underset{u \in \mathcal{C}^+}{\operatorname{argmin}} \int_0^1 (r(x) - u(x))^2 \, \mu(dx)$$

et

$$b^{(1)} = \underset{b \in \mathcal{C}^-}{\operatorname{argmin}} \|r - u^{(1)} - b\|^2 = \underset{b \in \mathcal{C}^-}{\operatorname{argmin}} \int_0^1 \left(r(x) - u^{(1)}(x) - b(x)\right)^2 \mu(dx).$$

Les deux dernières écritures supposent l'existence et l'unicité d'une solution aux problèmes de minimisation considérés. Les cônes $\mathcal{C}^+$ et $\mathcal{C}^-$ sont clairement des convexes de $L^2(\mu)$. Nous n'avons pas trouvé dans la littérature la preuve de leur fermeture pour la norme $\|.\|$. Nous en donnons une démonstration en Annexe E page 125 [1].

Les inégalités (2.28) et (2.29) correspondent à une décomposition biais-variance de l'estimateur obtenu après k itérations. Le terme $\|\hat{r}_n^{(k)} - r_n^{(k)}\|$ représente l'écart entre l'estimateur et l'application de l'algorithme aux données qui seraient non bruitées : on peut donc voir ce terme comme un terme de variance. Le second terme $\|r_n^{(k)} - r^{(k)}\|$ mesure l'écart entre le résultat obtenu sur les points de la courbe et le résultat obtenu sur la courbe elle-même : ne faisant pas intervenir les données Y, on désigne ce terme par le biais. Le nombre d'itérations k étant fixé, le contrôle de ces deux quantités permet de montrer que, lorsque n tend vers l'infini, $\hat{r}_n^{(k)}$ converge vers $r^{(k)}$, résultat de k itérations de la méthode sur la fonction r.

Cependant, k étant fixé, la fonction $r^{(k)}$ ne coïncide pas en général avec la fonction r. Nous appelons ainsi l'écart $\|r^{(k)} - r\|$ erreur d'approximation alors que $\|\hat{r}_n^{(k)} - r^{(k)}\|$ est appelée erreur d'estimation. Les termes en présence sont schématisés dans la formulation suivante :

1. Une autre façon de s'assurer de cette existence et unicité est d'appliquer le Théorème 1.1 donné par Anevski & Soulier (2011).

$$\|\hat{r}_n^{(k)} - r\| \leq \underbrace{\|\hat{r}_n^{(k)} - r^{(k)}\|}_{\text{Estimation}} + \underbrace{\|r^{(k)} - r\|}_{\text{Approximation}}$$

$$\leq$$

$$\|\hat{r}_n^{(k)} - r_n^{(k)}\| + \|r_n^{(k)} - r^{(k)}\|$$

$$\downarrow \qquad \downarrow$$

$$\text{Variance} + \text{Biais}$$

La figure 2.10 illustre notre démarche. Pour tout k fixé, nous montrons que l'erreur d'estimation tend vers 0 avec n. Plus précisément, nous montrons

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{r}_n^{(k)} - r^{(k)}\|^2\right] = 0.$$

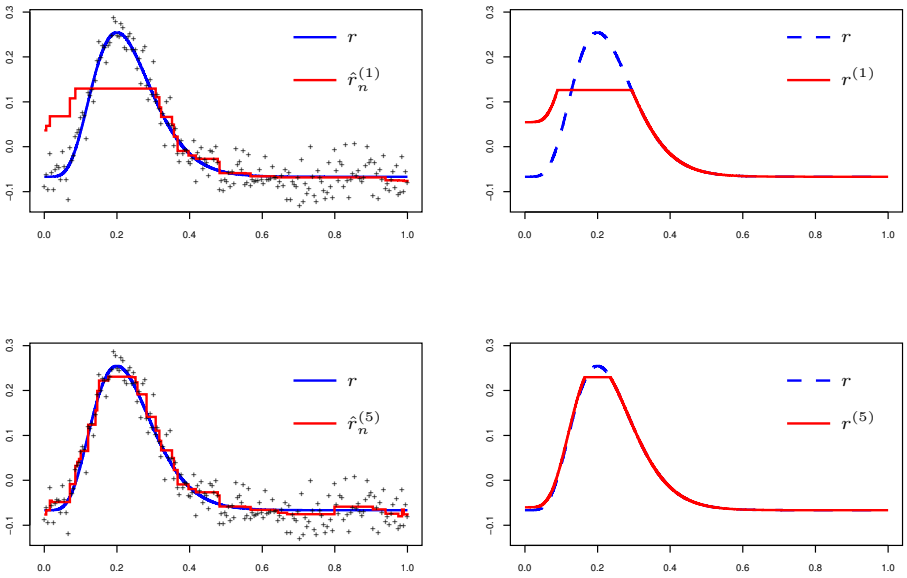


Fig. 2.10 – Erreur d'approximation et erreur d'estimation.

Pour ce faire, on analyse tout d'abord la première demi-étape de l'algorithme en prouvant que, lorsque la fonction de régression n'est pas isotonique, l'application de la régression isotonique sur des données bruitées fournit un estimateur qui converge vers la fonction isotonique la plus proche de la fonction de régression. Ce point est un résultat en lui-même car à notre connaissance, la question n'avait pas été envisagée jusqu'à présent, les travaux existants étudiant uniquement les propriétés de convergence de la régression isotonique pour une fonction de régression elle-même isotonique. Le résultat est énoncé en Théorème 2.5 page 2.5 et sa démonstration fait l'objet de la Section 2.3.2.

En Section 2.3.3, nous voyons ensuite comment les résultats se propagent au fil de l'algorithme, permettant le contrôle de l'erreur d'estimation. La figure 2.10 illustre également le fait que lorsque le nombre d'itérations k augmente, la fonction $r^{(k)}$ se rapproche de r. Ce résultat s'obtient en calquant la preuve du Théorème 2.1 qui démontre pour n fixé, l'interpolation des données quand k tend vers l'infini. En effet, nous nous appuyons dans cette preuve sur l'analyse de Bauschke

& Borwein (1994) de l'algorithme de Von Neumann. Or leurs résultats valent pour des convexes fermés dans des espaces de Hilbert. Ainsi en reprenant la démonstration avec $\mathbb{R}^n$ remplacé par $L_2(\mu)$, A par $\mathcal{C}^+$ et B par $r + \mathcal{C}^+$, on obtient

$$\lim_{k\to\infty} \|r^{(k)} - r\|$$

qui montre que l'erreur d'approximation tend vers 0 avec le nombre d'itérations.

Enfin, nous concluons en Section 2.3.4 en montrant comment, par un jeu de compromis entre le nombre de points et le nombre d'itérations, on obtient le Théorème 2.4. Pour faciliter la lecture, tous les points techniques sont reportés en Annexe.

2.3.2 Consistance de la régression isotonique pour une fonction non motonone

Nous montrons ici que l'estimateur de régression isotonique appliqué à des données suivant le modèle général (2.27) converge vers la fonction isotonique la plus proche de la fonction de régression.

Comme cela revient à analyser la première demi-étape de l'algorithme I.I.R., le nombre d'itérations k de la méthode n'intervient pas. Aussi, pour alléger les notations, nous nous affranchissons de l'exposant qui correspond au nombre d'itérations en notant :

$$\hat{u}_n := \hat{u}_n^{(1)}, \qquad u_n := u_n^{(1)}, \qquad u_+ := u^{(1)}.$$

Obtenir ce résultat de consistance nécessite l'emploi d'inégalités de concentration pour assurer le passage d'une norme à l'autre. Ces inégalités, qui constituent l'Annexe H, supposent de travailler avec des fonctions bornées. En prenant r à variation bornée et en supposant le bruit borné, il est clair d'après les "min-max formulas" et la généralisation donnée dans Anevski & Soulier (2011), que $\hat{u}_n$, u_n et u_+ sont bornées.

Le résultat s'énonce de la manière suivante :

Théorème 2.5 (Régression isotonique sur une fonction non monotone)
Sous l'hypothèse que le bruit ε du modèle (2.27) est borné, on a

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n - u_+\|^2\right] = 0$$

avec $\hat{u}_n = \operatorname{argmin}_{u\in\mathcal{C}_n^+} \|Y - u\|_n$ *et* $u_+ = \operatorname{argmin}_{u\in\mathcal{C}^+} \|r - u\|$.

La figure 2.11 donne une illustration. On observe la proximité entre la courbe représentant l'application de la régression isotonique sur les données et celle représentant l'application de la régression isotonique sur la fonction de régression.

Preuve
D'après le Lemme F.1 démontré en Annexe F, on a, pour le terme de biais

$$\lim_{n\to\infty} \|u_n - u_+\| = 0 \qquad p.s.$$

Par convergence dominée, on déduit pour l'espérance

$$\lim_{n\to\infty} \mathbb{E}\left[\|u_n - u_+\|^2\right] = 0.$$

Il faut supposer le bruit borné pour avoir la convergence du terme de variance. Par le Lemme G.1 (cf. Annexe G), on a alors

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n - u_n\|^2\right] = 0$$

et l'inégalité triangulaire permet de conclure. □

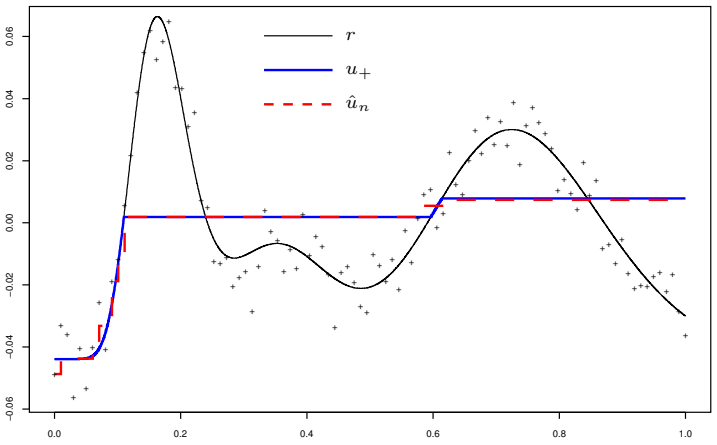


Fig. 2.11 – Régression isotonique sur une fonction non monotone.

2.3.3 Régression isotonique itérée : contrôle de l'erreur d'estimation

En reprenant les notations propres à l'algorithme, nous avons d'après la preuve du Théorème 2.5,

$$\lim_{n\to\infty} \mathbb{E}\left[\|u_n^{(1)} - u^{(1)}\|^2\right] = 0 \qquad \text{et} \qquad \lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n^{(1)} - u^{(1)}\|\right]^2 = 0.$$

Pour ce qui est du terme de variance, on a montré au passage (cf. équation (G.5) page 133) l'analogue pour la norme $\|.\|_n$

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n^{(1)} - u_n^{(1)}\|_n^2\right] = 0.$$

Par ailleurs, le Lemme H.2 de l'Annexe H permet de déduire de la convergence presque sûre vers 0 de $\|u_n^{(1)} - u^{(1)}\|$ celle de $\|u_n^{(1)} - u^{(1)}\|_n$ vers 0. Par convergence dominée, on obtient alors

$$\lim_{n\to\infty} \mathbb{E}\left[\|u_n^{(1)} - u^{(1)}\|_n^2\right] = 0$$

donc au total

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n^{(1)} - u^{(1)}\|_n^2\right] = 0. \tag{2.30}$$

Nous voyons dans ce qui suit comment itérer ces résultats selon l'algorithme backfitting. Nous détaillons la fin de la première étape puis la seconde avant de généraliser. Pour cela, nous manipulons pour $k \geq 1$ fixé, les éléments $r - u^{(k)}$ et $r - b^{(k)}$. Précisons une fois pour toutes que, toujours d'après Anevski et Soulier, ces éléments sont bornés, ce qui justifie l'utilisation des inégalités de concentration de l'Annexe H.

Fin de la première étape Nous montrons tout d'abord que $\mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|^2\right] \to 0$.
On a par définition

$$b^{(1)} = \operatorname*{argmin}_{b\in\mathcal{C}^-} \|r - u^{(1)} - b\| \qquad \text{et} \qquad \hat{b}_n^{(1)} = \operatorname*{argmin}_{b\in\mathcal{C}_n^-} \|Y - \hat{u}_n^{(1)} - b\|_n.$$

On introduit les vecteurs de $\mathbb{R}^n$

$$\tilde{Y} = Y - u^{(1)} \qquad \text{et} \qquad \tilde{b}_n^{(1)} = \underset{b \in \mathcal{C}_n^-}{\operatorname{argmin}} \|\tilde{Y} - b\|_n.$$

Remarquons que

$$\tilde{Y} = \left(r - u^{(1)}\right) + \varepsilon$$

et que

$$\tilde{b}_n^{(1)} = \underset{b \in \mathcal{C}_n^-}{\operatorname{argmin}} \|(r - u^{(1)}) + \varepsilon - b\|_n.$$

On comprend ainsi que l'on peut étudier $\|\tilde{b}_n^{(1)} - b^{(1)}\|$ selon le même schéma que celui suivi pour montrer le Théorème 2.5, $\tilde{b}_n^{(1)}$ jouant le rôle de $\hat{u}_n^{(1)}$ avec r remplacée par $r - u^{(1)}$ et la régression isotonique remplacée par la régression antitonique. D'après ce que l'on a dit au début de cette section, on a en particulier la consistance pour la norme $\|.\|_n$. Par ailleurs, la régression antitonique réduisant les distances, on a

$$\|\hat{b}_n^{(1)} - \tilde{b}_n^{(1)}\|_n \leq \|Y - \hat{u}_n^{(1)} - \tilde{Y}\|_n = \|\hat{u}_n^{(1)} - u^{(1)}\|_n$$

soit une majoration par un terme dont on a montré qu'il tend en espérance vers 0 (cf. équation (2.30)).
Ainsi,

$$\mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|_n^2\right] \leq 2 \times \left\{\mathbb{E}\left[\|\tilde{b}_n^{(1)} - b^{(1)}\|_n^2\right] + \mathbb{E}\left[\|\hat{b}_n^{(1)} - \tilde{b}_n^{(1)}\|_n^2\right]\right\} \to 0.$$

On peut alors obtenir $\lim_{n\to\infty} \mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|^2\right] = 0$ via l'utilisation du Lemme H.3 comme pour la fin de l'étude de la variance à la première demi-étape (cf. page 133). A l'issue de l'étape 1, on a donc

$$\mathbb{E}\left[\|\hat{r}_n^{(1)} - r^{(1)}\|^2\right] \leq 2 \times \left\{\mathbb{E}\left[\|\hat{u}_n^{(1)} - u^{(1)}\|^2\right] + \mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|^2\right]\right\} \to 0.$$

Seconde étape Cette fois, on a

$$\hat{u}_n^{(2)} = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|Y - \hat{b}_n^{(1)} - u\|_n \qquad \text{et} \qquad u^{(2)} = \underset{u \in \mathcal{C}^+}{\operatorname{argmin}} \|r - b^{(1)} - u\|.$$

On introduit

$$\tilde{Y} = Y - b^{(1)} = (r - b^{(1)}) + \varepsilon \qquad \text{et} \qquad \tilde{u}_n^{(2)} = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|\tilde{Y} - u\|_n = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|(r - b^{(1)}) + \varepsilon - u\|_n.$$

On est ainsi ramené à la preuve du Théorème 2.5 avec $r - b^{(1)}$ jouant le rôle de r et $\tilde{u}_n^{(2)}$ celui de $\hat{u}_n^{(1)}$. Retenant les résultats de convergence en norme $\|.\|_n$ on obtient

$$\lim_{n\to 0} \mathbb{E}\left[\|\tilde{u}_n^{(2)} - u^{(2)}\|_n^2\right] = 0.$$

De par la propriété de réduction des distances et d'après ce qu'on a conclu en fin d'étape 1, on obtient

$$\mathbb{E}\left[\|\hat{u}_n^{(2)} - \tilde{u}_n^{(2)}\|_n^2\right] \leq \mathbb{E}\left[\|Y - \hat{b}_n^{(1)} - ((r - b^{(1)}) + \varepsilon)\|_n^2\right] \leq \mathbb{E}\left[\|\hat{b}_n^{(1)} - b^{(1)}\|_n^2\right] \to 0.$$

Par conséquent,

$$\mathbb{E}\left[\|\hat{u}_n^{(2)} - u^{(2)}\|_n^2\right] \leq 2 \times \left\{\mathbb{E}\left[\|\tilde{u}_n^{(2)} - u^{(2)}\|_n^2\right] + \mathbb{E}\left[\|\hat{u}_n^{(2)} - \tilde{u}_n^{(2)}\|_n^2\right]\right\} \to 0$$

et par le Lemme H.3, Annexe H page 138, $\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n^{(2)} - u^{(2)}\|^2\right] = 0$.
Selon le même principe, on montre $\lim_{n\to\infty} \mathbb{E}\left[\|\hat{b}_n^{(2)} - b^{(2)}\|^2\right] = 0$ pour conclure

$$\mathbb{E}\left[\|\hat{r}_n^{(2)} - r^{(2)}\|^2\right] \leq 2 \times \left\{\mathbb{E}\left[\|\hat{u}_n^{(2)} - u^{(2)}\|^2\right] + \mathbb{E}\left[\|\hat{b}_n^{(2)} - b^{(2)}\|^2\right]\right\} \to 0.$$

En itérant le procédé, on montre

$$\forall k \geq 1 \qquad \lim_{n\to 0} \mathbb{E}\left[\|\hat{r}_n^{(k)} - r^{(k)}\|^2\right] = 0.$$

Autrement dit, à chaque étape, l'erreur d'estimation tend vers zéro lorsque la taille de l'échantillon tend vers l'infini.

2.3.4 Conclusion

La Section 2.3.3 a montré que

$$\forall k \geq 1 \qquad \lim_{n\to\infty} \mathbb{E}\left[\|\hat{r}_n^{(k)} - r^{(k)}\|\right] = 0.$$

Ainsi, pour $k = 1$, il existe n_1 tel que pour tout $n \geq n_1$

$$\mathbb{E}\left[\|\hat{r}_n^{(1)} - r^{(1)}\|\right] \leq \|r^{(1)} - r\|$$

donc

$$\mathbb{E}\left[\|\hat{r}_n^{(1)} - r\|\right] \leq 2\|r^{(1)} - r\|.$$

De même pour $k = 2$, il existe $n_2 > n_1$ tel que pour tout $n \geq n_2$,

$$\mathbb{E}\left[\|\hat{r}_n^{(2)} - r^{(2)}\|\right] \leq \|r^{(2)} - r\|$$

donc

$$\mathbb{E}\left[\|\hat{r}_n^{(2)} - r\|\right] \leq 2\|r^{(2)} - r\|.$$

En itérant sur k, nous construisons de proche en proche une suite $(n_k)_{k\geq 1}$ strictement croissante telle que pour tout $k \geq 1$ et tout $n \geq n_k$

$$\mathbb{E}\left[\|\hat{r}_n^{(k)} - r\|\right] \leq 2\|r^{(k)} - r\|.$$

Puisque par définition $\hat{r}_n^{(0)} = 0$, cette propriété peut être généralisée à $k = 0$ en convenant que $n_0 = 0$ et $r^{(0)} = 0$.

L'association $k \mapsto n_k$ étant strictement croissante, construisons la suite (k_n) construite de la façon suivante : $k_n = 0$ si $n < n_1$, $k_n = 1$ si $n_1 \leq n < n_2$, ... La suite k_n croît avec n, tend vers l'infini et l'on a donc pour tout n

$$\mathbb{E}\left[\|\hat{r}_n^{(k_n)} - r\|\right] \leq 2\|r^{(k_n)} - r\| \xrightarrow[n\to\infty]{} 0,$$

par la proriété de convergence de l'erreur d'approximation. Le Théorème 2.4 est donc prouvé.

Chapitre 3

Applications

Dans ce chapitre, nous voyons la mise en pratique de la régression isotonique itérée. La Section 3.1 montre son application sur des exemples simulés et aborde en particulier la question des critères d'arrêt. Nous considérons l'arrêt par validation croisée et adaptons, à notre cadre, des critères d'arrêt classiques pour des estimateurs linéaires. Sont également envisagés des critères intégrant des contraintes de forme portant sur le nombre de maxima locaux de l'ajustement. En comparant notre méthode à d'autres lisseurs univariés, nous observons son bon comportement au niveau des points de discontinuité pour les fonctions présentant des ruptures.

Dans la Section 3.2, nous l'appliquons ainsi à des données réelles issues de la génétique, données pour lesquelles la détection de ruptures est un enjeu. Les méthodes usuelles sont souvent des méthodes de segmentation qui nécessitent de fixer de nombreux paramètres. Notre estimateur étant constant par morceaux, nous choisissons de l'utiliser comme une méthode de segmentation. Il semble que l'on obtienne alors, de manière plus directe, des résultats comparables avec d'autres méthodes et que des perspectives d'améliorations existent selon la connaissance des données à traîter.

En Section 3.3, nous nous intéressons à un prolongement naturel de la régression isotonique : la régression unimodale. Il s'agit alors d'estimer la localisation du maximum d'une fonction croissante puis décroissante. D'un point de vue théorique, dans ce cas particulier, nous pouvons expliciter le résultat de l'application de k itérations de l'algorithme sur la fonction de régression et montrons que l'intervalle modal de cette fonction résultat contient le mode recherché. Nous montrons sur quelques cas simulés, comment la méthode peut alors servir de "pré-traitement" à la régression unimodale usuelle en donnant directement un nombre limité points candidats. Nous montrons enfin qu'étendre cette idée à la localisation de plusieurs modes semble offrir des perspectives intéressantes.

3.1 Simulations

Dans cette section, nous étudions le comportement de notre estimateur en l'appliquant sur des jeux de données simulées. Nous générons des points $(x_i, y_i)_{i=1\cdots n}$ i.i.d. selon le modèle

$$Y = r(X) + \varepsilon \tag{3.1}$$

avec r à variation bornée sur $[0,1]$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Nous envisageons 16 fonctions notées $r_1, \cdots, r_{16}$. Elles sont représentées en figure 3.1.

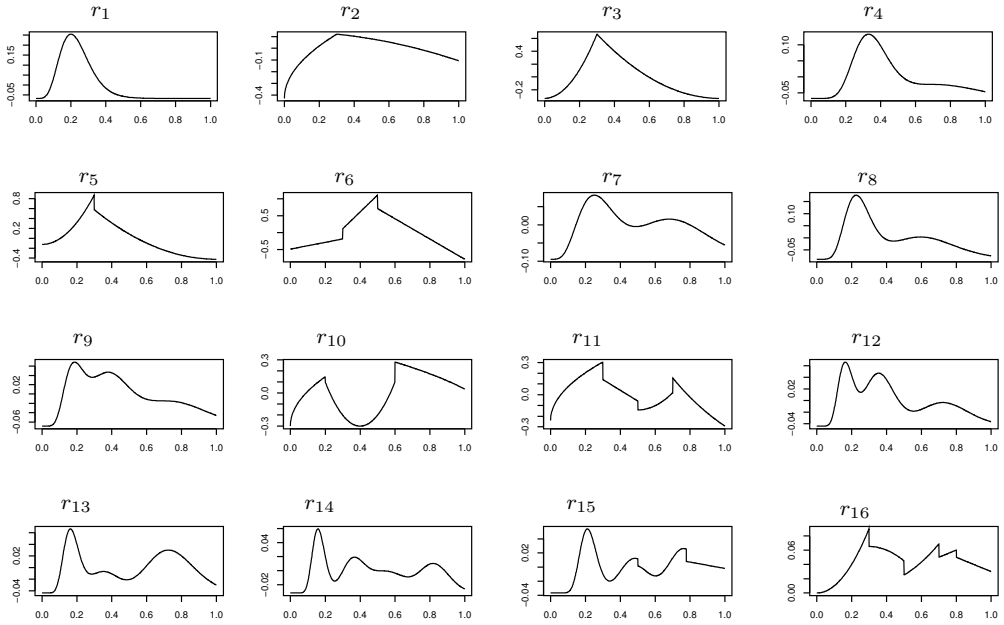


Fig. 3.1 – Les 16 fonctions considérées.

Six d'entre elles sont unimodales $(r_1, \cdots, r_6)$, cinq sont bimodales $(r_7, \cdots, r_{11})$ et cinq présentent 3 modes $(r_{12}, \cdots, r_{16})$. Les situations sont variées dans la mesure où les fonctions sont plus ou moins régulières et les modes plus ou moins faciles à détecter. Les fonctions r_4, r_9, r_{14} présentent chacune une tangente horizontale en un point qui n'est ni un mode ni un antimode. De plus, certaines fonctions présentent des points de rupture : r_5, r_6, r_{10}, r_{11}, r_{15} et r_{16}.

On se place en design fixe avec $x_1 = 0, \cdots, x_n = 1$ donc

$$x_i = \frac{i-1}{n-1} \qquad i = 1 \cdots n.$$

Nous considérons plusieurs tailles d'échantillons $(n = 50, 100, 200, 500, 1000)$ et trois niveaux de bruit :

$$\begin{aligned} \mathrm{var}(\varepsilon) &= 0.05\,\mathrm{var}(r) \\ \mathrm{var}(\varepsilon) &= 0.1\,\mathrm{var}(r) \\ \mathrm{var}(\varepsilon) &= 0.2\,\mathrm{var}(r) \end{aligned}$$

où $\mathrm{var}(r) = \frac{1}{n}\sum_{i=1}^{n}\left(r(x_i) - \bar{r}\right)^2$ avec $\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r(x_i)$. Pour fixer les idées, la fonction r_1 et $n = 100$ points aléatoires sont représentés pour ces trois niveaux de bruit en figure 3.2.

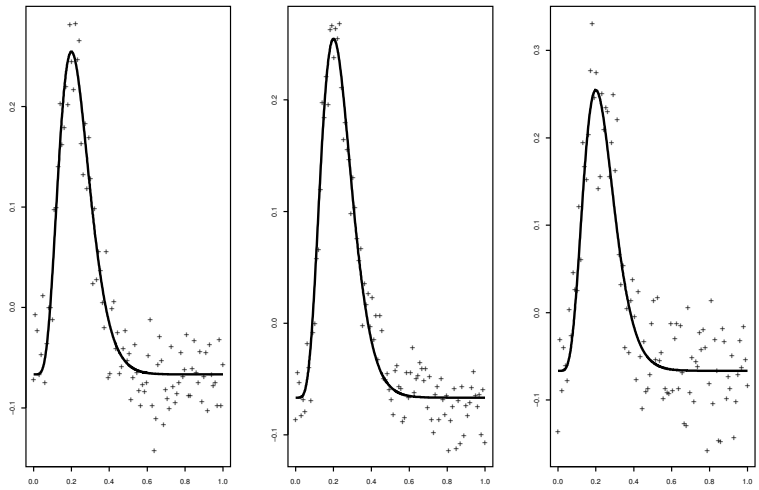


Fig. 3.2 – La fonction r_1 et les 3 niveaux de bruit ($n = 100$).

A partir d'un échantillon $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, on construit l'estimateur $\hat{r}$. On peut calculer en chaque point i, l'erreur d'ajustement $(r_i - \hat{r}_i)^2$ et en déduire l'erreur moyenne empirique

$$\|r - \hat{r}\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} (r_i - \hat{r}_i)^2 .$$

Cette quantité est aléatoire puisqu'elle dépend des points (x_i, y_i) de l'échantillon. En générant un grand nombre de jeux de données et en recalculant l'estimateur à chaque fois, on obtient une valeur approchée de l'espérance des erreurs. Pour $N_{\max}$ le nombre total de réplications, l'erreur quadratique moyenne au point x_i (notée MSE_i pour "Mean Square Error") est

$$\mathrm{MSE}_i = \mathbb{E}\left[(r(x_i) - \hat{r}(x_i))^2\right] \approx \frac{1}{N_{\max}} \sum_{N=1}^{N_{\max}} (r_i - \hat{r}_{i,N})^2 \tag{3.2}$$

où $\hat{r}_{i,N}$ désigne la valeur au point x_i de la $N^{\text{ème}}$ estimation. Pour l'erreur quadratique moyenne globale, on a

$$\mathrm{MSE} = \mathbb{E}\left[\|r - \hat{r}\|_n^2\right] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}_i \approx \frac{1}{N_{\max}} \sum_{N=1}^{N_{\max}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (r_i - \hat{r}_{i,N})^2 \right\}.$$

Notre estimateur tout comme ceux auquels nous le comparons peut présenter des effets de bord assez importants. Aussi n'avons-nous pris en compte que les points du design tombant dans l'intervalle $[0.1, 0.9]$ pour calculer les erreurs. Plutôt que la quantité précédente, nous avons ainsi calculé

$$\mathrm{MSE} = \mathbb{E}\left[\|r - \hat{r}\|_{n_t}^2\right] = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathrm{MSE}_i \approx \frac{1}{N_{\max}} \sum_{N=1}^{N_{\max}} \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} (r_i - \hat{r}_{i,N})^2 \right\}. \tag{3.3}$$

où l'indice t sert à préciser que l'on est dans cet ensemble test et où les observations tombant dans $[0.1, 0.9]$ sont supposées être réindexées de 1 à n_t. Le design étant fixe, cela revient à travailler avec environ 80% des données.

On sait que le fait d'augmenter le nombre d'itérations de notre méthode conduit à interpoler les données. Ce faisant, comme on l'a vu dans le chapitre précédent, l'erreur d'approximation (le biais) diminue. En contrepartie, on s'attend à ce que la variance de l'estimateur augmente avec le nombre d'itérations. En section 3.1.1, nous revenons sur ces considérations en détaillant le comportement de notre estimateur. Naturellement, un objectif de cette étude est de discuter des critères d'arrêt de la méthode. Nous envisageons trois façons de le faire : une procédure de validation croisée en section 3.1.2 ; l'application de critères pénalisés en section 3.1.3 ; l'application de critères basés sur la connaissance a priori du nombre de modes de la fonction de régression en section 3.1.4. Nous comparons enfin notre méthode à d'autres lisseurs univariés classiques (splines de lissage et polynômes locaux) en section 3.1.5.

Nous avons effectué $N_{\max} = 1000$ réplications des méthodes. Les simulations ont été effectuées avec le logiciel libre R. Plusieurs packages permettent de faire de la régression isotonique dont le package Iso que nous utilisons. En section 3.1.5, nous appliquons aussi aux données des estimateurs à noyaux et des splines de lissage. Pour cela nous avons utilisé le package locpol qui propose une méthode de régression polynomiale locale de degré 1 et la fonction smooth.spline du package stats qui ajuste une spline de lissage cubique.

3.1.1 Comportement de l'estimateur I.I.R. suivant le nombre d'itérations.

Dans cette section, nous cherchons à évaluer les termes de biais et de variance lorsque le nombre d'itérations augmente.

Notons $\hat{r}^{(k)} = \left[\hat{r}_1^{(k)}, \cdots, \hat{r}_n^{(k)}\right]'$ le vecteur des valeurs ajustées par la méthode I.I.R. après k itérations. La MSE cumule les termes de biais et de variance. Au point i, on a la décomposition suivante :

$$\begin{aligned} \mathrm{MSE}_i^{(k)} &= \mathbb{E}\left[\left(\hat{r}_i^{(k)} - r_i\right)^2\right] &&= \mathbb{E}\left[\left(\hat{r}_i^{(k)} - \mathbb{E}[\hat{r}_i^{(k)}]\right)^2\right] &&+ \left(\mathbb{E}\left[\hat{r}_i^{(k)} - r_i\right]\right)^2 \\ &&&= \mathrm{var}\left(\hat{r}_i^{(k)}\right) &&+ \mathrm{biais}^2\left(\hat{r}_i^{(k)}\right). \end{aligned} \tag{3.4}$$

Pour évaluer sur des données simulées les quantités en présence dans (3.4), on remplace les espérances par les moyennes empiriques sur les $N_{\max}$ réplications. Par exemple :

$$\mathbb{E}\left[\hat{r}_i^{(k)}\right] \approx \frac{1}{N_{\max}} \sum_{N=1}^{N_{\max}} \hat{r}_{i,N}^{(k)}.$$

En effectuant la moyenne des termes de l'équation (3.4) sur les points du design tombant dans $[0.1, 0.9]$, on obtient ensuite une mesure globale de l'erreur sur cette partie test.

La figure 3.3 représente, pour une fonction particulière, la suite $(\mathrm{MSE}^{(k)})$ sur une grille d'itérations allant de $k = 1$ à $k = 60$.

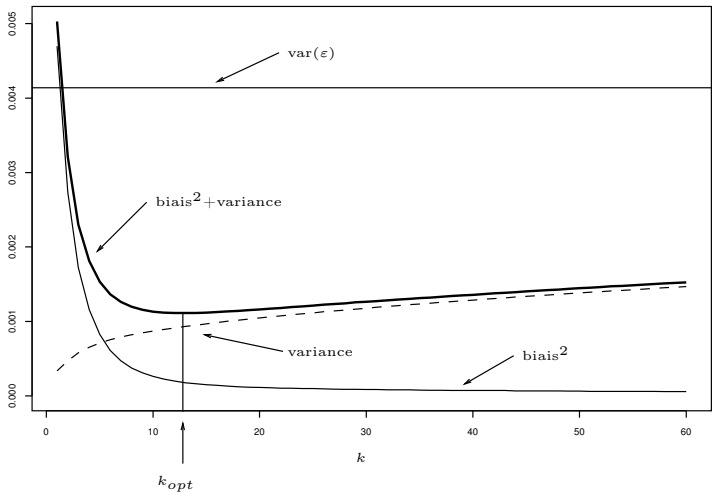


Fig. 3.3 – Décomposition biais-variance de $\hat{r}^{(k)}$ en fonction de k. Exemple de r_{11} pour $n = 100$ et pour le $3^{\text{ème}}$ niveau de bruit.

On retrouve dans les autres cas ce qu'on observe dans l'exemple représenté, à savoir que le biais décroît avec k alors que la variance augmente. En augmentant encore le nombre d'itérations, on observe aussi que la somme $\text{MSE}^{(k)}$ tend vers $\text{var}(\varepsilon)$, ce qui est cohérent avec le fait qu'augmenter le nombre d'itérations tend à reproduire les données. Le point important, et typique pour ces courbes, est que la suite $\left(\text{MSE}^{(k)}\right)_k$ décroît avant de croître et présente ainsi un minimum atteint pour un nombre d'itérations que nous notons k_{opt}.

Le nombre d'itérations solution de

$$\underset{k\in\mathbb{N}}{\text{argmin}}\, \mathbb{E}\left[\int_{0.1}^{0.9}\left(r(x)-\hat{r}^{(k)}(x)\right)^2\mu(dx)\right]$$

tout comme celui de

$$\underset{k\in\mathbb{N}}{\text{argmin}}\, \mathbb{E}\left[\frac{1}{n_t}\sum_{i=1}^{n_t}\left(r_i-\hat{r}_i^{(k)}\right)^2\right] \tag{3.5}$$

sont bien sûr inconnus en pratique puisqu'ils dépendent notamment de la fonction de régression et de la loi de ε qui sont inconnues. En revanche, sur la base de l'ensemble des réplications, pour une fonction, un niveau de bruit et une taille d'échantillon donnés, on peut en calculer une valeur approchée. La valeur k_{opt} représentée en figure 3.3 correspond ainsi à une approximation de la solution de (3.5), pour la fonction r_{11}, pour le premier niveau de bruit et pour $n = 100$.

Il est intéressant de voir comment varie ce nombre selon les paramètres envisagés. Nous n'avons pas calculé la valeur de k_{opt} dans chacun des cas de notre étude mais en figure 3.4 sont représentées, pour deux fonctions, les valeurs de k_{opt} selon la taille de l'échantillon et le nombre de points. On observe que k_{opt} augmente avec le nombre de points et diminue avec la variance de ε. Ajoutons que ce nombre d'itérations a également tendance à augmenter avec le nombre d'extrema locaux de la fonction de régression : une fonction à trois modes sera généralement associée à une valeur de k_{opt} plus élevée qu'une fonction unimodale, les tailles d'échantillon et le niveau de bruit étant par ailleurs égaux.

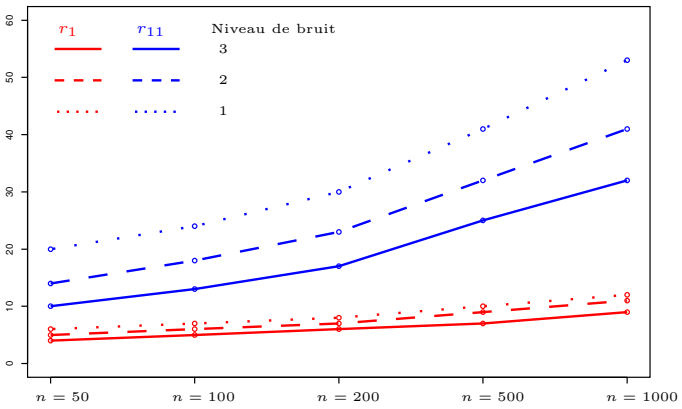


Fig. 3.4 – Représentation de k_{opt} en fonction de n et du niveau de bruit.

Dans la suite, ce n'est pas k_{opt} qui sert de référence pour mesurer la performance de nos estimateurs. Pour chacun des $N_{\max}$ ensembles de points, nous calculons

$$\hat{k}_{opt} = \underset{k}{\operatorname{argmin}} \frac{1}{n_t} \sum_{i=1}^{n_t} \left(r_i - \hat{r}_i^{(k)} \right)^2, \tag{3.6}$$

nombre d'itérations qui minimise, sur l'échantillon en question, l'écart aux vrais valeurs et que l'on peut voir comme une estimation ponctuelle de k_{opt}. Pour donner une idée de cet ajustement que nous qualifions d'optimal, nous représentons en figure 3.5 son allure $\hat{r}^{(\hat{k}_{opt})}$ sur un exemple.

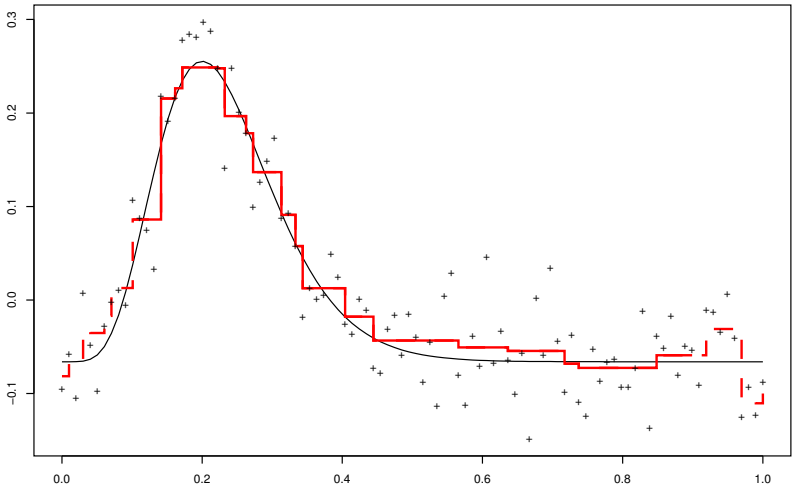


Fig. 3.5 – Exemple d'ajustement optimal pour un jeu de données simulé pour r_1, $n = 100$ et $3^{\text{ème}}$ niveau de bruit.

Nous voyons maintenant des critères permettant d'arrêter l'algorithme sur la seule base des données

3.1.2 Validation croisée

La validation croisée est souvent utilisée comme procédure de choix de modèle en régression (cf. Snee (1977) ou Picard & Cook (1984) par exemple), notamment lorsque l'on ne dispose pas de gros jeux de données et qu'on ne peut pas en acquérir de nouvelles pour tester les capacités prédictives de la méthode construite.

Le principe, dit de "data splitting", consiste à séparer les données en deux ensembles distincts : un échantillon d'apprentissage sur lequel on construit l'estimateur ; un échantillon de validation qui sert à tester l'estimateur construit. Plus précisément, on utilise les données (x_i, y_i) de l'échantillon d'apprentissage pour construire l'estimateur. On applique cet estimateur pour prédire des valeurs $\hat{y}_i$ aux points x_i de l'échantillon de validation et l'on mesure ensuite l'écart entre les valeurs prédites et les valeurs réellement observées sur cet ensemble de validation.

Pour éviter que les estimateurs ainsi obtenus aient une variance trop élevée, on peut améliorer la procédure de la manière suivante : on sépare l'échantillon en une partition de K sous-ensembles, chaque sous-ensemble servant tour à tour à tester l'estimateur qui est construit sur la réunion des $K-1$ autres et une moyenne (par exemple) des K erreurs de prédiction étant calculée à fin de chaque boucle. Ce principe est appelé "K-fold cross-validation" dans la terminologie anglo-saxonne. Le plus courant ("Leave-one-out cross validation") est celui où $K = n$; dans ce cas, chacune des valeurs de l'échantillon prise isolément sert tour à tour d'échantillon de validation pendant que toutes les autres forment l'échantillon d'apprentissage.

Plusieurs articles discutent les aspects pratiques et théoriques de ces procédures comme Snee (1977), Efron (1983), Breiman & Spector (1992) ou Wong (1983). On peut aussi citer Kohavi (1995) et Efron & Tibshirani (1997) pour des améliorations de type "bootstrap" par exemple. On peut retenir que la "leave-one-out cross-validation" présente un biais faible mais une forte variance et que de manière générale, le choix de la procédure est fortement lié au type de données et aux tailles d'échantillons. Par ailleurs, et c'est surtout cet aspect qui nous concerne, le coût opératoire peut s'avérer très élevé si la méthode de lissage employée est elle-même coûteuse. Pour ce qui nous concerne, si l'on admet que la régression isotonique sur n points se calcule en n opérations, k itérations de l'algorithme I.I.R nécessitent $2k \times n$ opérations et prendre $K = n$ pour la validation croisée, multiplie le tout par n. Si l'on choisit k de l'ordre de n, le total est donc de l'ordre de $\mathcal{O}(n^3)$. Cette difficulté liée au temps de calcul est d'ailleurs un inconvénient majeur de cette méthode et dans la pratique, on essaye si possible d'appliquer des critères plus économiques comme ceux que nous verrons en section suivante.

La procédure que nous proposons ici s'inspire de ce qui vient d'être dit. Nous séparons aléatoirement chacun des $N_{\max}$ échantillons en un échantillon de validation comprenant 5% des données, le reste constituant l'échantillon d'apprentissage. Sur l'échantillon d'apprentissage, nous appliquons l'algorithme de régression isotonique itérée sur une grille possible d'itérations allant de $k_{\max} = 50$ pour $n = 50$ à $k_{\max} = 150$ pour $n = 1000$. Pour chaque itération k, l'estimateur $\hat{r}^{(k)}$ est utilisé pour prédire Y en les x_i de l'échantillon de validation. Nous calculons ensuite l'erreur de prédiction sur cet échantillon et retenons l'itération produisant l'erreur de prédiction la plus faible (en prenant l'intervalle $[0.1, 0.9]$ comme référence). Pour une fonction, une taille d'échantillon et un niveau de bruit donnés, nous disposons ainsi d'une suite de $N_{\max}$ itérations notées $\left(\hat{k}_{vc}^N\right)_{N=1\cdots N_{\max}}$.

Nous analysons plus en détail le comportement de l'estimateur arrêté par cette procédure de validation croisée dans la section suivante en le comparant à celui obtenu par d'autres critères d'arrêt. Pour se faire une première idée, nous représentons sur un exemple la distribution des valeurs $\left(\hat{k}_{vc}^N - \hat{k}_{opt}^N\right)_{N=1\cdots N_{\max}}$ en figure 3.6 pour les cinq tailles d'échantillon considérées.

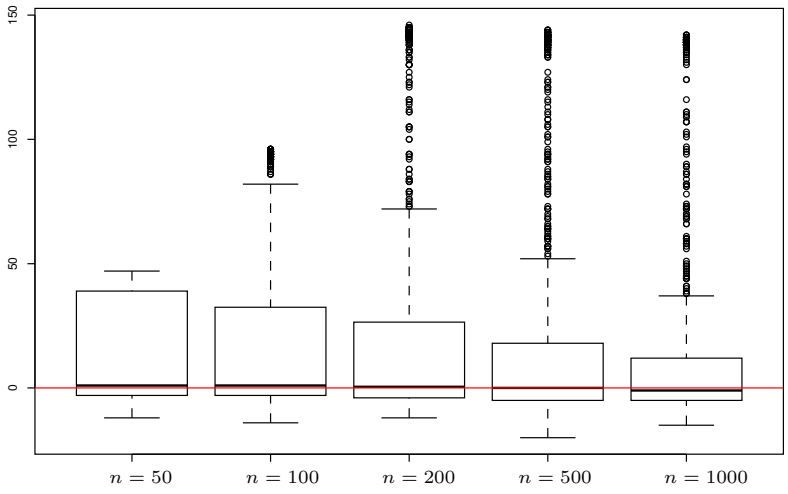


Fig. 3.6 – Distribution de $\left(\hat{k}_{vc}^N - \hat{k}_{opt}^N\right)_{N=1\cdots N_{\max}}$. Cas de r_4 pour le $3^{\text{ème}}$ niveau de bruit.

On observe (et cela se retrouve dans les autres cas) que le nombre d'itérations est bien souvent plus grand que le nombre souhaité. Cette tendance au sur-ajustement diminue cependant quand la taille des échantillons augmente.

3.1.3 Application de critères pénalisés

Nous voyons maintenant des critères d'arrêt qui résultent d'un compromis biais-variance. Nous considérons le critère AIC (Akaike Information Criterion, cf. Akaike (1973)), le critère AIC modifié AICc, proposé par Hurvich *et al.* (1998), le critère BIC (Bayesian Information Criterion, cf. Schwarz (1978)) et le critère GCV (Generalized Cross Validation, cf. Craven & Wahba (1978)). Pour des lisseurs linéaires, on peut les mettre sous la forme commune

$$\hat{k} = \underset{k}{\operatorname{argmin}} \log\left(\frac{\mathrm{RSS}_k}{n}\right) + \phi(p_k). \tag{3.7}$$

RSS désigne la somme des carrés des résidus :

$$\mathrm{RSS}_k = \sum_{i=1}^{n_t} (y_i - \hat{y}_i)^2 \tag{3.8}$$

et ϕ est une fonction croissante du nombre p_k de degrés de liberté du lisseur (ddl en abrégé). Ce nombre contrôle la proximité aux données. En régression linéaire, il s'agit du nombre de variables explicatives donc de la trace de la matrice de projection. Pour un lisseur linéaire, il correspond à la trace de la matrice de lissage qui augmente lorsque le paramètre de lissage diminue. Dans (3.7), le premier terme favorise ainsi la fidélité aux données alors que le second la pénalise.

Les fonctions ϕ associées à chacun des quatre critères mentionnés plus haut sont :

$$\begin{array}{ll} \phi_{AIC}(p_k) & = 2\frac{p_k}{n} \\ \phi_{BIC}(p_k) & = \frac{p_k}{n}\log n \\ \phi_{AICc}(p_k) & = 1 + 2\frac{p_k+1}{n-p_k-2} \\ \phi_{GCV}(p_k) & = -2\log\left(1 - \frac{p_k}{n}\right). \end{array} \tag{3.9}$$

La régression isotonique n'étant pas un estimateur linéaire (cf. page 10), l'estimateur de régression isotonique itérée n'est pas linéaire non plus. Il n'est donc pas possible de faire référence à une matrice de lissage pour appliquer ces critères pénalisés. Nous choisissons de prendre comme nombre équivalent de degrés de liberté le nombre de valeurs différentes (nombre de sauts $+1$ pour être plus précis) prises par l'estimateur. Outre le fait qu'il n'est pas contre-intuitif de considérer que le nombre de sauts est en effet associé à la régularité de l'estimateur (dans le sens où plus le nombre est important et plus on est proches des données), nous nous appuyons, pour motiver ce choix, sur les travaux de Meyer & Woodroofe (2000). Les auteurs introduisent la divergence de l'estimateur définie par

$$D = \operatorname{div}(\hat{r}) = \sum_{i=1}^{n} \frac{\partial}{\partial y_i} \hat{r}_i(y)$$

et qu'ils associent à la dimension effective du modèle. Cette opérateur de divergence correspond à la trace de la matrice jacobienne de l'application

$$\text{iso} \left\{ \begin{array}{ccc} \mathbb{R}^n & \to & \mathbb{R}^n \\ y = (y_1, \cdots, y_n)' & \mapsto & \hat{r} = (\hat{r}_1, \cdots, \hat{r}_n)' \end{array} \right.$$

Ils montrent par ailleurs que dans le cas de la régression isotonique, elle correspond au nombre de valeurs distinctes dans $(\hat{r}_1, \cdots, \hat{r}_n)'$. Ce point a une interprétation assez intuitive si l'on se souvient que l'estimation $\hat{r}_i$ au point x_i est le résultat d'une moyenne locale autour de ce point. Reprenons pour comprendre le petit exemple de la page 12 donné en Section 1.1.2 où

$$\hat{r}_1 = \hat{r}_2 = \frac{1}{2}, \qquad \hat{r}_3 = \hat{r}_4 = \hat{r}_5 = \frac{2}{3}, \qquad \hat{r}_6 = 2.$$

On a ainsi

$$\hat{r}_1 = \hat{r}_2 = \frac{1}{2}\left(y_1 + y_2\right),$$

donc

$$\sum_{j=1}^{2} \frac{\partial}{\partial y_j} \hat{r}_j(y) = 2 \times \frac{1}{2} = 1.$$

De même

$$\hat{r}_3 = \hat{r}_4 = \hat{r}_5 = \frac{1}{3}\left(y_3 + y_4 + y_5\right),$$

donc

$$\sum_{j=3}^{5} \frac{\partial}{\partial y_j} \hat{r}_j(y) = 3 \times \frac{1}{3} = 1,$$

et

$$\frac{\partial}{\partial y_6} \hat{r}_6(y) = 1.$$

Finalement D correspond au nombre de blocs : $D = 3$.

Pour un lisseur linéaire, puisque la matrice jacobienne est la matrice de lissage, on retrouve avec cette quantité la trace de la matrice de lissage. De plus, Meyer & Woodroofe (2000) établissent les deux inégalités suivantes qui généralisent des inégalités classiques vérifiées dans le cas linéaire et qui renforcent encore l'analogie :

$$\mathbb{E}[\|\hat{r} - r\|_n^2] \leq \frac{1}{n}\sigma^2 \mathbb{E}[D] \text{ et } \mathbb{E}[\|y - \hat{r}\|^2] \leq \frac{1}{n}\sigma^2 \mathbb{E}[n - D].$$

Notons au passage que les auteurs en déduisent que, pour la régression isotonique, $\mathbb{E}[D] \leq Cn^{1/3}$ avec C une constante ne dépendant que de $\Delta = (r(x_n) - r(x_1))/\sigma$ et qu'ils retrouvent ainsi la vitesse de convergence de Brunk (1970) pour une fonction lipschitzienne.

Pour renforcer ce choix, notons que la même idée a été reprise très récemment par Tibshirani *et al.* (2011) qui proposent un estimateur approchant de façon presque monotone les données. Ils formalisent en considérant le problème de minimisation :

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda \sum_{i=1}^{n-1} (\beta_i - \beta_{i+1})_+$$

où $x_+ = \max(x, 0)$. Le terme de droite pénalise les paires adjacentes violant la monotonie et le premier mesure la fidélité aux données. Pour $\lambda = 0$, la solution interpole les données et pour $\lambda \to \infty$, on obtient la régression isotonique. Entre ces deux extrêmes, la solution présente donc un compromis entre respect de la monotonie et fidélité au données. A λ fixé, l'estimateur est construit sur une variante de l'algorithme PAVA et le choix de λ se fait en prenant là encore le nombre de sauts de l'estimateur comme nombre de degrés de liberté.

La figure 3.7 présente pour les trois niveaux de bruit et pour les différentes valeurs de n, le nombre de sauts moyen en fonction du nombre d'itérations pour la méthode I.I.R. Est aussi représenté un "intervalle de confiance" de ce nombre moyen (la moyenne $\pm$ l'écart-type) obtenu sur les $N_{\max} = 1000$ réplications.

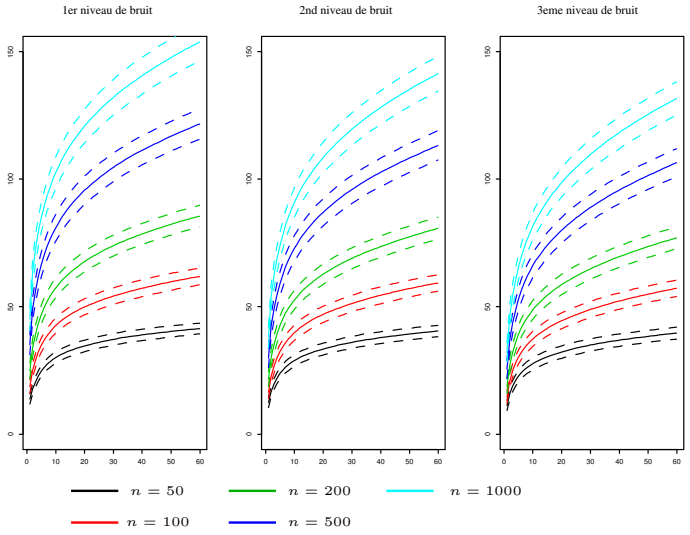


Fig. 3.7 – ddl de l'estimateur I.I.R. en fonction du niveau du nombre d'itérations, pour trois niveaux de bruit et les valeurs de n. Cas de la fonction r_1.

On constate que, pour un même nombre d'itérations, plus le nombre de points est grand et plus le nombre de sauts produits par l'estimateur est grand. On remarque que, le niveau de bruit diminuant, le nombre de sauts a tendance à augmenter, le nombre de points et le nombre d'itérations étant fixés. On peut proposer l'explication suivante : lorsque le niveau de bruit baisse, cela signifie que les points se rapprochent de la courbe sous-jacente ; imaginons une simple régression isotonique avec des points aléatoires autour d'une fonction de régression croissante. Si le bruit

est faible, la suite de points simulés sera elle même presque croissante conduisant à de nombreux sauts. Au contraire, si le bruit a un niveau important, la régression isotonique aura tendance à regrouper plus de points produisant un estimateur prenant moins de valeurs différentes. On peut imaginer que le phénomène est comparable lorsque la régression isotonique est itérée. Ces remarques valent pour les autres fonctions. La comparaison faite entre les différentes fonctions n'apporte pas d'élément particulier supplémentaire.

Nous voyons maintenant comment ces critères d'arrêt se comportent en pratique. Nous procédons comme en section précédente, à savoir que pour chacun des $N_{\max}$ ensembles de n points (x_i, y_i) et pour chacun des quatre critères, on calcule le nombre d'itérations minimisant le critère (3.7). On dispose donc, pour chacun des critères, d'une suite de $N_{\max}$ nombre d'itérations retenus, et ce pour chaque fonction, chaque niveau de bruit et chaque taille d'échantillon donnés. Nous les notons $\left(\hat{k}^N_{aicc}\right)$, $\left(\hat{k}^N_{bic}\right)$, $\left(\hat{k}^N_{aic}\right)$, $\left(\hat{k}^N_{gcv}\right)$, l'indice N allant de 1 à $N_{\max}$. La figure 3.8 représente sur un des cas simulés, la distribution des $N_{\max}$ valeurs de rapports $\hat{k}_{aicc}/\hat{k}_{opt}$, $\hat{k}_{bic}/\hat{k}_{opt}$, $\hat{k}_{aic}/\hat{k}_{opt}$ et $\hat{k}_{gcv}/\hat{k}_{opt}$ où $\hat{k}_{opt}$ est défini en équation (3.6).

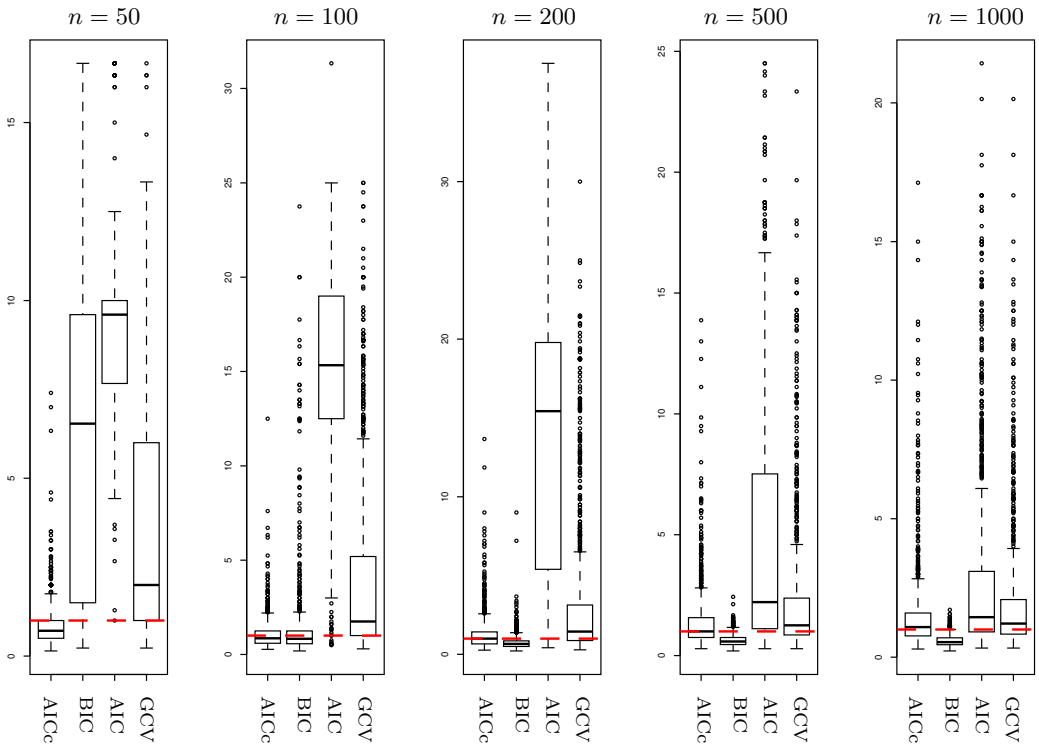


Fig. 3.8 – Distribution de $\hat{k}/\hat{k}_{opt}$ pour les 4 critères (fonction r_4, 3$^{\text{ème}}$ niveau de bruit).

On observe que le critère AICc présente les nombres d'itérations les plus proches de $\hat{k}_{opt}$. Il produit semble-t-il assez souvent un nombre trop grand d'itérations sauf quand $n = 50$. Le critère BIC se comporte plutôt bien lorsque $n \geq 100$ bien que généralement le nombre d'itérations soit inférieur à $\hat{k}_{opt}$; en revanche il ne semble pas adapté aux petits échantillons. Les critères AIC et GCV semblent produire de trop grands nombres d'itérations. Ajoutons que l'on retrouve ces remarques sur la plupart des cas de figure envisagés. Retenons que le critère AICc semble meilleur dans la

plupart des cas mais que le critère BIC produit des estimateurs sans doute plus robustes pour les échantillons assez grands.

Il peut être intéressant de comparer l'estimateur arrêté par validation croisée à ceux que nous venons de voir. Nous retenons les critères AICc et BIC qui se dégagent et comparons sur le même cas que précédemment avec les rapports $\hat{k}_{vc}/\hat{k}_{opt}$ en figure 3.9. Sur ce cas comme sur les autres, il semble que la validation croisée telle que nous l'avons programmée n'apporte rien de plus que le critère AICc. Si la médiane des rapports est assez proche de 1, on note néanmoins une tendance nette à produire beaucoup d'itérations.

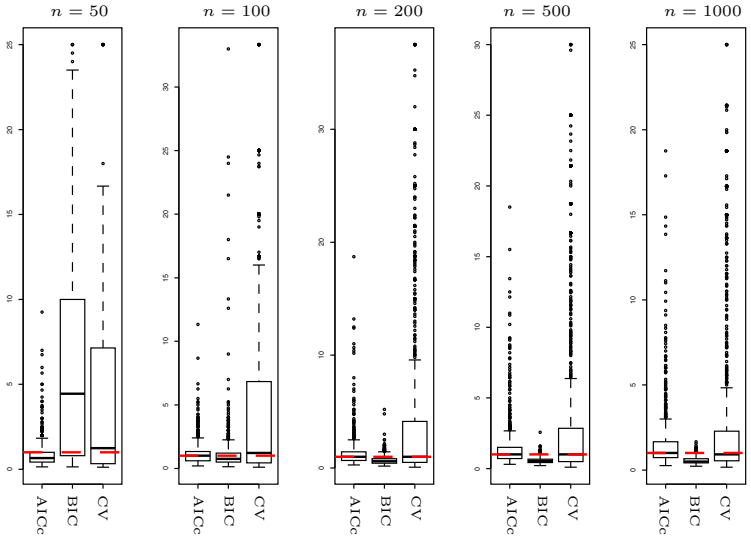


Fig. 3.9 – Distribution de $\hat{k}/\hat{k}_{opt}$ pour les critères AICc, BIC et validation croisée (fonction r_4, $3^{\text{ème}}$ niveau de bruit).

3.1.4 Estimations sous contrainte de forme

Comme nous utilisons comme base la régression isotonique qui est conçue pour l'estimation sous contrainte de forme, il est naturel de s'interroger sur l'intérêt qu'il pourrait y avoir à arrêter notre algorithme sur la base de connaissances a priori concernant la fonction de régression. Nous supposons connu le nombre de modes et proposons ainsi d'arrêter l'algorithme I.I.R. dès l'apparition de ce nombre de modes (estimateur noté inf) ou juste avant l'apparition d'un mode supplémentaire (estimateur noté sup). Dans certains cas, il est possible que cette idée ne puisse pas s'appliquer car la courbe ajustée ne fournit le bon nombre de modes pour aucune itération particulière. Dans ce cas d'ailleurs, aucun des deux estimateurs ne rend de résultat. Nous commençons par recenser l'importance de ce phénomène dans le tableau 3.1 (la recherche de modes se fait sur l'intervalle $[0.1, 0.9]$).

Tableau 3.1 – Recensement des cas d'absences de résultat pour les estimateurs contraints.

niveau 1																
$r =$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$n = 50$	1						3	40	125	8		4	62			30
$n = 100$								23	155				57			13
$n = 200$								13	142				22			11
$n = 500$								1	80				5			1
$n = 1000$									12				3			
niveau 2																
$r =$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$n = 50$	3						16	62	149	26		8	74	3	1	42
$n = 100$							6	56	130	5		2	68			37
$n = 200$							1	36	162	1			52			15
$n = 500$								6	140				21			3
$n = 1000$								1	78				1			
niveau 3																
$r =$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$n = 50$	25				2		34	84	175	39		38	69	2	5	67
$n = 100$	4						18	78	139	29		13	74	1	1	45
$n = 200$							5	47	133	3		8	70			40
$n = 500$								25	185				44			16
$n = 1000$								6	152				28			7

Mis à part pour la fonction r_9, la méthode fonctionne dans au moins 90% des cas. Il y a aussi le cas des fonctions r_8 et r_{13} pour lesquels le taux d'absence de résultat varie entre 5 et 10%. Ces fonctions possèdent la particularité qu'un de leur mode a une très faible amplitude ce qui le rend difficile à déceler. Une analyse plus détaillée de quelques cas pathologiques montre qu'en général, la courbe de l'estimateur restitue dans un premier temps les modes principaux de la vraie fonction. Si le niveau de bruit est assez important par rapport à l'amplitude d'un mode difficile à mettre en évidence, il arrive qu'à un moment, plusieurs modes supplémentaires locaux apparaissent d'un seul coup (le bon est d'ailleurs en général localisé). Il n'est pas étonnant dès lors que l'on observe que le nombre de situations litigieuses diminue lorsque le niveau de bruit baisse et que le nombre de points du design augmente.

Comme pour les estimateurs pénalisés, on peut comparer la distribution des itérations $\hat{k}_{\inf}$ et $\hat{k}_{\sup}$ à celle de $\hat{k}_{opt}$. Nous représentons en figure 3.10 la distribution des $N_{\max} = 1000$ rapports $\hat{k}_{\inf}/\hat{k}_{opt}$ et $\hat{k}_{\sup}/\hat{k}_{opt}$ pour la fonction r_4 déjà utilisée et le 3$^{\text{ème}}$ niveau de bruit. En guise de comparaison avec les résultats de la section précédente, nous représentons aussi celle des $\hat{k}_{aicc}/\hat{k}_{opt}$.

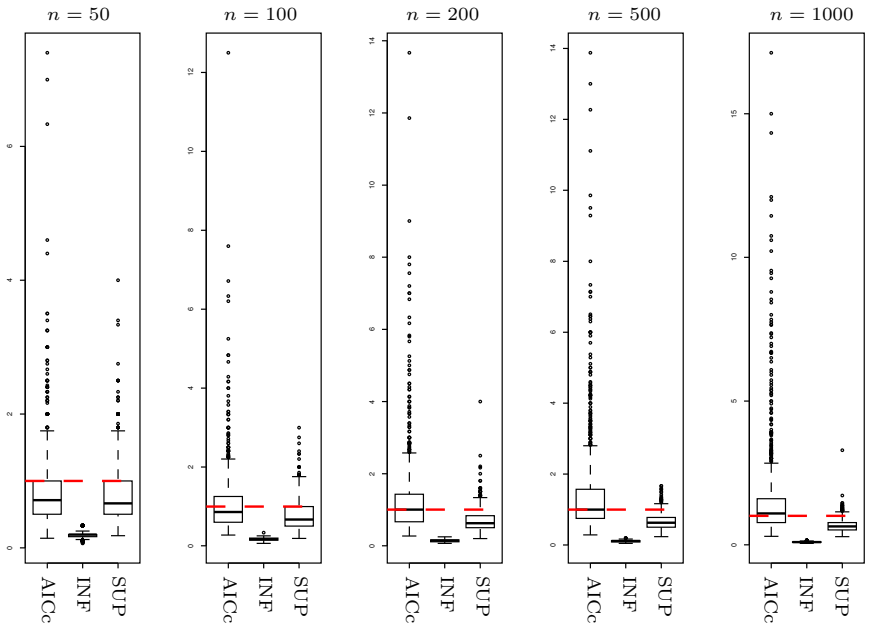


Fig. 3.10 – Distribution de $\hat{k}/\hat{k}_{opt}$ pour les critères AICc, inf et sup (fonction r_4, $3^{\text{ème}}$ niveau de bruit).

On observe que $\hat{k}_{\text{inf}}$ est systématiquement inférieur à $\hat{k}_{opt}$. C'est aussi en général le cas pour $\hat{k}_{\text{sup}}$ bien que cela, et c'est logique, soit moins marqué. Ce point semble assez normal dans la mesure où l'on peut considérer que ces estimateurs, qui ne sont pas construits pour respecter un compromis biais-variance, mais qui s'arrêtent "dès" l'apparition du bon nombre de modes itèrent moins longtemps. On constate quand même, et ce constat peut être fait pour la plupart des fonctions et des niveaux de bruit, que le fait d'arrêter l'algorithme avant l'apparition d'un mode supplémentaire donne des résultats assez comparables à l'arrêt par le critère AICc.

Pour conclure cette partie, nous présentons en tableau 3.2 un récapitulatif comparant les erreurs à la fonction de régression pour les estimateurs sup et ceux arrêtés par les critères AICc et BIC. Pour constituer ce tableau, nous avons calculé sur l'ensemble des $N_{\max}$ réplications les erreurs

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\left(\hat{r}_{i,N}-r_{i,N}\right)^2, \qquad N=1\cdots N_{\max}$$

pour chacun des lisseurs. Nous avons ensuite fait la moyenne de ces $N_{\max}$ valeurs puis, pour chaque cas envisagé, calculé le rapport à la moyenne la plus faible. Par exemple, pour r_1 et $n = 50$, c'est le critère sup qui présente l'erreur moyenne la plus faible (marquée en rouge). L'erreur moyenne pour le critère AICc est presque égale et pour le critère BIC, elle est 20% supérieure. Les résultats sont présentés pour le niveau intermédiaire de bruit.

	$n=50$			$n=100$			$n=200$			$n=500$			$n=1000$		
	aicc	bic	sup	aicc	bic	sup	aicc	bic	sup	aicc	bic	sup	aicc	bic	sup
r_1	1.01	1.20	1	1	1.06	1.18	1	1.12	1.24	1	1.19	1.26	1	1.23	1.23
r_2	1.20	1.51	1	1.06	1.11	1	1.06	1.12	1	1.05	1.16	1	1.04	1.17	1
r_3	1.22	1.31	1	1.01	1.07	1	1	1.04	1.01	1	1.10	1.03	1	1.14	1.03
r_4	1.05	1.24	1	1	1.06	1.09	1	1.04	1.09	1	1.10	1.08	1	1.12	1.07
r_5	1.41	1.20	1	1	1.04	1.04	1	1.06	1.12	1	1.10	1.17	1	1.13	1.17
r_6	1.30	1.24	1	1	1.04	1	1	1.05	1.05	1	1.08	1.06	1	1.09	1.07
r_7	1.25	1.34	1	1.09	1.17	1	1.05	1.23	1	1.04	1.31	1	1.04	1.39	1
r_8	1.20	1.34	1	1.08	1.12	1	1.04	1.15	1	1.03	1.24	1	1.02	1.31	1
r_9	1.21	1.40	1	1.06	1.12	1	1.01	1.13	1	1	1.24	1.06	1	1.34	1.09
r_{10}	1.42	1.35	1	1.10	1.17	1	1.05	1.19	1	1.03	1.28	1	1.02	1.33	1
r_{11}	2.26	1.17	1	1.11	1.11	1	1.01	1.07	1	1.01	1.14	1	1	1.17	1
r_{12}	1.59	1.16	1	1.12	1.16	1	1.03	1.23	1	1.01	1.35	1	1	1.41	1
r_{13}	1.51	1.20	1	1.10	1.13	1	1.05	1.12	1	1.04	1.25	1	1.03	1.31	1
r_{14}	1.95	1.09	1	1.15	1.11	1	1.04	1.14	1	1.02	1.23	1	1.01	1.33	1
r_{15}	2.42	1.03	1	1.18	1.06	1	1	1.07	1	1	1.19	1.02	1	1.25	1.03
r_{16}	2.22	1.08	1	1.21	1.10	1	1	1.08	1	1	1.17	1.04	1	1.23	1.06

Tableau 3.2 – Comparaison des erreurs moyennes pour l'algorithme I.I.R. selon 3 critères d'arrêt.

On observe que la connaissance du nombre de modes apporte un intérêt significatif lorsque le nombre de points est faible. Cet avantage est nettement moins marqué dès que les échantillons dépassent la taille 50 : les moyennes d'erreurs sont alors à peu près égales voire supérieures à ce que l'on obtient par le critère AICc. Le critère BIC semble le critère le moins adapté aux situations qui ont été simulées.

3.1.5 Comparaison avec des lisseurs classiques

Dans cette section, nous comparons les estimateurs I.I.R. que nous avons construits dans les sections précédentes à deux lisseurs classiques : un estimateur à noyau (méthode de régression linéaire locale) et une spline cubique de lissage. Nous ne retenons pour établir ces comparaisons que le meilleur estimateur pénalisé i.e. l'estimateur arrêté par le critère AICc et le meilleur estimateur construit sur une connaissance a priori du nombre de modes, l'estimateur noté sup.

Pour éviter les redondances dans la présentation des résultats, nous avons sélectionné 6 fonctions sur les 16 : deux fonctions unimodales (r_3 et r_6), deux fonctions bimodales (r_8 et r_{11}) et deux fonctions trimodales (r_{12} et r_{16}). Dans chacune de ces trois situations, la première fonction est continue alors que la seconde présente au moins deux ruptures (cf. figure 3.1 page 60).

Le tableau 3.3 présente, pour les fonctions sélectionnées, pour $n = 100, 500, 1000$ et pour le niveau bruit intermédiaire, les informations analogues à celles du tableau précédent.

Tableau 3.3 – Comparaisons à d'autres lisseurs.

	$n=100$				$n=500$				$n=1000$			
	spl	ker	aicc	sup	spl	ker	aicc	sup	spl	ker	aicc	sup
r_3	1	1.03	2.05	2.02	1	1.08	2.44	2.51	1	1.09	2.59	2.68
r_6	1	1.01	1.39	1.39	1	1.01	1.04	1.11	1.10	1.11	1	1.08
r_8	1	1.19	2.02	1.88	1	1.38	2.82	2.74	1	1.45	3.24	3.19
r_{11}	1	1.02	1.17	1.06	1.22	1.23	1	1	1.39	1.40	1	1.01
r_{12}	1	1.14	1.96	1.75	1	1.30	2.50	2.49	1	1.37	2.85	2.86
r_{16}	1	1.02	1.25	1.04	1.22	1.23	1	1.04	1.39	1.39	1	1.06

Il est clair que pour les fonctions régulières (fonctions r_3, r_8, r_{12}), l'estimateur à noyau et surtout la spline de lissage sont meilleurs que la méthode I.I.R. En revanche, pour les fonctions présentant des sauts, l'intérêt du critère AICc est manifeste dès que $n \geq 500$ et équivalent sinon (performances marquées en vert dans le tableau). On aboutit ainsi par exemple à des erreurs moyennes produites par l'estimateur à noyau ou la spline supérieures de 40% pour les fonctions r_{11} et r_{16}. Ajoutons pour compléter que les écarts s'accentuent avec un niveau de bruit qui baisse.

Il est intéressant d'essayer de mesurer l'impact que les points de ruptures ont sur ces écarts. Pour cela, nous représentons, pour chaque estimateur, la MSE en tout point i, présentée en équation (3.4) page 62 et que nous rappelons :

$$\begin{aligned} \mathrm{MSE}_i^{(k)} &= \mathbb{E}\left[\left(\hat{r}_i^{(k)} - r_i\right)^2\right] &&= \mathbb{E}\left[\left(\hat{r}_i^{(k)} - \mathbb{E}[\hat{r}_i^{(k)}]\right)^2\right] &&+ \left(\mathbb{E}\left[\hat{r}_i^{(k)} - r_i\right]\right)^2 \\ &&&= \mathrm{var}\left(\hat{r}_i^{(k)}\right) &&+ \mathrm{biais}^2\left(\hat{r}_i^{(k)}\right). \end{aligned}$$

On estime ces quantités en approchant les espérances par les moyennes sur les $N_{\max}$ réplications. Pour les quatre estimateurs considérés, les MSE sont représentées en figure 3.11. On a pris le cas de la fonction r_{11} pour le 3ème niveau de bruit (car les différences sont plus faciles à visualiser). La courbe de la fonction est rappelée en partie supérieure de la figure.

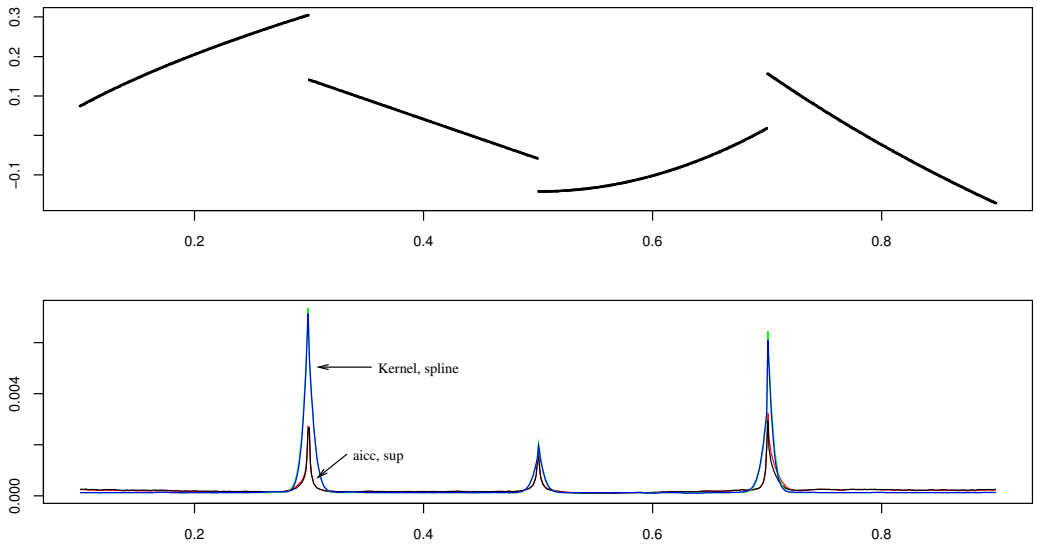


Fig. 3.11 – La MSE en tout point. Cas de r_{11}, $n = 1000$, 3ème niveau de bruit.

On observe une grande similarité des courbes associées aux estimateurs sup et AICc d'un côté et à l'estimateur à noyau et à la spline de lissage de l'autre. On peut penser que les différences proviennent essentiellement des écarts aux niveaux des sauts. On peut affiner l'analyse en représentant séparément les termes de biais et de variance (cf. figure 3.12).

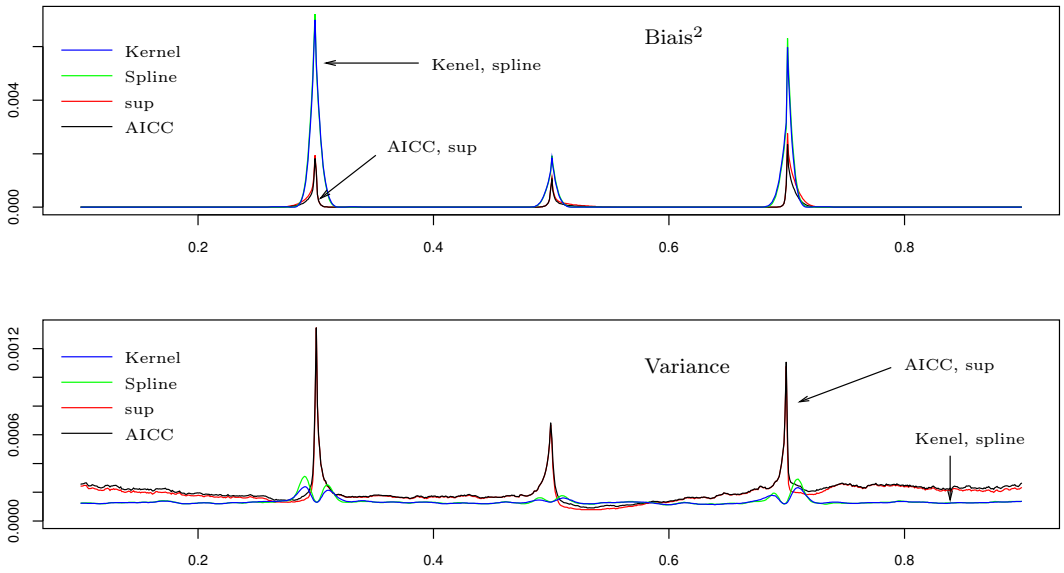


Fig. 3.12 – biais^2 et variance en tout point. Cas de r_{11}, $n = 1000$, 3$^{\text{ème}}$ niveau de bruit.

On voit que la supériorité de la méthode I.I.R. au niveau des sauts provient d'un biais plus faible, bien que l'estimateur soit plus variable que ses concurrents en ces endroits. On peut compléter en notant que, lorsque le niveau de bruit diminue, les écarts entre les termes de biais s'accroissent encore (au profit de la méthode I.I.R.), alors que les rapports de variance restent dans les mêmes proportions. Précisons enfin que, lorsque les fonctions sont plus régulières, l'estimateur contraint par le critère AICC présente un biais important au niveau des modes et antimodes, et une variance importante dans les parties monotones : ceci est illustré en figure 3.13 pour la fonction r_{12}.

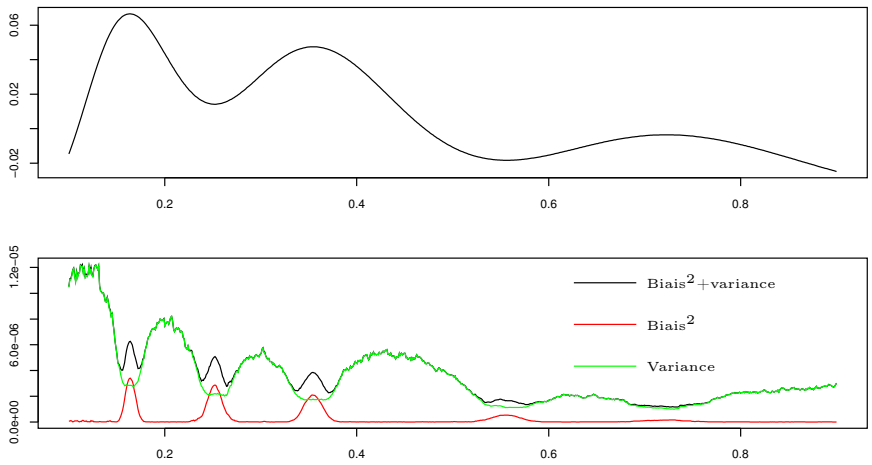


Fig. 3.13 – Décomposition biais-variance en tout point. Cas de r_{12}, $n = 1000$, 3$^{\text{ème}}$ niveau de bruit.

En conclusion, on peut retenir qu'au regard des simulations qui ont été faites, le critère AICc semble se distinguer des autres. Sans doute cependant faudrait-il poursuivre les comparaisons avec des tailles d'échantillons plus grandes. Par ailleurs, la procédure de validation croisée qui a été implémentée est très basique. L'améliorer, en sélectionnant dans chaque boucle plusieurs échantillons de validation pour moyenner les résultats, permettrait peut-être de diminuer les fortes variances constatées et de rendre ce critère compétitif. Il apparaît que notre méthode a un comportement intéressant en présence de ruptures sur la fonction de régression. On peut s'attendre à ce qu'elle puisse être employée à détecter de telles ruptures dans certains pratiques. Nous l'appliquons en ce sens à des données réelles dans la section qui suit. Enfin, un package R qui permet d'appliquer la méthode a été programmé. Il est disponible à l'adresse suivante :
http ://www.sites.univ-rennes2.fr/laboratoire-statistique/JEGOU/index.html

3.2 Localisation d'aberrations chromosomiques à partir de profils CGH

Dans la Section 3.2.1, nous présentons de façon générale les données provenant de puces CGH et les problématiques qu'elles soulèvent. En Section 3.2.2, nous évoquons quelques méthodes classiques. Parmi elles, la méthode GLAD (Hupé *et al.* (2004)) donne un accès facile à ses résultats illustrés sur un jeu de données public. Nous montrons en Section 3.2.3, qu'utilisée comme une procédure de segmentation, la régression isotonique itérée fournit de manière directe des résultats comparables sur ces mêmes données. Nous appliquons ensuite notre méthode à la recherche de gènes impliqués dans le retard mental chez les enfants et proposons quelques pistes spécifiques d'améliorations.

3.2.1 Problématique

Actuellement, l'utilisation des techniques basées sur les puces à ADN permet l'exploration à l'échelle du génome de différents niveaux moléculaires. Les biopuces ou *microarrays* permettent en une seule expérience la détection simultanée du niveau d'expression de milliers de gènes. Les puces d'hybridation génomique comparative ou puces CGH (*Comparative Genomic Hybridization*) sont utilisées pour la détection des anomalies du nombre de copies de l'ADN (cf. Pinkel *et al.* (1998), Ishkanian *et al.* (2004)). Cette technologie est notamment utile dans la recherche sur le cancer (Solinas-Toldo *et al.* (1997)) car la plupart des cancers présentent un contenu chromosomique anormal résultant

– soit en des gains et amplifications de zones de l'ADN susceptibles de porter des oncogènes (catégories de gènes favorisant la survenue des cancers),
– soit en des délétions partielles ou totales d'autres zones portant au contraire des gènes suppresseurs de tumeurs.

L'hybridation génomique comparative permet, après traitement, de mesurer les différences d'expression de séquences génomiques entre des cellules saines (même chez les patients malades), en général des lymphocytes, et des cellules potentiellement atteintes. Les profils obtenus correspondent au log du rapport des intensités mesurées sur les cellules tumorales et les cellules saines en des zones particulières analogues du génome (positions révélées par des sondes). En figure 3.14 sont représentés quatre exemples de profils CGH [1]. Les abscisses des points correspondent à la position des sondes et les ordonnées aux log-ratios des intensités.

1. Données fournies par Marie De Tayrac, Institut Génétique et Développement de Rennes.

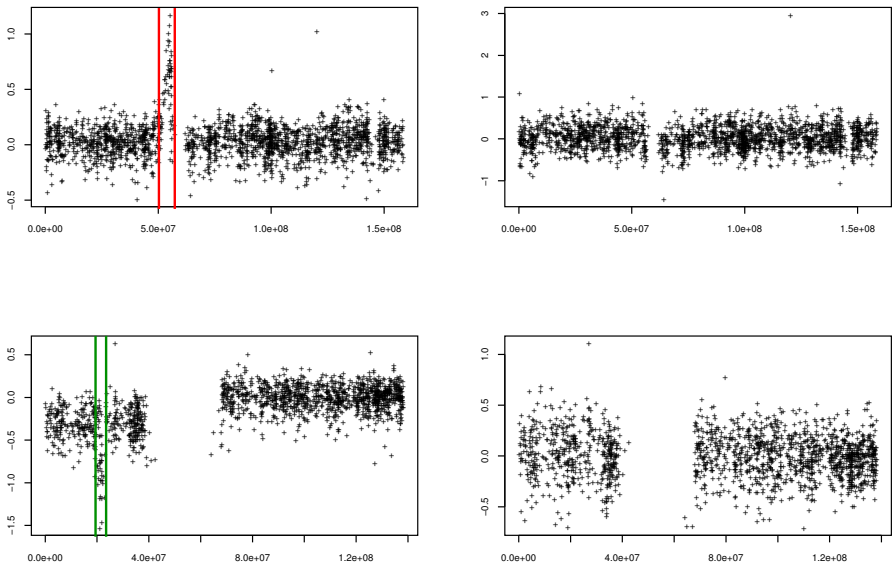


Fig. 3.14 – Exemples de profils CGH.

Les deux graphes de la partie supérieure concernent le chromosome 7, celui de gauche résultant des mesures effectuées sur un patient atteint d'un cancer du cerveau, celui de droite sur un patient sain. La partie délimitée par les deux bandes rouges est associée à une zone du génome portant le gène *EGFR* impliqué dans le développement des tumeurs. Chez le malade, on observe une amplification de cette zone correspondant à un nombre élevé de copies. La partie inférieure concerne le chromosome 9. On remarque une zone de délétion sur la tumeur du patient de gauche dont les cellules sont atteintes : cette zone emporte les gènes $p16^{INK4A}$ et $p15^{IN4KB}$ qui sont suppresseurs de tumeurs. Les profils CGH normaux, comme ceux de droite, correspondent à un signal bruité autour d'une constante (en général 0) résultant de rapports d'intensité normaux. Les sources de variabilité sont nombreuses. Elles proviennent en particulier d'aléas inhérents aux conditions expérimentales et sont également liées à la technologie bi-couleurs employée (biais dû au fluorochrome, cf. Rosenzweig *et al.* (2004)).

Une question importante est en tout cas de localiser les zones pour lesquelles dosages géniques sont significativement différents de 0 et donc de détecter en particulier les points de rupture, de sorte à identifier les régions impliquées. Dans la section qui suit, nous évoquons quelques outils développés en ce sens.

3.2.2 Méthodes existantes

Les méthodes envisagées considèrent en général la présence d'un nombre inconnu K de zones à l'intérieur desquelles la distribution du log-ratio Y suit un modèle gaussien

$$Y \sim \mathcal{N}(\mu_k, \sigma_k^2), \qquad k = 1 \cdots K.$$

Il s'agit alors d'estimer le nombre K de zones (ou "sets"), leur localisation, ainsi que les paramètres de moyenne et de variance pour chacune d'elle. Les approches principales ont trait aux méthodes de segmentation ou aux modèles de Markov cachés.

La procédure développée dans Fridlyand *et al.* (2004) est basée sur l'estimation de modèles de Markov cachés. L'estimation du maximum de vraisemblance des paramètres est pénalisée par le nombre de ruptures et se fait par un algorithme de type EM. Leur méthode suppose de définir un nombre maximum de ruptures possibles, ce qui semble tout à fait raisonnable avec la connaissance génétique des problèmes qu'ils envisagent. Elle suppose aussi de définir le critère de pénalisation. Plusieurs critères comme les critères BIC ou AIC ont été testés par simulation. De plus, à l'intérieur de la procédure, il est décidé de regrouper des ensembles contigus pour lesquels la différence des médianes des observations est inférieure à un certain seuil. Ce seuil, qui est fixé par l'utilisateur et qui peut varier selon les tumeurs étudiées, dépend spécifiquement de la connaissance que l'on a sur les données, notamment concernant la pureté de signal.

Les méthodes relevant de la segmentation visent également à déterminer la localisation des zones ainsi que leur nombre. Par exemple, les travaux de Jong *et al.* (2003) ou Olshen *et al.* (2004) s'appuient respectivement sur des critères de vraisemblance et de sommes partielles. Dans les deux cas, les points de rupture sont déterminés a posteriori sur des critères empiriques. L'algorithme d'optimisation utilisé dans Jong *et al.* (2003) est emprunté à Moscato (1989). On incrémente le nombre de sets possibles et pour chacun des nombres envisagés, on distribue aléatoirement les localisations possibles des points de rupture. La méthode déplace la position de ces points d'une unité à droite ou à gauche et calcule un score basé sur un compromis entre la vraisemblance des paramètres et une pénalisation du nombre de zones. Le modèle retenu est celui maximisant la fonction de score. L'algorithme utilisé par Olshen *et al.* (2004) s'appuie lui sur une variante de la segmentation binaire appelée *Circular Binary Segmentation* (CBS).

Mentionnons aussi la procédure proposée dans Picard *et al.* (2005) qui reprend des principes comparables (estimation du maximum de vraisemblance avec pénalisation du nombre de segments). Cependant, considérant que les critères classiques tels BIC ou AIC ont tendance à produire un nombre trop important de segments, ils proposent une façon adaptative de choix d'une constante venant pondérer le terme de pénalisation. La constante est choisie, parmi les pentes de la fonction de vraisemblance du nombre de segments, comme celle ne conduisant plus à une amélioration significative de la vraisemblance lorsque le nombre de segments est augmenté. Cependant, et c'est un point de critique admis par les auteurs, il est à nouveau nécessaire de fixer un seuil en dessous duquel on postule la non-significativité de l'amélioration, et la détermination de ce seuil résulte d'un choix subjectif de l'utilisateur.

La méthode de Hupé *et al.* (2004) a aussi retenu notre attention. La procédure se décompose en deux parties principales. La première vise à détecter les points de rupture dans un profil de sorte à délimiter différentes régions. La seconde consiste à assigner à chaque région une valeur commune révélant le nombre de copies associées à la zone. Pour la première phase, on postule un modèle gaussien autour d'une fonction constante par morceaux. L'estimation des niveaux μ_k et des points de ruptures se fait par maximisation de vraisemblance pénalisée. La procédure, qui est itérative, est basée une méthode de lissage appelée AWS pour "Adaptative Weights Smoothing" (cf. Polzehl & Spokoiny (2000)). Hupé *et al.* (2004) adaptent cette procédure à leur cas de figure. A chaque étape, une nouvelle valeur de la vraisemblance des paramètres est calculée à partir d'un voisinage autour de chaque point plus grand qu'à l'étape précédente et des poids affectés à chaque observation actualisés. Cette partie suppose le réglage de nombreux paramètres : choix du noyau pour le calcul des poids, choix de la fenêtre, paramètre mesurant l'augmentation de la taille de voisinage... Une fois obtenues les valeurs ajustées par cette méthode, les auteurs regroupent les segments successifs associés à des valeurs ajustées proches. A nouveau, cette étape suppose de fixer un seuil dont l'obtention n'est pas précisée. L'ensemble de la procédure est implémenté sous R via le package GLAD (Gain and Loss Analysis of DNA). Dans la section qui suit, nous comparons notre méthode avec cette dernière. Nous reviendrons alors sur la seconde partie de la

procédure GLAD.

3.2.3 Comparaison avec la méthode GLAD

Le package GLAD met en œuvre la méthode développée par Hupé *et al.* (2004). Celui-ci contient des données publiques (cf. Snijders *et al.* (2001)) issues de 15 cellules humaines dont les caryotypes sont connus. La phase d'hybridation est faite à partir de 2276 sondes, ce qui donne (aux valeurs manquantes près parfois) autant de points répartis sur l'ensemble des 23 chromosomes. Une visualisation globale pour l'ensemble des cellules ainsi que l'ajustement par la méthode GLAD sont donnés en figure 3.15. Un code couleur est attribué aux chromosomes dans cette représentation globale.

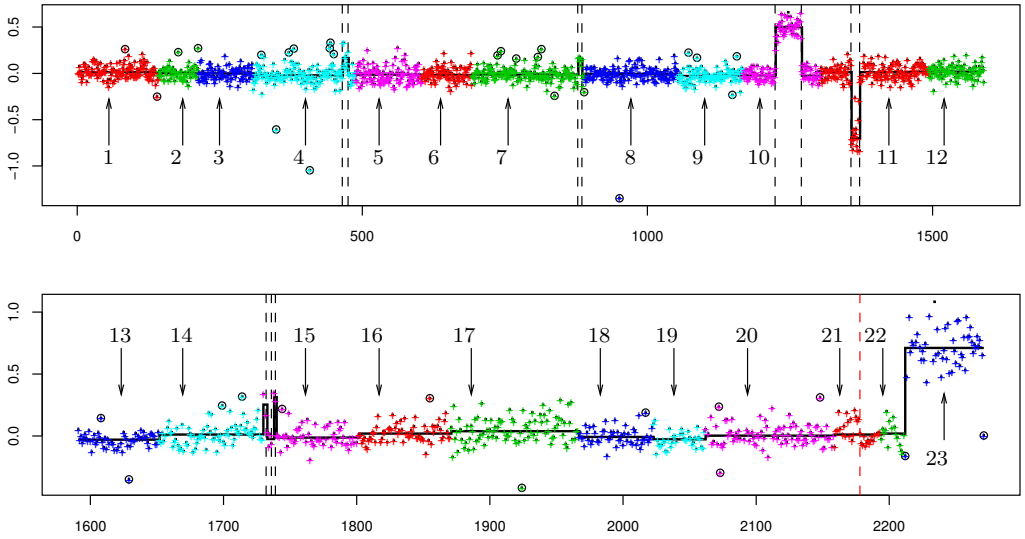


Fig. 3.15 – Profil de la cellule gm05296 et ajustement par le package GLAD.

La procédure mise en œuvre par Hupé *et al.* (2004) est précédée d'une phase de détection de points aberrants qui sont enlevés avant la mise en route : ces points sont entourés d'un cercle sur la figure 3.15. A l'issue de la première phase de la méthode GLAD, on dispose, pour chaque chromosome, d'un ensemble de clusters au sein desquels l'ajustement est constant. Les points de rupture séparant ces clusters sont figurés par les traits pointillés verticaux sur la figure. La suite de la méthode vise à regrouper certains clusters. Pour cela, les auteurs proposent d'effectuer une classification hiérarchique sur cet ensemble de clusters, classification pondérée par le nombre d'observations appartenant à chaque ensemble. Ils utilisent ensuite l'estimation des paramètres associés aux possibilités de coupure du dendogramme comme autant de valeurs d'entrée pour le calcul d'une fonction de vraisemblance des paramètres pénalisée par l'importance des sauts. Le nombre de segments associé à la valeur maximale de la vraisemblance est finalement retenu. En revanche, il n'est pas précisé comment fixer le coefficient réglant l'importance du terme de pénalité. Ainsi, dans la figure 3.15, le point de rupture représenté par le trait vertical rouge, qui marque initialement la séparation entre deux clusters, disparaît, les deux ensembles étant réunis dans le même.

Ajoutons que finalement, la valeur assignée à un segment particulier est la médiane des observations de ce segment. Il faut également souligner qu'en raison du biais induit par la technologie

bi-couleur utilisée, il arrive que les observations ne soient pas centrées sur 0. Cela peut être en particulier le cas pour des chromosomes de petite taille qui sont associés à un faible nombre de sondes (le chromosome 23 pour l'exemple représenté). Précisons enfin que, si le graphe 3.15 donne une vue globale des résultats, la méthode s'applique bien sûr chromosome par chomosome.

Nous proposons d'appliquer l'algorithme I.I.R. sur ces données. Nous utilisons essentiellement la méthode comme une méthode de segmentation ici. Certains segments d'intérêt peuvent se trouver au début ou à la fin des profils CGH. La méthode I.I.R. présentant certains effets de bord, nous avons décidé d'ajouter des points (en nombre équivalent à 5% de l'ensemble des points) de part et d'autre des données dont l'ordonnée commune est égale à la médiane des observations. Par simulation, nous avons vérifié auparavant que cela permettait d'atténuer les effets de bord et ne semblait pas déteriorer la détection d'éventuelles zones d'intérêt situées aux extrêmes. Par ailleurs, le package GLAD fournit les points considérés comme aberrants : nous décidons de les remplacer par la médiane des observations pour le chromosome avant de lancer l'algorithme. Ceci est illustré en figure 3.16 : les points aberrants sont figurés en noir ; ils sont remplacés par les points représentés en rouge.

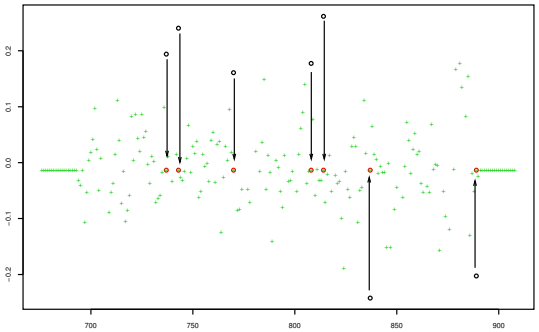


Fig. 3.16 – Préparation des données en vue de l'application de I.I.R. Chromosome 7, cellule gm05296.

Parmi les critères programmés pour notre algorithme, c'est le critère BIC qui semble le plus adapté à la situation. Les critères portant sur les contraintes de formes supposent de connaître le nombre de segments par chromosome, ce qui n'est pas le cas ici. Les autres critères pénalisés donnent des ajustements qui paraissent trop variables, en particulier trop proches de certains points extrêmes. Notre ajustement sépare les observations en plusieurs ensembles au sein desquels l'ajustement est constant. On peut voir ces ensembles comme des clusters dont les bords sont autant de points de rupture à étudier. Il serait également possible de mettre en œuvre une procédure de type classification hiérarchique pour réunir au sein de mêmes groupes les ensembles associés à des valeurs non significativement différentes, mais nous proposons plutôt la règle simple qui permet d'obtenir des résultats assez semblables : nous estimons la variance des observations en calculant, pour le chromosome considéré

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 . \tag{3.10}$$

Nous prenons ensuite "l'intervalle de confiance" centré sur la médiane des observations

$$\mathrm{med}(y) \pm \hat{\sigma} \tag{3.11}$$

et retenons les segments qui sortent de cette bande.

L'ajustement fourni par GLAD après l'étape finale de clustering décrite plus haut et celui obtenu par la méthode I.I.R. sont représentés en figure 3.17 dans un cas particulier.

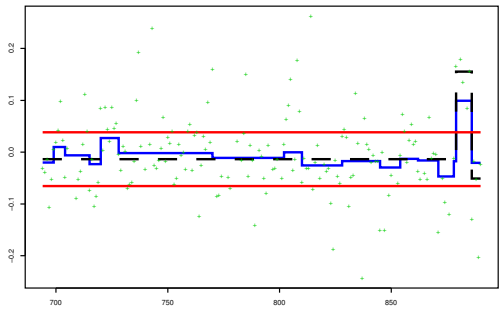


Fig. 3.17 – Ajustement par GLAD en pointillés, par I.I.R. en traits pleins. Cas du chromosome 7, cellule gm05296.

Nous avons aussi représenté en figures 3.18 et 3.19 les ajustements par les deux méthodes pour l'ensemble des chromosomes de la cellule gm05296. Sur cet exemple, la méthode I.I.R. permet d'identifier les mêmes segments. En revanche, pour le chromosome 22, une zone est détectée par notre méthode alors qu'elle ne l'est pas par la méthode GLAD (zone entourée sur la figure 3.19). Il n'est pas sûr que cela soit véritablement un problème. En effet, un usage fréquent pour l'aide au diagnostic avec ce type de données est d'analyser point par point chaque nuage. Le praticien examine la représentation des observations et choisit ensuite un certain nombre de segments qu'il estime susceptibles d'être associés à des régions altérées. Le médecin analyse alors les bornes des segments proposés pour identifier les gènes compris entre ces bornes et voir s'il s'agit de variations en nombres de copies connues.

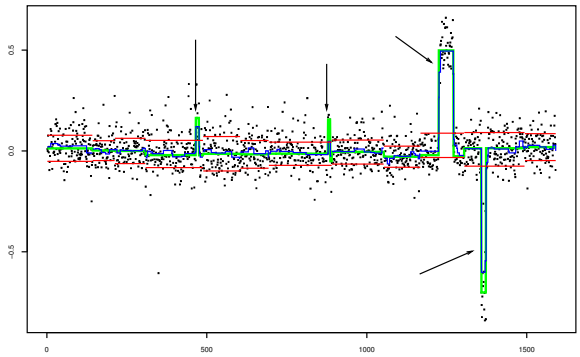


Fig. 3.18 – Ajustements par GLAD en vert, par I.I.R. en bleu pour les chromosomes 1 à 12, cellule gm05296.

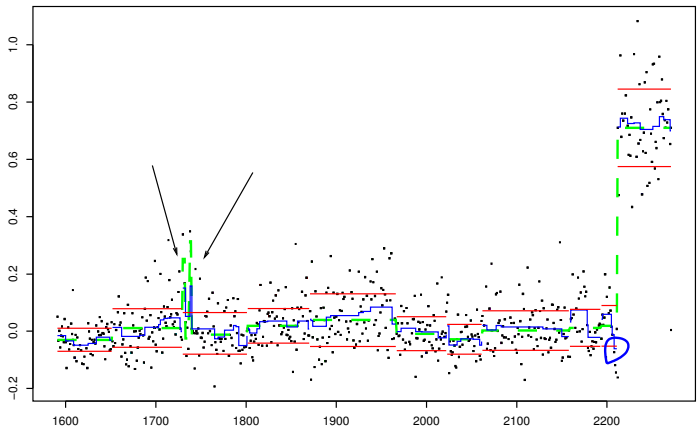


Fig. 3.19 – Ajustements par GLAD vert, par I.I.R. en bleu pour les chromosomes 13 à 23, cellule gm05296.

Pour conclure, disons que la méthode I.I.R. semble fournir des résultats assez encourageants sur ce jeu de données. Sur l'ensemble des 15 cellules, seuls 5 points de rupture sur les 35 donnés par GLAD sont "oubliés". La méthode GLAD suppose de régler un grand nombre de paramètres et le réglage de bon nombre d'entre eux semble relever de critères empiriques. Notre idée présente l'intérêt d'une mise en œuvre nettement plus directe puisque, en dehors du choix du critère d'arrêt, seule la largeur de l'intervalle de confiance repose sur un choix subjectif. On peut ajouter d'ailleurs qu'en diminuant légèrement la largeur de cet intervalle, on retrouve la plupart des points laissés de côté avec l'intervalle précédent.

3.2.4 Application à la recherche de régions du génome incriminées dans le retard mental chez les enfants

Des données de type puces CGH sont utilisées au Laboratoire de génétique médicale du CHU de Rennes pour déterminer des régions du génome susceptibles d'être liées à certains troubles mentaux chez les enfants (Jaillard *et al.* (2010)). Par rapport aux données provenant de tumeurs cancéreuses où les rapports sont calculés à partir de mesures faites sur les cellules d'un même individu, la procédure d'acquisition des données est ici un peu différente puisque le calcul se base sur la comparaison de cellules prélevées chez des enfants présentant des troubles à des cellules prélevées sur des enfants n'en présentant pas.

Le Laboratoire nous a fourni les données pour les 22 chromosomes d'intérêt pour 5 individus nommés :

```
"X2273RM_noise"  "X3543X"   "X9574G"   "X1048G"  "X3580X"
```

Nous disposons ainsi de 110 jeux de données dont la taille varie d'environ 300 à 14500. Les données ont été bornées manuellement selon la procédure décrite en section précédente : le praticien analyse chaque nuage point par point et choisit un certain nombre de segments (une vingtaine plus ou moins selon les cas) qui dont ensuite analysés en comparaison à des bases de données connues. C'est une procédure longue et coûteuse où finalement on ne retient au maximum qu'un segment par échantillon. Ces données ont déjà été analysées au sein du Laboratoire. Nous avons appliqué notre méthode "en aveugle", c'est-à-dire sans connaître les bornes d'intérêt. En

revanche, nous savons que certaines peuvent se situer en début ou fin de segment. Nous ajoutons donc des points de part et d'autre des données selon le même schéma qu'en section précédente.

La procédure mise en place est très comparable à ce que nous avons fait dans la Section 3.2.3. Cependant, nous souhaitons pouvoir relancer plusieurs fois la méthode sur le même échantillon. Le premier objectif est d'augmenter la robustesse de la détection des segments et le second d'associer en quelque sorte une "confiance" à un segment détecté. Si un segment apparaît dans 95% des réplications, sans doute doit-on avoir plus confiance dans le fait qu'il correspond à une éventuelle variation du nombre de copies que s'il n'apparaît que dans 5% des cas. Nous insérons ainsi une étape de sélection aléatoire d'un sous-échantillon, sur lequel la méthode est appliquée pour nous relançons la procédure :

- on sélectionne aléatoirement un échantillon d'apprentissage contenant 95% des points (hors points ajoutés aux extrêmes),
- on applique la régression isotonique itérée avec le critère d'arrêt BIC,
- on estime la variance sur cet échantillon d'apprentissage :

$$\hat{\sigma}^2 = \frac{1}{n_{app}} \sum_{i=1}^{n_{app}} \left(\hat{y}_i - y_i\right)^2 ,$$

- on considère l'intervalle centré sur la médiane des observations :

$$\mathrm{med}(y) \pm \hat{\sigma},$$

- on marque les bornes des segments sortant des intervalles précédents,
- on réplique $N = 500$ fois la méthode.

A l'issue de cette procédure, nous disposons, pour chaque réplication, des bornes des segments qui ont été éventuellement détectés. Suivant les cas, nous mettons en évidence entre 0 et 2 segments. Les cas où l'on obtient 2 segments sont peu nombreux, aussi ne figurent-ils pas dans le tableau 3.4 qui regroupe les résultats. Il est donc clair que la procédure ne permet pas de détecter l'ensemble des variations pathologiques et non pathologiques que le praticien. analyse. En revanche, il convient d'analyser si les variations pathologiques qui nous ont été fournies par le Laboratoire sont détectées.

Tableau 3.4 – Détection de variations pathologiques par la méthode I.I.R.. Chromosomes 1 à 11 en haut, 12 à 22 en bas.

	1	2	3	4	5	6	7	8	9	10	11
X3543X	0	0	0	0	0	0	0	0	0	0	0
X9574G	0	0	0	0	0	0	0	0	0	0	0
X1048G	1	0	0	0	0	0	0	0	0	0	0
X3580X	0	0	0	0	0	0	0	1	0	0	0
X2273RMnoise	0	0	0	0	0	0	0	1	0	0	0

	12	13	14	15	16	17	18	19	20	21	22
X3543X	0	0	0	1	0	0	0	0	0	0	0
X9574G	0	0	0	0	0	0	0	0	0	0	1
X1048G	0	0	0	1	0	0	0	0	0	0	0
X3580X	0	0	0	1	0	0	0	0	0	0	0
X2273RMnoise	0	0	0	0	0	0	0	1	0	0	0

En rouge sont indiquées les variations pathologiques que notre méthode retrouve. Par exemple, pour l'individu X9574G, nous mettons en évidence un segment dont les bornes correspondent à celles retenues par l'équipe médicale. Ce cas est représenté en figure 3.20. Ajoutons que la distribution des bornes est très peu variable (les abscisses retenues d'une réplication à l'autre sont identiques dans plus de 95% des cas).

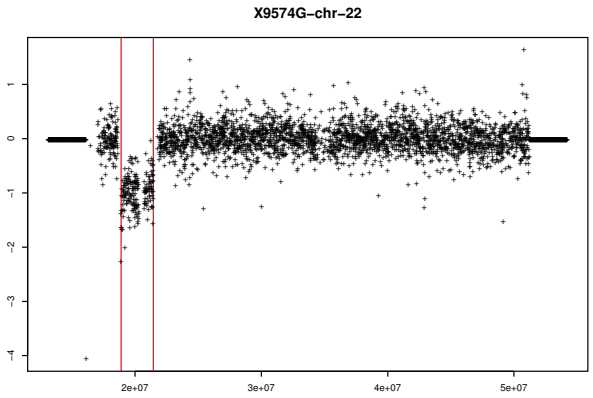


Fig. 3.20 – Une variation pathologique détectée.

En bleu sont indiqués les cas où nous donnons un segment correspondant à une variation non pathologique. Notons que ces segments font également partie de ceux qui ont été analysés. Un exemple est donné figure 3.21.

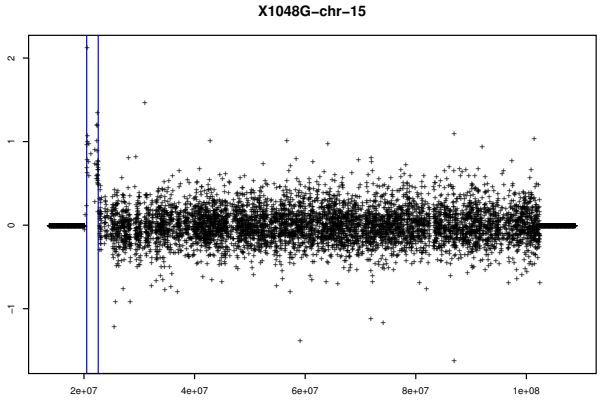


Fig. 3.21 – Une variation non pathologique.

Enfin, le cas mentionné en vert dans le tableau correspond à la seule variation pathologique qui n'a pas été détectée. L'examen du graphe montre qu'il s'agit d'une région située en fin de séquence et que le nuage semble présenter un niveau de bruit assez important (cf. figure 3.22).

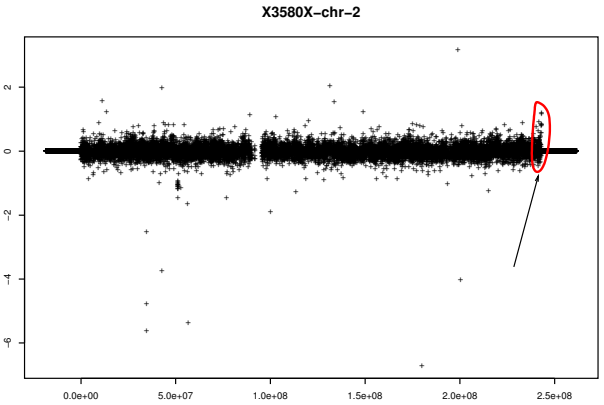


Fig. 3.22 – Une variation non détectée.

La méthode fournit donc des résultats assez encourageants quant à la détection de variations pathologiques puisque cinq des six identifiées ont été mises en évidence. Le fait d'en avoir laissé échapper une pose en revanche problème. En effet, les médecins préfèrent disposer d'un ensemble assez large de segments contenant toutes les zones d'intérêt plutôt que d'un ensemble plus restreint qui ne les contienne pas toutes.

On peut imaginer certaines améliorations de la méthode. Lorsque l'on représente l'ajustement sur le cas précédent, sans représenter le nuage des points, on observe que la zone d'intérêt correspond à la valeur où l'ajustement est le plus grand (cf. figure 3.23 avec l'ajustement en rouge et le début de la zone d'intérêt pointé par la flêche bleue). Une piste serait de prendre en compte sans distinction les abscisses des bornes des segments comme autant de points de rupture possibles et de laisser le soin au médecin de trancher sur leur pertinence. Dans ce cas cependant, les segments les plus élevés (ou les plus bas) seraient à examiner prioritairement. On peut également imaginer

d'affecter à chaque segment la médiane des valeurs qui y figurent, comme pour GLAD (courbe en traits pointillés verts). En général, les valeurs ajustées sont en valeurs absolues plus grandes qu'avec l'ajustement par I.I.R., ce qui facilite la détection et rejoint l'idée que la méthode peut être utilisée ici comme une procédure de segmentation plus que de régression. On peut aussi imaginer de tracer une zone de confiance en considérant deux bandes de part et d'autre de la ligne horizontale figurant la médiane des y_i et d'amplitude $2 \times \sqrt{\text{var}(\hat{Y})}$ (traits pointillés noirs). Ces idées sont actuellement à l'étude.

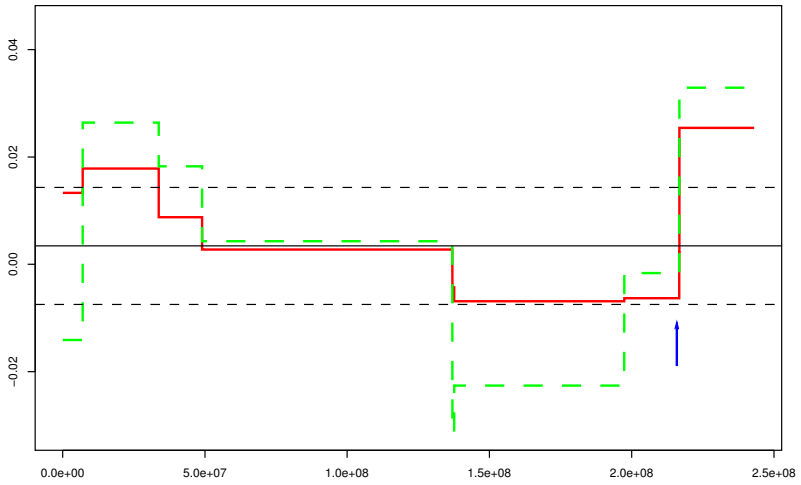


Fig. 3.23 – Pistes d'améliorations possibles.

3.3 Application à la régression unimodale

Dans certaines applications, il arrive que l'on suppose que la fonction de régression est unimodale, c'est-à-dire croissante puis décroissante. Par exemple, dans Schmidt (1993), on explique que la capacité de développement d'une variété d'arbre dans un secteur donné croît avec la densité avant de décroître en raison notamment de l'appauvrissement de la quantité de lumière disponible pour chaque arbre. Dans Frisén (1986), on cite une étude où des images télévisées contenant un motif particulier ont été présentées à des patients. La plupart du temps, les patients ne peuvent pas distinguer le motif si le réglage du constraste de l'image est trop fort ou si le réglage est trop faible et on suppose que la sensibilité de l'œil est une fonction unimodale de l'intensité du contraste. Outre l'estimation de la fonction de régression, un enjeu important est de localiser la position du maximum. Dans l'exemple précédent, le contraste associé à la plus grande acuité visuelle donne une mesure de la densité de recepteurs de la rétine, densité qui ne peut pas être mesurée directement sur l'œil.

Nous considérons ici que la fonction $r : [0,1] \to \mathbb{R}$ du modèle

$$Y = r(X) + \varepsilon$$

est unimodale. Autrement dit, on suppose qu'il existe α dans $]0,1[$ tel que

$$\forall x, x' \in [0,\alpha] : x \leq x' \Rightarrow r(x) \leq r(x') \qquad \text{et} \qquad \forall x, x' \in [\alpha,1] : x \leq x' \Rightarrow r(x) \geq r(x').$$

Plusieurs procédures existent pour estimer de telles fonctions. Certaines sont paramétriques comme celle consistant à ajuster des polynômes de faible degré (cf. Hotelling (1941)), d'autres non paramétriques, basées par exemple sur l'utilisation de splines de lissage (cf. Silverman (1985)). Notons également la procédure adaptative proposée par Reboul (2005).

Si la position du mode α est connue, il est bien évident que l'on peut effectuer une régression isotonique avant ce mode et une régression antitonique après pour construire un estimateur de la fonction. La procédure décrite par Turner & Wollan (1997) s'inspire de celle-ci dans le cas d'un mode inconnu. Disposant d'observations $(x_i, y_i)_{i=1,\cdots,n}$ de (X, Y) rangées dans l'ordre des x_i, l'idée est la suivante :

- on envisage tour à tour chaque x_i comme mode potentiel,
- on effectue une régression isotonique sur les valeurs précédentes et une régression antitonique sur les suivantes,
- On calcule l'erreur quadratique correspondant à cet ajustement

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 ,$$

- on retient enfin comme estimateur de α, l'abscisse x_i associée à l'erreur d'ajustement la plus faible.

Il faut noter que l'on n'a pas en général unicité de la solution, ce qui n'est pas surprenant dans la mesure où, si l'ensemble des fonctions unimodales est bien un cône, ce n'est pas un convexe. On n'hérite donc pas de la propriété d'unicité de la projection sur cet ensemble.

Dans leur article, Turner & Wollan (1997) montrent que la méthode donne un estimateur consistant de la position du mode et le comparent à d'autres techniques sur des exemples simulés. Il faut bien avouer que leurs résultats ne permettent pas cependant de prouver la supériorité indiscutable de cette approche. La principale conclusion est que la performance de tel ou tel estimateur semble liée à de nombreux paramètres comme la forme de la fonction sous-jacente ou le niveau de bruit appliqué.

Le coût opératoire de la procédure est problématique. La régression isotonique sur n points pouvant être réalisée en $\mathcal{O}(n)$ opérations (cf. Best (1984)), cette approche nécessite $\mathcal{O}(n^2)$ opérations ce qui devient vite très lourd pour de grands échantillons. Il semble ainsi que la majorité des développements de la régression unimodale faisant intervenir la régression isotonique soient d'ordre algorithmique. Citons les travaux de Geng & Shi (1990) et ceux de Stout (2008). Ce dernier réorganise de façon astucieuse les étapes de l'algorithme PAVA pour obtenir une complexité linéaire $\mathcal{O}(n)$ pour le cas des coûts L_2 et L_∞ et une complexité quasi-linéaire en $\mathcal{O}(n \log n)$ pour le coût L_1. Récemment, Liu & Ubhaya (2009) ont proposé un algorithme au coût comparable dans le cas unimodal mais avec la contrainte supplémentaire que les valeurs ajustées soient entières. Ils s'appuient sur des applications possibles dans Goldstein & Kruskal (1976) et sur un algorithme qu'avait auparavant développé l'un des deux auteurs (cf. Ubhaya (1987)).

Dans ce qui suit, nous proposons quelques perspectives pratiques de la régression isotonique itérée. Après avoir explicité la régression isotonique sur une fonction unimodale en Section 3.3.1, nous montrons en Section 3.3.2, que la régression isotonique itérée appliquée sur une telle fonction donne un estimateur qui présente un intervalle modal, intervalle dont les bornes encadrent le mode de r. En Section 3.3.3, nous montrons, sur quelques exemples simulés, que la méthode fournit un intervalle incluant le mode avec une bonne confiance. Nous proposons enfin en Section 3.3.4 d'étendre cette idée à la localisation de plusieurs modes.

3.3.1 Régression isotonique sur une fonction unimodale

Nous considérons dans cette section le modèle

$$Y = r(X) + \varepsilon$$

avec une fonction r continue sur $[0,1]$, d'intégrale nulle, strictement croissante sur $[0,\alpha]$ puis strictement décroissante sur $[\alpha,1]$. Le mode de r est donc situé à l'abscisse α.

Nous allons expliciter la régression isotonique d'une telle fonction, autrement dit, expliciter la solution au problème de minimisation

$$\underset{u\in\mathcal{C}^+}{\operatorname{argmin}} \int_0^1 \left(r(x)-u(x)\right)^2 \mu(dx).$$

Dans toute cette section, nous supposons X distribuée selon la loi uniforme sur $[0,1]$. Le problème précédent s'écrit donc

$$\underset{u\in\mathcal{C}^+}{\operatorname{argmin}} \int_0^1 \left(r(x)-u(x)\right)^2 dx. \tag{3.12}$$

Pour ce faire, nous montrons en Proposition 3.1 qu'il existe un réel $\beta \in]0,\alpha[$ dont l'image par r est égale à la valeur moyenne de r sur $[\beta,1]$:

$$r(\beta) = \frac{1}{1-\beta}\int_\beta^1 r(x)dx.$$

En Proposition 3.2, nous montrons que la fonction égale à r jusqu'à l'abscisse β puis constante égale à $r(\beta)$ sur le reste de l'intervalle est la solution de (3.12). Ainsi, nous montrons que la fonction [2]

$$u_+ \left\{ \begin{array}{lll} [0,1] & \to & \mathbb{R} \\ x & \mapsto & u_+(x) = r(x)\mathbb{1}_{[0,\beta]}(x) + r(\beta)\mathbb{1}_{]\beta,1]}(x) \end{array} \right.$$

est la régression isotonique de r. On représente un exemple en figure 3.24.

2. Nous reprenons les notations de la Section 2.3 où l'on a étudié la consistance de l'estimateur de régression isotonique itérée.

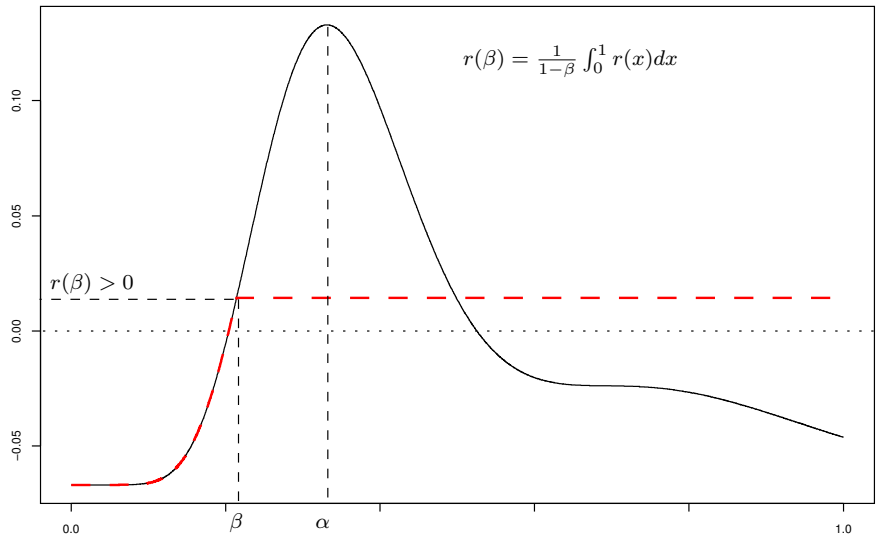


Fig. 3.24 – Régression isotonique sur une fonction unimodale.

Proposition 3.1
Soit r une fonction définie et continue sur $[0,1]$, strictement croissante sur $[0,\alpha]$ et strictement décroissante sur $[\alpha,1]$. On suppose

$$r(0) < 0 \qquad et \qquad \int_0^1 r(x)d(x) = 0,$$

alors il existe un unique réel $\beta \in]0,\alpha[$ tel que

$$r(\beta) = \frac{1}{1-\beta}\int_\beta^1 r(x)d(x).$$

On a de plus $r(\beta) > 0$.

Preuve
Considérons la fonction

$$\psi : x \in]0,\alpha[\mapsto (1-x)r(x) - \int_x^1 r.$$

La fonction r est strictement décroissante et continue sur $[\alpha,1]$ donc $r(\alpha) > \frac{1}{1-\alpha}\int_\alpha^1 r(x)d(x)$. Ainsi

$$\psi(\alpha) = (1-\alpha)r(\alpha) - \int_\alpha^1 r(x)dx > 0.$$

La fonction ψ est continue et comme $\psi(0) = r(0) < 0$, il existe $\beta \in]0,\alpha[$ qui annule ψ donc tel que

$$r(\beta) = \frac{1}{1-\beta}\int_\beta^1 r(x)dx.$$

Remarquons que β est tel que $r(\beta) > 0$. En effet, si $r(\beta) \leq 0$ alors $\int_\beta^1 r(x)dx \leq 0$. Or on a également $\int_0^\beta r(x)d(x) < 0$ puisque r croît sur $[0,\beta]$ et ceci contredit l'hypothèse $\int_0^1 r(x)dx = 0$. L'existence est démontrée.

Pour montrer l'unicité de β, nous introduisons

$$B = \{\beta \in [0,1] : \psi(\beta) = 0\} \qquad \text{et} \qquad \beta^\star \text{ la borne inférieure de } B.$$

Comme ψ est continue, $\psi(\beta^\star) = 0$. Montrons que ψ est strictement croissante sur $[\beta^\star, \alpha]$. Considérons deux réels x_1 et x_2 avec $\beta^\star \leq x_1 < x_2 \leq \alpha$. Nous avons

$$\begin{aligned}\psi(x_1) - \psi(x_2) &= (1-x_1)r(x_1) - (1-x_2)r(x_2) - \int_{x_1}^1 r(x)dx + \int_{x_2}^1 r(x)dx \\ &= (1-x_1)r(x_1) - (1-x_2)r(x_2) - \int_{x_1}^{x_2} r(x)dx.\end{aligned}$$

Comme nécessairement $r(\beta^\star) > 0$ et que r est strictement croissante sur$[x_1, x_2] \subseteq [\beta^\star, \alpha]$, nous avons

$$(x_2 - x_1)r(x_1) < \int_{x_1}^{x_2} r(x)dx < (x_2 - x_1)r(x_2).$$

On en déduit

$$\begin{aligned}\psi(x_1) - \psi(x_2) &< (1-x_1)r(x_1) - (1-x_2)r(x_2) - (x_2 - x_1)r(x_1) \\ &< (1-x_2)\left(r(x_1) - r(x_2)\right) \\ &< 0\end{aligned}$$

donc ψ est strictement croissante $[\beta^\star, \alpha]$. Nous ne pouvons donc avoir $\beta > \beta^\star$ tel que $\psi(\beta) = 0$: une telle valeur est unique sur $[0, \alpha]$. □

Remarque
Nous supposons r strictement croissante avant α car cela allège l'étude de l'unicité. Cependant le seul cas qui contredit l'unicité est celui où l'on aurait $\beta^\star$ tel que $r(\beta^\star) = \frac{1}{1-\beta^\star}\int_{\beta^\star}^1 r(x)dx$ puis r constante sur un intervalle B de borne inférieure $\beta^\star$. Il est en effet facile de vérifier qu'alors

$$\forall t \in B, \qquad r(t) = \frac{1}{1-t}\int_t^1 r(x)dx$$

ce qui confère à tout t de B la propriété évoquée. Ainsi, on aurait pu, dans la Proposition 3.1, supposer r croissante avant α sans imposer qu'elle le soit strictement, et préciser, qu'à part le cas particulier précédent, on a bien unicité de β.

Proposition 3.2 (Régression isotonique d'une fonction unimodale)
Soit r une fonction continue définie sur $[0,1]$, strictement croissante sur $[0,\alpha]$ et strictement décroissante sur $[\alpha, 1]$. On suppose $r(0) < 0$ et $\int_0^1 r(x)d(x) = 0$, alors la fonction u_+ définie par

$$u_+(x) = r(x)\mathbb{1}_{[0,\beta]}(x) + r(\beta)\mathbb{1}_{]\beta,1]}(x)$$

où β est la valeur introduite en Proposition 3.1, est la régression isotonique de r.

Preuve
On s'appuie sur le résultat de Anevski & Soulier (2011) que nous avons rappelé en page 18. On introduit la fonction cumulative de r :

$$R : t \mapsto \int_0^t r(x)dx.$$

On rappelle que le Greatest Convex Minorant de R est la fonction convexe maximale dominée par R (cf. figure 3.25). Formellement, c'est la fonction R_+ définie par

$$R_+ = \text{argmax}\,\{H \text{ convexe et } \forall t \in [0,1], H(t) \leq R(t)\}$$

et la dérivée à gauche en tout point de cette fonction caractérise la régression isotonique de r :

$$(R_+)' = \underset{u \in \mathcal{C}^+}{\text{argmin}} \int_0^1 (r(x) - u(x))^2 \, dx.$$

On considère la fonction cumulative de u_+ :

$$U_+ : t \mapsto \int_0^t u_+(x)dx.$$

U_+ est convexe puisque u_+ est croissante. Sur $[0, \beta]$, elle correspond nécessairement à la fonction convexe maximale dominée par R puisqu'elle coïncide avec R qui est elle-même convexe sur cet intervalle. Ainsi

$$\forall t \in [0, \beta] : \qquad U_+(t) = R_+(t) = R(t).$$

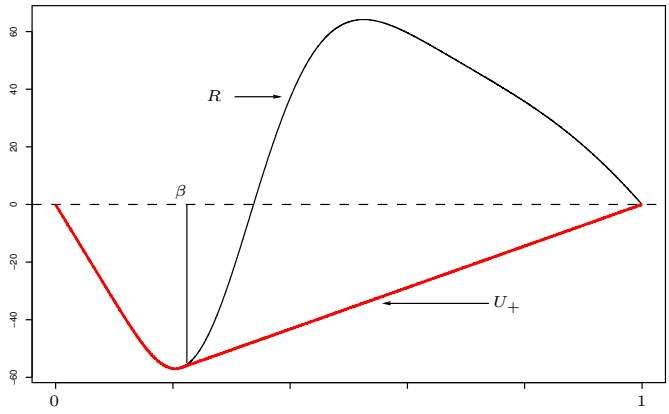


Fig. 3.25 – Régression isotonique sur une fonction unimodale : représentation cumulée.

u_+ est constante sur $[\beta, 1]$ et comme u_+ est d'intégrale nulle par construction, la courbe de U_+ est, sur cet intervalle, un segment de droite joignant $(\beta, U_+(\beta))$ à $(1, 0)$. La régression isotonique de r est nécessairement d'intégrale nulle donc la courbe de R_+, en plus de passer par $(\beta, U_+(\beta))$, passe également par $(1, 0)$. Il est clair qu'une fonction qui joindrait ces deux points et qui passerait par un point intermédiaire situé strictement au-dessus de la courbe de U_+ ne serait pas convexe. Ainsi,

$$\forall t \in [\beta, 1] : \qquad R_+(t) = U_+(t)$$

et $t \mapsto (U_+)'(t) = u_+(t)$ est bien la régression isotonique de r. □

Nous avons l'analogue des Propositions 3.1 et 3.2 pour la régression antitonique. La situation est représentée en figure 3.26.

Corollaire 3.1
Soit r une fonction définie et continue sur $[0,1]$, strictement croissante sur $[0,\alpha[$ et strictement décroissante sur $[\alpha,1]$. On suppose

$$r(1) < 0 \qquad et \qquad \int_0^1 r(x)dx = 0.$$

Alors il existe une unique réel $\gamma \in]\alpha,1[$ tel que

$$r(\gamma) = \frac{1}{\gamma}\int_0^\gamma r(x)dx.$$

On a de plus $r(\gamma) > 0$.

Corollaire 3.2 (Régression antitonique d'une fonction unimodale)
Soit r une fonction continue définie sur $[0,1]$, strictement croissante sur $[0,\alpha]$ strictement décroissante sur $[\alpha,1]$ telle que $r(1) < 0$ et $\int_0^1 r(x)dx = 0$, alors la fonction b_- définie sur $[0,1]$ par

$$b_-(x) = r(\gamma)\mathbb{1}_{[0,\gamma[}(x) + r(x)\mathbb{1}_{[\gamma,1]}(x),$$

où γ est la valeur introduite en Corollaire 3.1, est la régression antitonique de r, i.e.

$$b_- = \underset{b\in\mathcal{C}^-}{\operatorname{argmin}} \int_0^1 \left(r(x) - b(x)\right)^2 dx.$$

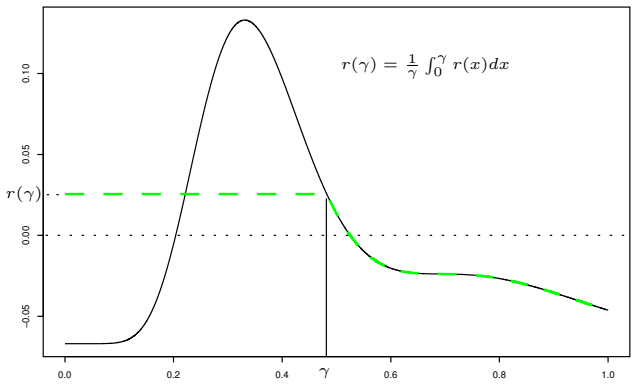


Fig. 3.26 – Régression antitonique sur une fonction unimodale.

3.3.2 Régression isotonique itérée sur une fonction unimodale

Dans cette section, nous voyons comment les résultats précédents permettent, lorsque l'on déroule les étapes de l'algorithme de régression isotonique itérée, de mettre en évidence deux suites adjacentes (β_k) et (γ_k) qui convergent vers le mode α. Nous considérons une fonction r définie et continue sur $[0,1]$, strictement croissante sur $[0,\alpha]$ puis strictement décroissante sur $[\alpha,1]$. Nous supposons qu'elle est d'intégrale nulle et qu'elle vérifie $r(0) < 0$, ce qui permet de se reporter directement aux résultats de la section précédente. Dans le cas contraire, i.e. avec $r(0) \geq 0$, on a nécessairement $r(1) < 0$ pour respecter la condition portant sur la nullité de l'intégrale. On peut alors se reporter au cas décrit ci-dessous en considérant la fonction $\tilde{r}$ définie par $\tilde{r}(x) = r(1-x)$.

Nous reprenons les notations utilisées pour l'étude de la consistance : travaillant dans le cadre fonctionnel, nous notons les estimations à l'issue de la $k^{\text{ème}}$ étape $u^{(k)}$, $b^{(k)}$ et $r^{(k)}$. Plus précisément, le résultat obtenu s'exprime de la manière suivante :

Théorème 3.1
Soit pour $k \geq 1$, les fonctions $u^{(k)}$ et $b^{(k)}$ résultant de l'application de l'algorithme I.I.R sur la fonction r. Alors, pour tout $k \geq 1$, il existe un unique réel $\beta_k \in [0, \alpha[$ tel que

$$u^{(k)}(\beta_k) = \frac{1}{1-\beta_k} \int_{\beta_k}^{1} \left(r - b^{(k-1)}\right)(x)dx$$

et un unique réel γ_k tel que

$$b^{(k)}(\gamma_k) = \frac{1}{\gamma_k} \int_{0}^{\gamma_k} \left(r - u^{(k)}\right)(x)dx.$$

La suite (β_k) est croissante vers α et la suite (γ_k) est décroissante vers α.

L'existence des suites (β_k) et (γ_k) ainsi que la plupart des éléments utiles se montrent par récurrence. Nous détaillons dans la suite les deux premières étapes de l'algorithme, la généralisation des résultats se faisant sur le même principe. L'existence des termes β_k et γ_k provient directement des Propositions 3.1 et 3.2 lorsque l'on applique la régression isotonique et de leur corollaires lorsque l'on applique la régression antitonique. Pour établir la convergence des suites (β_k) et (γ_k), nous montrons que pour tout $k \geq 1$,

$$r(\beta_k) < r(\gamma_k) < r(\beta_{k+1}) < r(\gamma_{k+1}) \tag{3.13}$$

avec pour tout $k \geq 1$, $\beta_k \leq \alpha$ et $\gamma_k \geq \alpha$. Ainsi (β_k) et (γ_k) sont deux suites convergentes et en passant à la limite, on a

$$r(\beta_\infty) = r(\gamma_\infty). \tag{3.14}$$

Nous montrons également que pour tout $k \geq 1$,

$$\gamma_k r(\gamma_k) - \beta_k r(\beta_k) = \int_{\beta_k}^{\gamma_k} r(x)dx. \tag{3.15}$$

Le passage à la limite sur cette égalité entraîne

$$\gamma_\infty r(\gamma_\infty) - \beta_\infty r(\beta_\infty) = \int_{\beta_\infty}^{\gamma_\infty} r(x)dx$$

donc grâce à (3.14)

$$(\gamma_\infty - \beta_\infty) r(\beta_\infty) = \int_{\beta_\infty}^{\gamma_\infty} r(x)dx.$$

Il est alors clair, r étant supposée strictement monotone de part et d'autre de α, que cette dernière égalité ne peut se produire qu'à condition que

$$\beta_\infty = \gamma_\infty = \alpha.$$

Donnons le détail des deux premières étapes de l'algorithme.

Etape 1 Le déroulement de cette étape est illustré en figure 3.27. Nous effectuons tout d'abord la régression isotonique sur r. En vertu de la Proposition 3.1, il existe un unique réel noté $\beta_1 \in]0, \alpha[$ tel que

$$r(\beta_1) = \frac{1}{1-\beta_1} \int_{\beta_1}^{1} r(x)dx$$

et la régression isotonique de r est définie par

$$u^{(1)}(x) = r(x) \times \mathbb{1}_{[0,\beta_1]}(x) + r(\beta_1) \times \mathbb{1}_{]\beta_1,1]}(x) \tag{3.16}$$

avec comme propriété $r(\beta_1) > 0$. Nous considérons maintenant $r - u^{(1)}$ et lui appliquons la régression antitonique. Nous avons

$$(r - u^{(1)})(x) = 0 \times \mathbb{1}_{[0,\beta_1]}(x) + (r(x) - r(\beta_1)) \times \mathbb{1}_{]\beta_1,1]}(x).$$

C'est une fonction continue, croissante sur $[0, \alpha]$, décroissante sur $[\alpha, 1]$ et d'intégrale nulle. Par ailleurs, elle est nulle en 0 et prend une valeur positive en α donc sa valeur en 1 est nécessairement négative. Nous pouvons donc lui appliquer les Corollaires 3.1 puis 3.2 :

$$\exists! \gamma_1 \in]\alpha, 1[, \qquad (r - u^{(1)})(\gamma_1) = \frac{1}{\gamma_1} \int_0^{\gamma_1} (r - u^{(1)})(x)dx$$

et $(r - u^{(1)})(\gamma_1) > 0$. Ainsi, comme $u^{(1)}(\gamma_1) = r(\beta_1)$, on a $r(\gamma_1) > r(\beta_1)$. La première estimation antitonique est

$$\begin{aligned} b^{(1)}(x) &= (r - u^{(1)})(\gamma_1) \times \mathbb{1}_{[0,\gamma_1[}(x) + (r - u^{(1)})(x) \times \mathbb{1}_{[\gamma_1,1]}(x) \\ &= (r(\gamma_1) - r(\beta_1)) \times \mathbb{1}_{[0,\gamma_1[}(x) + (r(x) - r(\beta_1)) \times \mathbb{1}_{[\gamma_1,1]}(x) \end{aligned}$$

et à l'issue de la première étape, l'estimation $r^{(1)}$ de r est définie par

$$\begin{aligned} r^{(1)}(x) = &(r(x) + r(\gamma_1) - r(\beta_1)) \times \mathbb{1}_{[0,\beta_1]}(x) \\ &+ r(\gamma_1)\mathbb{1}_{]\beta_1,\gamma_1[}(x) \\ &+ r(x) \times \mathbb{1}_{[\gamma_1,1]}(x). \end{aligned}$$

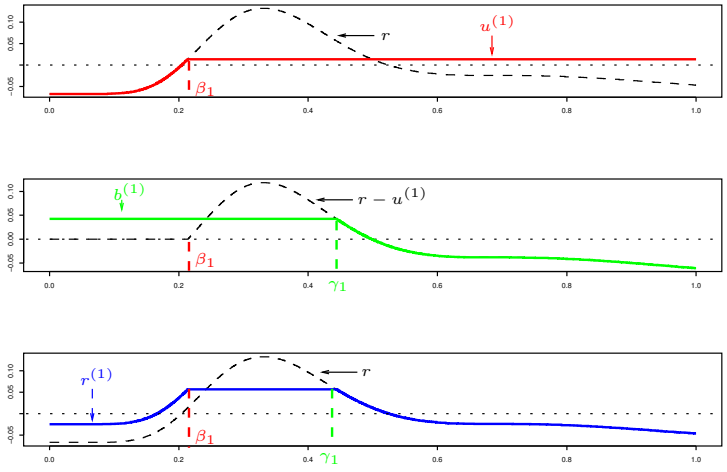


Fig. 3.27 – Première étape.

Etape 2 Le déroulement de la seconde étape est représenté en figure 3.28. Nous appliquons la régression isotonique à

$$\left(r - b^{(1)}\right)(x) = (r(x) - (r(\gamma_1) - r(\beta_1))) \times \mathbb{1}_{[0,\gamma_1[}(x) + r(\beta_1) \times \mathbb{1}_{[\gamma_1,1]}(x).$$

Comme $r(0) < 0$ et $r(\gamma_1) - r(\beta_1) > 0$, la valeur de cette fonction est négative en 0. Son maximum est positif, atteint en α. Elle est d'intégrale nulle donc les conditions d'application de la Proposition 3.1 sont réunies :

$$\exists! \beta_2 \in]0, \alpha[: \qquad (r - b^{(1)})(\beta_2) = \frac{1}{1-\beta_2} \int_{\beta_2}^{1} (r - b^{(1)})(x) dx.$$

Après β_2, la fonction $r - b^{(1)}$ croît jusqu'à α puis décroît jusqu'à la valeur $r(\beta_1)$. Comme $(r - b^{(1)})(\beta_2)$ représente la valeur moyenne de $r - b^{(1)}$ après β_2, nécessairement

$$(r - b^{(1)})(\beta_2) > r(\beta_1).$$

Par conséquent,

$$\begin{aligned} r(\beta_2) &> r(\beta_1) + b^{(1)}(\beta_2) \\ &> r(\beta_1) + r(\gamma_1) - r(\beta_1) \\ &> r(\gamma_1) \end{aligned}$$

et finalement,

$$r(\beta_2) > r(\gamma_1) > r(\beta_1). \tag{3.17}$$

Comme r est croissante avant α, nous déduisons

$$\beta_2 > \beta_1. \tag{3.18}$$

La seconde estimation croissante est alors

$$u^{(2)}(x) = (r(x) - (r(\gamma_1) - r(\beta_1))) \times \mathbb{1}_{[0,\beta_2]}(x) + (r(\beta_2) - (r(\gamma_1) - r(\beta_1)))\, \mathbb{1}_{]\beta_2,1]}(x).$$

Sur le même principe, on applique la régression antitonique à $r - u^{(2)}$. Le Corollaire 3.1 permet de mettre en évidence un réel $\gamma_2 > \alpha$ qui vérifie $r(\gamma_2) > r(\beta_2)$ donc

$$r(\gamma_2) > r(\beta_2) > r(\gamma_1). \tag{3.19}$$

Les équations (3.17) et (3.19) donnent

$$r(\beta_1) < r(\gamma_1) < r(\beta_2) < r(\gamma_2)$$

qui correspond à (3.13) pour $k = 1$. Par ailleurs, comme r décroît après α, on a bien

$$\gamma_2 < \gamma_1. \tag{3.20}$$

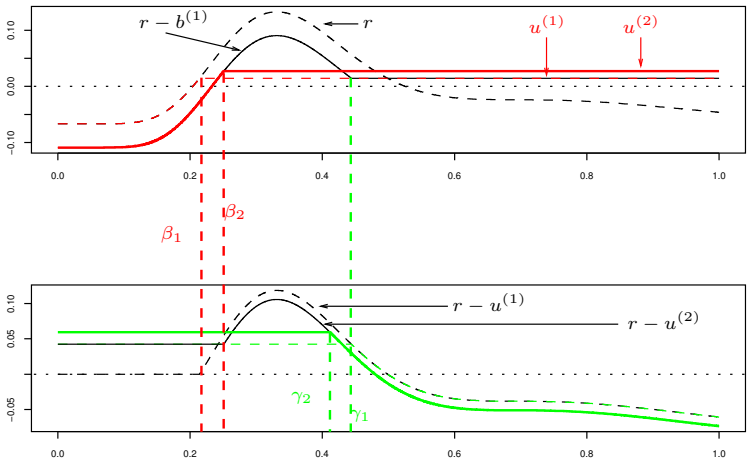


Fig. 3.28 – Deuxième étape.

Vérifions maintenant l'équation (3.15) pour ces premières itérations. Nous reprenons la définition des termes. Ainsi, γ_1 est caractérisé par

$$(r - u^{(1)})(\gamma_1) = \frac{1}{\gamma_1} \int_0^{\gamma_1} (r - u^{(1)})(x)dx.$$

Nous en déduisons en reprenant l'expression de $u^{(1)}$ (cf. équation (3.16))

$$\begin{aligned} r(\gamma_1) - r(\beta_1) &= \frac{1}{\gamma_1} \int_{\beta_1}^{\gamma_1} (r(x) - r(\beta_1))dx \\ &= \frac{1}{\gamma_1} \int_{\beta_1}^{\gamma_1} r(x)dx - \frac{\gamma_1 - \beta_1}{\gamma_1} r(\beta_1), \end{aligned}$$

qui conduit à

$$\gamma_1 r(\gamma_1) - \beta_1 r(\beta_1) = \int_{\beta_1}^{\gamma_1} r(x)dx$$

cas particulier de (3.15) pour $k = 1$. De même, reprenons la caractérisation de γ_2 et introduisons β_2 dans les bornes de l'intégrale par la relation de Chasles :

$$\begin{aligned} \frac{1}{\gamma_2} \int_0^{\gamma_2} (r - u^{(2)})(x)dx &= \frac{1}{\gamma_2} \int_0^{\beta_2} (r - u^{(2)})(x)dx + \frac{1}{\gamma_2} \int_{\beta_2}^{\gamma_2} (r - u^{(2)})(x)dx \\ &= \frac{\beta_2}{\gamma_2}(r(\gamma_1) - r(\beta_1)) + \frac{1}{\gamma_2} \int_{\beta_2}^{\gamma_2} r(x)dx - \frac{\gamma_2 - \beta_2}{\beta_2} u^{(2)}(\beta_2) \\ &= \frac{\beta_2}{\gamma_2}(r(\gamma_1) - r(\beta_1)) + \frac{1}{\gamma_2} \int_{\beta_2}^{\gamma_2} r(x)dx - \frac{\gamma_2 - \beta_2}{\gamma_2}(r(\beta_2) - r(\gamma_1) + r(\beta_1)). \end{aligned}$$

Par conséquent,

$$\begin{aligned} r(\gamma_2) &= u^{(2)}(\gamma_2) + \frac{1}{\gamma_2}\int_0^{\gamma_2} r(x)dx \\ &= r(\beta_2) - r(\gamma_1) + r(\beta_1) + \frac{\beta_2}{\gamma_2}(r(\gamma_1) - r(\beta_1)) + \frac{1}{\gamma_2}\int_{\beta_2}^{\gamma_2} r(x)dx - \frac{\gamma_2 - \beta_2}{\gamma_2}(r(\beta_2) - r(\gamma_1) + r(\beta_1)) \\ &= r(\beta_2)\frac{\beta_2}{\gamma_2} + \frac{1}{\gamma_2}\int_{\beta_2}^{\gamma_2} r(x)dx. \end{aligned}$$

En multipliant par γ_2, il vient finalement

$$\gamma_2 r(\gamma_2) - \beta_2 r(\beta_2) = \int_{\beta_2}^{\gamma_2} r(x)dx.$$

3.3.3 Application à l'estimation du mode

Les résultats de la section précedente montrent que l'application de k itérations de l'algorithme I.I.R. sur une fonction de régression r fournit une fonction $r^{(k)}$ croissante sur $[0, \beta_k]$, constante sur l'intervalle $[\beta_k, \gamma_k]$ qui contient le mode α puis décroissante sur $[\gamma_k, 1]$. La fonction $r^{(k)}$ présente ainsi un seul intervalle modal qui encadre α et le nombre d'itérations augmentant, l'amplitude de cet intervalle se réduit autour de α (cf. figure 3.29).

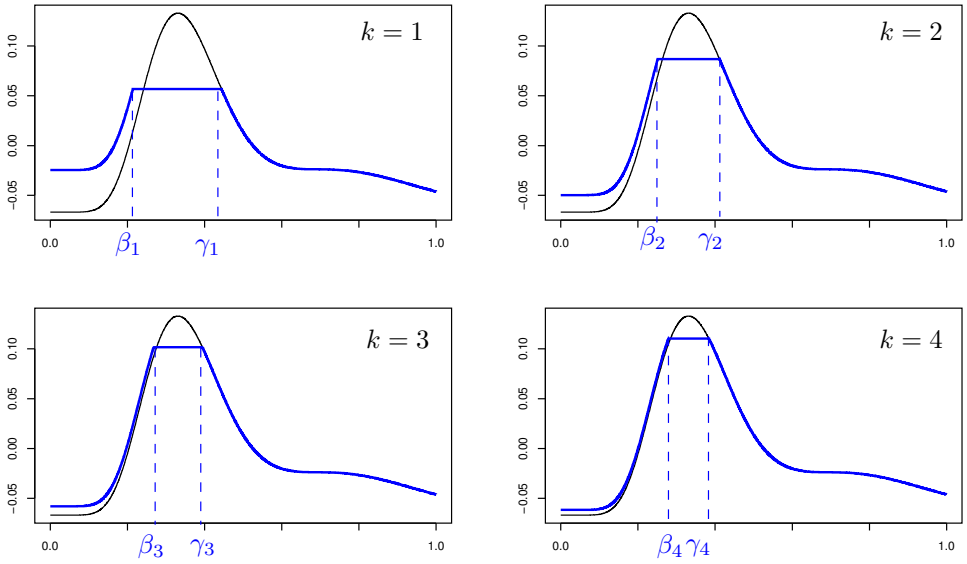


Fig. 3.29 – Déroulement de l'algorithme.

Dans cette section, nous montrons à travers quelques simulations comment notre méthode peut s'appliquer à la régression unimodale. Nous considérons les quatre fonctions unimodales continues r_1, r_2, r_3 et r_4 qui ont été utilisées en Section 3.1. Elles sont représentées en figure 3.30.

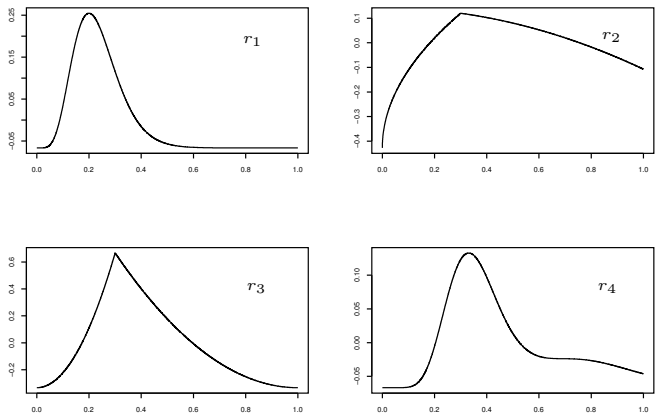


Fig. 3.30 – Les fonctions unimodales considérées.

Nous cherchons à savoir si notre méthode fournit en pratique un intervalle modal qui contienne en effet α. En Section 3.1, nous avons envisagé plusieurs manières d'arrêter l'algorithme : arrêt par application de critères pénalisés (AIC, BIC, GCV, AICc) ; arrêt par contrainte portant sur le nombre de modes (inf qui arrête l'algorithme dès l'apparition du nombre de modes souhaité, sup qui l'arrête avant l'apparition d'un mode supplémentaire). Nous retenons ici les résultats pour le critère AICc et le critère sup dont nous avons vu qu'ils présentaient les meilleures performances pour l'estimation de la fonction de régression. Bien qu'en appliquant le critère sup, nous essayons d'intégrer la contrainte portant sur le nombre de modes, il arrive que l'estimation présente plus d'un extremum local pour tout k. Nous retenons dans ce cas l'intervalle associé au maximum global de l'estimation. Le même principe est appliqué pour le critère AICc puisque, en général, l'estimation associée ne présente pas qu'un seul mode.

Nous considérons deux tailles d'échantillon, $n = 100$, $n = 1000$, les x_i répartis de façon équidistante sur $[0, 1]$ et les mêmes niveaux de bruit qu'en Section 3.1

$$\begin{aligned}\mathrm{var}(\varepsilon) &= 0.05\,\mathrm{var}(r)\\ \mathrm{var}(\varepsilon) &= 0.1\,\mathrm{var}(r)\\ \mathrm{var}(\varepsilon) &= 0.2\,\mathrm{var}(r)\end{aligned}$$

où $\mathrm{var}(r) = \frac{1}{n}\sum_{i=1}^{n}\left(r(x_i) - \bar{r}\right)^2$ avec $\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r(x_i)$. Nous répliquons $N = 1000$ fois les procédures.

Le tableau 3.5 résume les résultats. Il fait figurer les pourcentages de succès dans la localisation de α à l'intérieur de l'intervalle modal de l'estimateur. On observe que les pourcentages de réussite sont meilleurs pour l'arrêt contraint mais ils sont tout de même très satisfaisants lorsque l'on arrête l'algorithme par le critère AICc. En contrepartie, la largeur des intervalles est en moyenne plus grande pour le critère sup, la différence s'accentuant lorsque le nombre de points augmente. Il faut ajouter que la largeur des intervalles ne varie pratiquement pas avec le nombre de points pour cet estimateur, au contraire de ce que l'on observe avec le critère AICc où la taille des intervalles diminue sensiblement. Pour les autres niveaux de bruit, les résultats sont comparables.

Tableau 3.5 – Localisation du mode de r dans l'intervalle modal de l'estimation (entre parenthèses, la largeur moyenne de l'intervalle). Second niveau de bruit

$n = 100$	r_1	r_2	r_3	r_4
AICc	97.1%(0.07)	86.8%(0.15)	95.9%(0.07)	98.2%(0.08)
sup	100%(0.15)	95.9%(0.22)	100%(0.15)	100%(0.14)

$n = 1000$	r_1	r_2	r_3	r_4
AICc	97.3%(0.05)	89.8%(0.08)	98.3%(0.04)	98.7%(0.05)
sup	100%(0.16)	100%(0.28)	100%(0.18)	100%(0.16)

Sur les exemples traités, il semble donc que la régression isotonique itérée puisse donner, avec une bonne confiance, un intervalle contenant le mode de la fonction de régression. On peut imaginer appliquer la procédure en préalable à la régression unimodale habituelle :

- la méthode fournit dans un premier temps un intervalle contenant de façon probable le mode à estimer ;
- on peut ensuite s'inspirer de la démarche de Turner & Wollan (1997) décrite en début de Section 3.3 pour déduire une estimation ponctuelle de α : chaque x_i appartenant à l'intervalle modal est considéré comme un mode potentiel ; on effectue une régression isotonique sur les points situés à sa gauche puis une régression antitonique sur les points de droite, l'estimation retenue étant celle engendrant l'erreur quadratique d'ajustement la plus faible.

Le nombre d'itérations est assez faible (pour le critère AICc, la médiane du nombre d'itérations n'excède pas 5 pour $n = 100$, pas 11 pour $n = 1000$), et l'intervalle étant de taille raisonnable, on peut considérer que le nombre d'appels à la régression isotonique ne risque pas d'être trop grand. Il serait intéressant d'étudier plus en détail la complexité d'une telle procédure et de la comparer aux algorithmes de référence existants comme celui de Stout (2008).

3.3.4 Application à la recherche de plusieurs modes

A notre connaissance, la régression isotonique n'a pas été utilisée pour le cas où il y aurait plusieurs modes à estimer. On peut penser que la raison est le coût de calcul qui dériverait de procédures s'inspirant de la méthode proposée par Turner & Wollan (1997). En effet, si l'on imagine une fonction avec deux modes, il faudrait envisager tous les (x_i, x_j, x_k) avec $i < j < k$ comme autant de triplets (mode1,antimode,mode2) candidats, ce qui nécessiterait de l'ordre de $\mathcal{O}(n^3)$ appels à la régression isotonique.

L'estimation des modes, en particulier en estimation de densité, est pourtant un sujet très étudié. La présence de plusieurs modes peut en effet révéler un mélange de populations dans la distribution (cf. Good & Gaskins (1980)). Par exemple, des tests sur le nombre de modes comme celui de Silverman (1981) ont été développés et étendus au cadre de la régression (cf. Heckman (1992), Harezlak (1998)). Ce dernier motive ses travaux en prenant comme exemple la vitesse de croissance chez les enfants (données disponibles via le package fda de R) où les courbes présentent manifestement plusieurs modes. Les exemples sont très nombreux en régression multivariée : les banques peuvent ainsi chercher à définir les critères sur plusieurs variables qui permettent de maximiser les chances de voir le client rembouser son emprunt ; en médecine, on peut chercher à expliquer différents degrés de sévérité pour une maladie. Certaines de ces questions ont notamment conduit aux travaux regroupés sous le nom de "Bump Hunting" (cf. Friedman & Fisher (1999)).

La figure 3.31 montre l'application de la régression isotonique itérée sur une fonction bimodale (en haut) et une fonction trimodale (en bas), pour $k = 1$ à gauche et $k = 15$ à droite. S'il semble qu'expliciter la fonction résultant de l'application de la régression isotonique sur une fonction à plusieurs modes soit plus difficile que pour un seul mode, la figure suggère qu'en pratique, notre estimateur peut servir à localiser les modes comme dans le cas précédent.

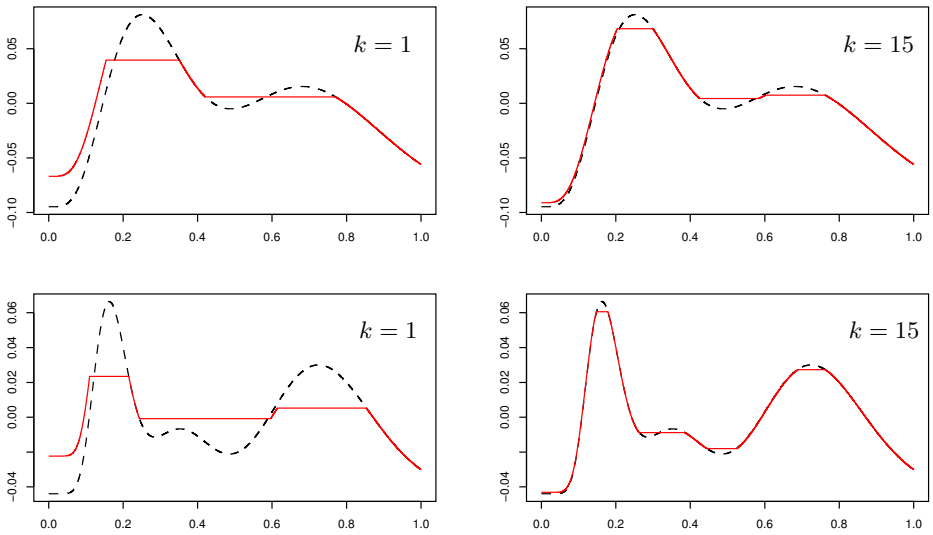


Fig. 3.31 – Application de la régression isotonique itérée sur une fonction à deux modes en haut, à trois modes en bas.

Nous présentons ici quelques simulations réalisées dans le cas de plusieurs modes. Nous considérons les fonctions continues à 2 ou 3 modes parmi l'ensemble des fonctions envisagées en Section 3.1. Elles sont représentées en figure 3.32

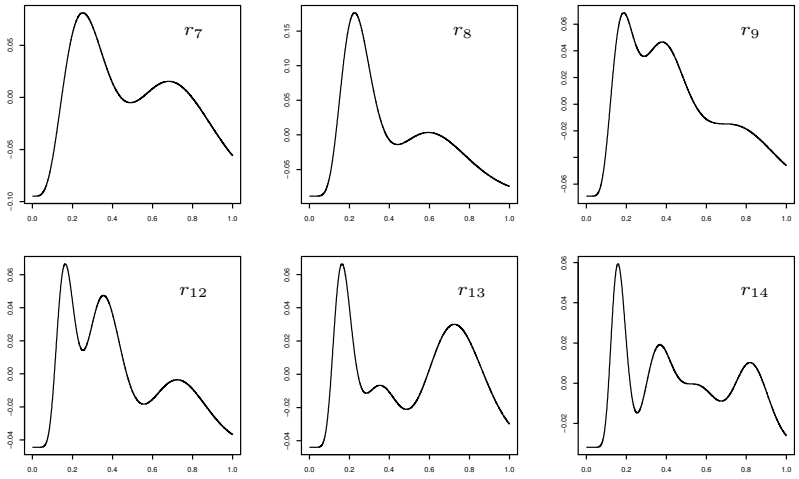


Fig. 3.32 – Les fonctions bimodales (en haut) et trimodales (en bas) considérées.

Nous présentons les résultats pour le second niveau de bruit dans le tableau 3.6. Sont donnés, sur les 1000 réplications, les pourcentages de bonnes localisations individuelles et simultanées des modes. Seules les simulations avec le critère AICc ont été effectuées.

Tableau 3.6 – Localisation des modes de r dans les intervalles de l'estimation. Cas bimodal en haut et trimodal en bas. Second niveau de bruit.

$n = 100, n = 1000$	Mode 1	Mode 2	Modes 1 et 2
r_7	95.9% 96.3%	77.8% 88.0%	75.3% 86.0%
r_8	97.2% 97.3%	44.1% 86.6%	42.1% 84.5%
r_9	96.9% 96.3%	19.4% 90.8%	17.9% 87.7%

$n = 100, n = 1000$	Mode 1	Mode 2	Mode 3	Modes 1, 2, et 3
r_{12}	96.0% 96.1%	92.4% 96.7%	71.2% 74.0%	65.9% 69.9%
r_{13}	97.2% 98.0%	4.30% 51.4%	58.2% 88.4%	2.40% 43.8%
r_{14}	98.2% 98.4%	98.8% 97.3%	78.4% 71.7%	76.7% 69.6%

Mis à part ceux concernant les fonctions r_8 et r_{13} où la détection simultanée des modes n'est pas satisfaisante (voir très mauvaise pour la fonction r_{13} et $n = 100$), les résultats sont assez bons. L'échec de la méthode pour les deux cas précédents vient de la difficulté à capturer les modes de faible amplitude. On remarque qu'avec des échantillon de taille $n = 1000$, la régression isotonique itérée localise les modes au moins sept fois sur dix (sauf pour r_{13}). Notons que les nombres d'itérations sont plus importants pour trois modes que pour deux : la médiane varie de $k = 5$ pour $n = 100$ et deux modes à $k = 30$ pour $n = 1000$ et trois modes.

Il serait intéressant d'approfondir les simulations dans ce cadre d'estimation de modes en envisageant plus de fonctions et en analysant les résultats avec d'autres critères d'arrêt. On peut également penser que la méthode pourrait être utilisée pour construire un test sur le nombre de modes.

Conclusion

La caractérisation de Jordan des fonctions à variation bornée est un résultat d'analyse bien connu. Considérer cette caractérisation comme un modèle additif dans un cadre de régression univariée est l'idée nouvelle qui constitue le point de départ des travaux présentés dans ce document. L'estimateur proposé combine ainsi l'algorithme backfitting destiné à estimer les modèles additifs et la régression isotonique dédiée à l'estimation des fonctions monotones. L'étude théorique est scindée en deux parties : tout d'abord, nous analysons l'estimateur lorsque le nombre de points n est fixé ; nous montrons ensuite sa consistance lorsque n tend vers l'infini.

Lorsque n est fixé, nous montrons que le fait d'augmenter le nombre d'itérations tend à reproduire les données. Une façon directe d'obtenir ce résultat est de voir notre algorithme comme une suite de Von Neumann de projections alternées. Par contre, cela ne suffit pas à caractériser les limites individuelles des termes de la somme constituant l'estimateur. Nous parvenons à caractériser ces limites en montrant que la régression isotonique itérée selon l'algorithme backfitting coïncide avec l'algorithme consistant à utiliser la régression isotonique pour réduire itérativement le biais.

En ce qui concerne la consistance, on ne peut pas utiliser les résultats connus pour la régression isotonique car ceux-ci supposent la fonction de régression monotone. En premier lieu, nous avons donc généralisé la propriété de consistance de la régression isotonique en montrant qu'elle converge vers la projection de la fonction de régression sur le cône des fonctions croissantes. Il s'ensuit que, le nombre d'itérations k étant fixé, notre estimateur converge vers la fonction résultant de l'application de l'algorithme à la fonction de régression elle-même. Cette fonction ne coïncidant pas en général avec la fonction de régression, il subsiste une erreur d'approximation dont on montre qu'elle tend vers 0 avec k. Finalement, nous parvenons à montrer l'existence d'une suite d'itérations (k_n) qui croît avec n, qui tend vers l'infini et telle que l'erreur quadratique moyenne tend bien vers 0 quand n tend vers l'infini.

En pratique, il n'est pas souhaitable de faire tendre le nombre d'itérations vers l'infini puisque cela conduit au sur-ajustement des données. Dans la première partie des applications, nous appliquons l'estimateur dans plusieurs cas simulés et comparons ainsi divers critères d'arrêt pour la méthode. Cette étude simulée a permis d'illustrer le bon comportement de l'estimateur au niveau des discontinuités pour des fonctions présentant des ruptures. Les puces CGH sont utilisées pour la détection du nombre de copies de l'ADN. Détecter les ruptures dans un profil CGH est un enjeu en génétique médicale car leur localisation fournit au médecin des régions du génome qui peuvent être impliquées dans le développement de tumeurs cancéreuses par exemple. Nous avons appliqué notre méthode sur deux jeux de données réelles issue de puces CGH. Les résultats sont encourageants dans la mesure où une utilisation basique de la méthode retrouve la plupart des ruptures attendues. Comme les procédures usuelles pour ces problématiques intègrent un certain nombre de paramètres spécifiques aux données (niveau de bruit, nombre attendu de ruptures, etc.), un axe pratique de développement de la méthode est sans doute de voir comment l'adapter aux problèmes particuliers soulevés par ce genre de données.

Concernant les perspectives, il semble difficile d'espérer obtenir des vitesses de convergence pour notre méthode dans un cadre général. En effet, il n'y a pas, à notre connaissance, de résultat de vitesse en dimension infinie pour l'algorithme de Von Neumann ce qui rend impossible le contrôle du terme d'approximation dans la décomposition de l'erreur quadratique. En revanche, pour une régression unimodale, le fait de réussir à caractériser le résultat de la régression isotonique itérée doit permettre de contrôler le terme d'approximation. Il semble dès lors raisonnable d'espérer calculer une vitesse pour notre estimateur dans ce cas.

La régression unimodale offre également des perspectives pratiques intéressantes car il paraît possible d'utiliser l'estimateur comme étape préalable à la régression unimodale usuelle. Il nous faut encore examiner plus en détail les questions relatives à la complexité de sorte à nous comparer aux méthodes existantes. Une autre piste également envisageable est d'utiliser la régression isotonique itérée pour la détection de plusieurs modes.

Une question naturelle est celle du prolongement multivarié de la méthode. Nous envisageons dans un premier temps de considérer les modèles additifs impliquant une somme de fonctions à variation bornée dans chaque direction de l'espace. L'estimateur serait alors construit en composant deux applications de l'algorithme backfitting et le résultat consisterait en un partitionnement de l'espace des varibles explicatives où l'ajustement serait constant. Il ne paraît pas raisonnable d'espérer concurrencer les méthodes additives classiques pour l'estimation des fonctions régulières. En revanche, il serait intéressant de voir si, à la manière du "Bump Hunting", on réussit à identifier des zones de l'espace où la variable à expliquer prend de grandes valeurs.

Annexe A

Unicité de la décomposition d'une fonction à variation bornée

Dans cette annexe, nous montrons le théorème 0.1 énoncé en page 3 et tiré de Revuz & Yor (2005). Il donne des conditions qui l'assurent l'unicité de la décomposition d'une fonction à variation bornée comme somme d'une fonction croissante et d'une fonction décroissante. Avant cela, nous revenons sur la notion de mesure de Stieltjes qui est utile pour la preuve. Celle-ci généralise la mesure de Lebesgue sur $\mathbb{R}$ en attribuant à chaque intervalle $]a, b]$, non pas sa longueur, mais une masse $F(b) - F(a)$ où F est une application croissante de $\mathbb{R}$ dans $\mathbb{R}$ continue à droite. Nous commençons par donner le théorème suivant, tiré de Briane & Pagès (2006), qui justifie l'existence d'une telle mesure sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ où $\mathcal{B}(\mathbb{R})$ désigne l'ensemble des boréliens de $\mathbb{R}$:

Théorème A.1
Soit $F : \mathbb{R} \to \mathbb{R}$ une fonction croissante continue à droite. Il existe une unique mesure μ_F sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, appelée mesure de Stieltjes associée à F, vérifiant :

$$\forall a, b \in \mathbb{R} \qquad \mu_F(]a, b]) = F(b) - F(a).$$

Ce théorème caractérise les mesures de Stieltjes sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Dès lors, pour tout $a < b$, on a

$$\begin{aligned}
\mu_F(]a, b]) &= F(b) - F(a) \\
\mu_F([a, b[) &= F(b_-) - F(a_-) \\
\mu_F([a, b]) &= F(b) - F(a_-) \\
\mu_F(\{a\}) &= F(a) - F(a_-).
\end{aligned}$$

Le théorème et ces relations sont encore valables sur $([0, 1], \mathcal{B}([0, 1]))$ en convenant par exemple que $F(0_-) = 0$, ce que nous ferons dans la suite.

Rappelons par ailleurs que s'il existe un ensemble $\mathcal{X} \in \mathcal{B}(\mathbb{R})$ tel que, pour une mesure μ, on a $\mu(A \cap \mathcal{X}) = \mu(A)$ pour tout $A \in \mathcal{B}(\mathbb{R})$, alors la mesure μ est dite portée par $\mathcal{X}$. Rappelons également que s'il existe deux ensembles disjoints portant respectivement une mesure μ_1 et une mesure μ_2, alors ces deux mesures sont dites mutuellement singulières.

Nous en venons maintenant au théorème justifiant l'unicité de la décomposition évoquée plus haut.

Théorème A.2
Soit f une fonction définie sur $[0,1]$, à variation bornée et continue à droite en tout point de $[0,1]$. Il existe une unique décomposition $f = F^+ - F^-$ avec F^+ et F^- croissantes sur $[0,1]$ pour lesquelles les mesures de Stieltjes sont mutuellement singulières et définies en 0 par

$$\left\{ \begin{array}{rcl} F^+(0) & = & f(0) \times \mathbb{1}_{f(0)>0} \\ F^-(0) & = & -f(0) \times \mathbb{1}_{f(0)<0} \end{array} \right.$$

Preuve
Soit f une fonction à variation bornée sur $[0,1]$ et une décomposition $f = H^+ - H^-$ où H^+ et H^- sont croissantes. Nous supposons que f est continue à droite en tout point et pouvons considérer que H^+ et H^- le sont également. Introduisons les mesures de Stieltjes ν^+ et ν^- associées à ces deux fonctions. La fonction $H = H^+ + H^-$ étant aussi croissante et continue à droite en tout point, nous pouvons considérer sa mesure de Stieltjes ν. Comme f est à variation bornée, ces trois mesures, en plus d'être positives, sont bornées. Par ailleurs elles vérifient

$$\nu^+ \ll \nu \text{ et } \nu^- \ll \nu. \tag{A.1}$$

On considère désormais l'espace mesuré $([0,1], \mathcal{B}([0,1]), \nu)$. ν^+ (resp. ν^-) est la mesure de Stieltjes associée aux variations de la partie croissante (resp. décroissante) de f mais ces deux mesures ne sont pas a priori mutuellement singulières. Par le théorème de Radon-Nikodym, il existe une fonction de densité $g^+ \in L^1(\nu)$ associée à ν^+ et une fonction de densité $g^- \in L^1(\nu)$ associée à ν^- telles que : $\forall B \in \mathcal{B}([0,1])$,

$$\nu^+(B) = \int_B g^+ d\nu \qquad \nu^-(B) = \int_B g^- d\nu.$$

On pose $g = g^+ - g^-$. Cette fonction est la densité de f par rapport à la mesure ν. En effet, $\forall t \in [0,1]$:

$$\begin{aligned} f(t) &= H^+(t) - H^-(t) \\ &= \int_0^t g^+(s)\nu(ds) - \int_0^t g^-(s)\nu(ds) \\ &= \int_0^t (g^+(s) - g^-(s))\nu(ds) \\ &= \int_0^t g(s)\nu(ds). \end{aligned}$$

Introduisons maintenant

$$f^+(t) = \max(g(t), 0) = g(t)\mathbb{1}_{g(t)>0} \qquad f^-(t) = -\min(g(t), 0) = -g(t)\mathbb{1}_{g(t)<0}.$$

Ce sont deux fonctions positives mesurables telles que $g = f^+ - f^-$ et $\forall B \in \mathcal{B}([0,1])$:

$$\int_B g d\nu = \int_B f^+ d\nu - \int_B f^- d\nu.$$

Nous pouvons associer deux mesures γ^+ et γ^- à chacun des termes de cette somme (cf. Rudin (1975), théorème 1.29, page 23), mesures qui sont mutuellement singulières car à supports

disjoints. On a alors $\forall t \in [0,1]$:

$$\begin{aligned} f(t) &= \int_0^t g(s)\nu(ds) \\ &= \int_0^t f^+(s)\nu(ds) - \int_0^t f^-(s)\nu(ds) \\ &= \int_0^t d\gamma^+(s) - \int_0^t d\gamma^-(s) \\ &= \gamma^+([0,t]) - \gamma^-([0,t]) \\ &:= F^+(t) - F^-(t). \end{aligned}$$

Quitte à imposer des conditions initiales sur F^+ et F^-, on dispose donc d'une écriture de f comme différence de deux fonctions croissantes (puisque définies par des intégrales de fonctions positives) et les mesures de Stieltjes associées à ces deux fonctions sont mutuellement singulières.

L'existence est acquise, montrons maintenant l'unicité. Soit une décomposition $f = F^+ - F^-$ vérifiant les conditions de l'énoncé du théorème avec γ^+ et γ^- les mesures associées. On définit la mesure $\mu = \gamma^+ - \gamma^-$. On peut remarquer que $\mu(t)$ ne dépend pas de la décomposition. En effet, soit $t \in [0,1]$. On a :

$$\begin{aligned} \mu([0,t]) &= \gamma^+([0,t]) - \gamma^-([0,t]) \\ &= F^+(t) - F^-(t) \\ &= f(t). \end{aligned}$$

Montrons maintenant que pour tout borélien B de $[0,1]$, $\gamma^+(B)$ et $\gamma^-(B)$ ne dépendent que de μ. Soit B un borélien. Nous avons, pour tout borélien $B' \subset B$:

$$\begin{aligned} \gamma^+(B) &\geq \gamma^+(B') \\ &\geq \mu(B'). \end{aligned}$$

et donc

$$\gamma^+(B) \geq \sup_{B' \subset B} \mu(B').$$

Notons $\mathcal{X}^+$ un ensemble portant γ^+. On considère une suite (B_n) croissante de boréliens inclus dans B telle que $B = \cup_{n=1}^{\infty} B_n$. On a $\gamma^+(B) = \lim_{n\to\infty} \gamma^+(B_n)$ car γ^+ est une mesure positive. Or $\forall n$,

$$\gamma^+(B_n) = \gamma^+(B_n \cap \mathcal{X}^+) = \mu(B_n \cap \mathcal{X}^+) \leq \sup_{B' \subset B} \mu(B').$$

En passant à la limite, il vient

$$\gamma^+(B) = \lim_{n\to\infty} \gamma^+(B_n) \leq \sup_{B' \subset B} \mu(B').$$

Finalement $\gamma^+(B) = \sup_{B' \subset B} \mu(B')$ dépend de f mais pas de sa décomposition. On obtient de la même façon un résultat analogue pour $\gamma^-(B)$. □

Le point essentiel tient au fait que les mesures γ^+ et γ^- des parties respectivement croissante et décroissante de f sont mutuellement singulières. Par conséquent, il existe deux ensembles boréliens disjoints $\mathcal{X}^+$ et $\mathcal{X}^-$ associés à chacune de ces mesures portant, pour le premier, la partie croissante de f, pour le second, la partie décroissante. De façon pratique, on peut donc

découper l'intervalle $[0,1]$ de sorte que f s'écrive comme la somme d'une fonction croissante et d'une fonction décroissante et de sorte que, lorsque la fonction croissante est strictement croissante, la fonction décroissante est constante et vice-versa. La proposition assure de plus que ce découpage de $[0,1]$ est unique. Nous représentons en figure A.1 une telle décomposition.

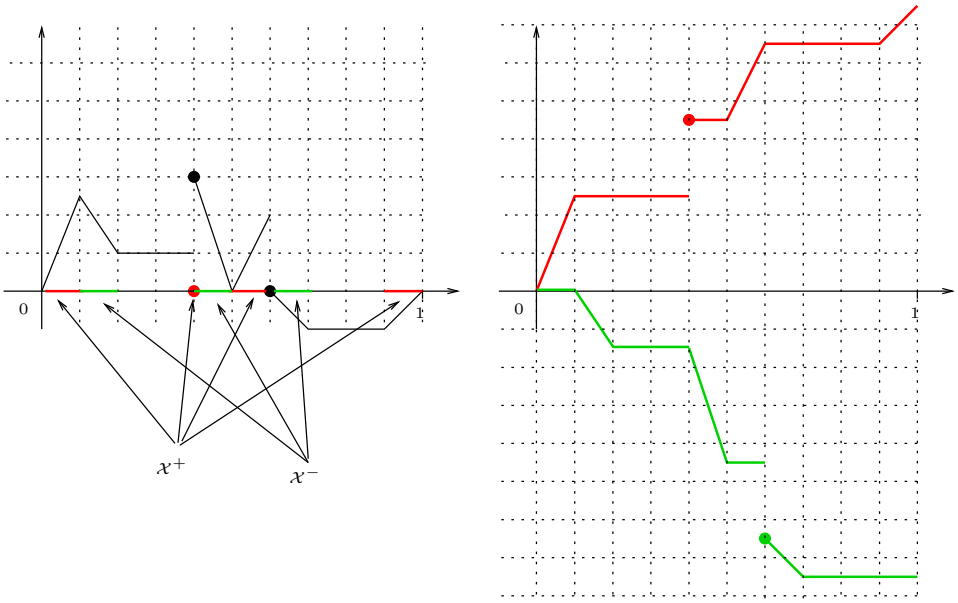


Fig. A.1 – Décomposition d'une fonction à variation bornée.

Avec les conditions initiales sur F^+ et F^-, seul l'un des deux termes de la décomposition "capture" la valeur de f en 0 alors que l'autre est forcément nul en ce point. Sans cette hypothèse, l'unicité dans le théorème 0.1 n'a lieu qu'à une constante additive près, c'est-à-dire à une translation opposée près des deux courbes de la figure A.1. Les mesures de Stieltjes des fonctions de la décomposition restent cependant mutuellement singulières sauf éventuellement au point 0. On est dans ce cas par exemple si l'on considère une fonction f d'intégrale nulle et que l'on décide d'ajuster les points de départ des courbes de F^+ et F^- de sorte à ce que les intégrales de ces deux fonctions soient elles aussi nulles.

Annexe B

Propriétés de $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$

Dans cette annexe, nous montrons que les cônes $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ ont la propriété d'être des cônes minces. De façon générale, considérons un cône convexe fermé $\mathcal{K}$ de $\mathbb{R}^n$. Rappelons auparavant que le cône polaire

$$\mathcal{K}^\star = \{y \in \mathbb{R}^n : \langle x, y\rangle_n \leq 0, \forall x \in \mathcal{K}\} = \{y \in \mathbb{R}^n : P_{\mathcal{K}}(y) = 0\}$$

et qu'un vecteur de $\mathcal{K}$ est sur un rayon extremal (ou arête) si on ne peut pas l'obtenir comme combinaison convexe stricte de deux vecteurs de $\mathcal{K}$. Par ailleurs $\mathcal{K}$ est dit propre si $\mathcal{K} \cap -\mathcal{K} = \{0\}$ et il est dit générateur si $\mathcal{K} - \mathcal{K} = \mathbb{R}^n$. La définition d'un cône mince est la suivante :

Définition B.1 (Cône mince)
On dit qu'un cône $\mathcal{K}$ de $(\mathbb{R}^n, \|.\|_n)$ convexe, fermé, propre et générateur est mince si pour tous vecteurs u et $\tilde{u}$ situés sur deux rayons extrémaux quelconques de son cône polaire $\mathcal{K}^\star$, on a $\langle u, \tilde{u}\rangle_n \leq 0$.

La figure B.1 représente un cône mince dans $\mathbb{R}^2$: c'est un cône dont les arêtes forment des angles aigus (cône aigu). Dans un espace de dimension 2, il est clair que mince équivaut à aigu pour un cône.

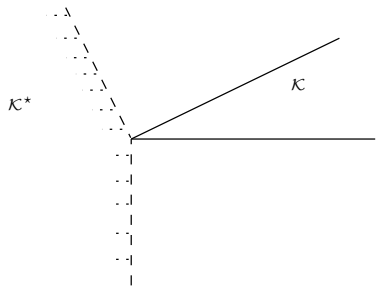


Fig. B.1 – Un cône mince de $\mathbb{R}^2$ et son cône polaire.

En dimension supérieure, si tout cône mince est bien aigu (cf. Isac & Németh (1986)), la réciproque est fausse. En effet, on sait d'après Schrijver (1986) par exemple, que les vecteurs sur les rayons extrémaux du cône polaire sont les vecteurs normaux (orientés correctement) aux faces du cône. La figure B.2 illustre dans $\mathbb{R}^3$ un exemple. Les vecteurs définis dans la base canonique

(i, j, k) par $e_1 = (0, 0, 2)'$, $e_2 = (1, 0, 2)'$ et $e_3 = (-1, 1, 2)'$ sont sur des arêtes de $\mathcal{K}$. Les produits scalaires entre ces vecteurs sont deux à deux positifs : $\mathcal{K}$ est aigu. Posons f_3 le vecteur normal à la face engendrée par e_1 et e_2, f_2 le vecteur normal à la face engendrée par e_1 et e_3 et f_1 le vecteur normal à la face engendrée par e_2 et e_3. Il est facile de calculer $f_3 = (0, -1, 0)'$, $f_2 = (-1, -1, 0)'$ et $f_1 = (1, 2, -1/2)'$ et de constater que f_3 et f_2 forment un angle aigu : $\mathcal{K}$ n'est pas mince.

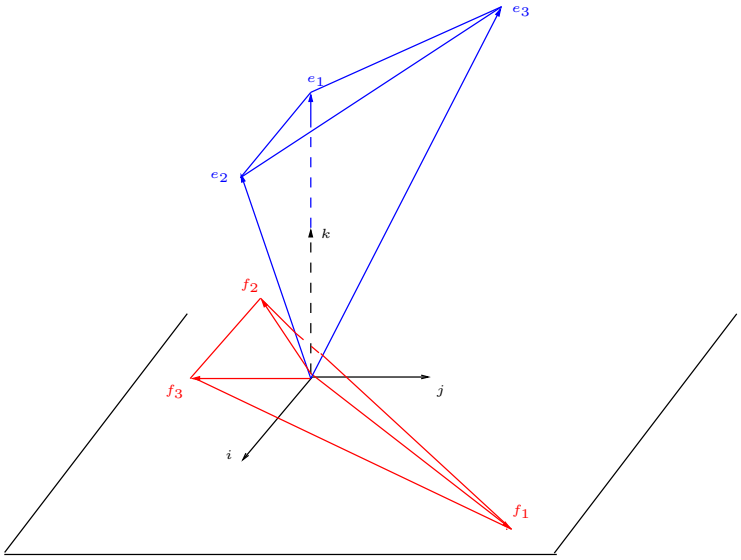


Fig. B.2 – Un cône aigu n'est pas toujours mince.

Nous en venons au résultat principal :

Lemme B.1
Les cônes $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ sont des cônes minces de $\mathcal{E}$, sous-espace de $\mathbb{R}^n$ formé par les vecteurs dont les composantes sont de moyenne nulle.

Preuve
$\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ sont clairement propres, leur intersection étant réduite aux séquences à la fois croissantes et décroissantes de $\mathbb{R}^n$ et de moyenne nulle, c'est-à-dire au vecteur nul. En considérant l'analogue de la décomposition de Jordan pour un vecteur, il est clair que $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$ sont générateurs de $\mathcal{E}$. Reste le dernier point à vérifier. Cela passe par la caractérisation des arêtes des cônes polaires de $\mathcal{C}_n^+$ et $\mathcal{C}_n^-$.

Prenons le cas de $\mathcal{C}_n^+$ et caractérisons tout d'abord ses arêtes. $\mathcal{C}_n^+$ est ici l'ensemble des vecteurs $u \in \mathbb{R}^n$ tels que

$$u_1 \leq u_2 \leq \cdots \leq u_n \text{ et } \sum_{i=1}^{n} u_i = 0.$$

C'est un cône polyédral (c'est-à-dire qu'il possède $n-1$ arêtes supportées par des vecteurs linéairement indépendants) car défini par des intersections d'hyperplans de $\mathbb{R}^n$. La seconde condition implique qu'il est contenu dans $\mathcal{E}$, hyperplan orthogonal au vecteur $e_n = (1, \cdots, 1)'$. Puisque les $(n-1)$ équations $u_i = u_{i+1}$ définissent des hyperplans, $\mathcal{C}_n^+$ est l'intérieur de ces $(n-1)$

hyperplans : il possède $(n-1)$ arêtes $e_1, \cdots, e_{n-1}$ telles que

$$\forall u \in \mathcal{C}_n^+, \exists (\alpha_1, \cdots, \alpha_{n-1}) \in \mathbb{R}^{n-1} : u = \sum_{i=1}^{n-1} \alpha_i e_i.$$

Posons pour $k = \{1, \cdots, n-1\}$:

$$e_k = \quad \underbrace{\left[-\frac{n-k}{n}, \cdots, -\frac{n-k}{n}\right.}_{k} \quad , \quad \underbrace{\left.\frac{k}{n}, \cdots, \frac{k}{n}\right]'}_{n-k}$$

On vérifie facilement que

$$\forall u \in \mathcal{C}_n^+ : u = (u_2 - u_1)e_1 + \cdots + (u_n - u_{n-1})e_{n-1}$$

et les coordonnées sont telles que $u_i - u_{i-1} \geq 0$. Les vecteurs $(e_1, \cdots, e_{n-1})$ forment ainsi les arêtes du cône $\mathcal{C}_n^+$.

Caractérisons maintenant les arêtes de son cône polaire $(\mathcal{C}_n^+)^\star$. Considérons pour $k = \{1, \cdots, n-1\}$, les vecteurs :

$$f_k = \quad \underbrace{[0, \cdots, 0, 1}_{k} \quad , \quad \underbrace{-1, 0, \cdots, 0]'}_{n-k}$$

Il est clair que les f_k forment une base de $\mathcal{E}$. Soit $v \in (\mathcal{C}_n^+)^\star$. On peut le décomposer dans cette base :

$$v = \sum_{k=1}^{n-1} \alpha_k f_k.$$

De plus, $v \in (\mathcal{C}_n^+)^\star \Rightarrow \langle v, e_j \rangle_n \leq 0$ pour tout $j \in \{1, \cdots, n-1\}$. Comme

$$\langle v, e_j \rangle_n = \alpha_j \langle f_j, e_j \rangle_n = -\alpha_j / n,$$

on a finalement

$$v \in \left(\mathcal{C}_n^+\right)^\star \Rightarrow \forall j \in \{1, \cdots, n-1\} : \alpha_j \geq 0.$$

Réciproquement, si $v = \sum_{k=1}^{n-1} \alpha_k f_k$ avec les $\alpha_k \geq 0$, alors

$$\forall j \in \{1, \cdots, n-1\} : \langle v, e_j \rangle_n = -\alpha_j / n \leq 0$$

et l'on a bien $v \in (\mathcal{C}_n^+)^\star$. On conclut ainsi que les vecteurs f_k forment les arêtes de $(\mathcal{C}_n^+)^\star$. Puisque $\langle f_k, f_{k'} \rangle_n = -\delta_{k,k'}/n \leq 0$, le cône $\mathcal{C}_n^+$ est bien mince. On montre selon le même principe que $\mathcal{C}_n^-$ est mince. □

Annexe C

Conditions d'identifiabilité

On a déjà vu qu'en ce qui concerne les conditions d'identifiabilité (2.7) et (2.15) des algorithmes 3 page 37 et 4 page 42, celles portant sur les moyennes étaient vérifées grâce aux propriétés de la régression isotonique. Nous montrons ici que les autres conditions d'identifiabilité résultent également des propriétés de la régression isotonique. Rappelons qu'elles s'écrivent

$$\Delta\left(\hat{u}^{(k)}\right) \circ \Delta\left(\hat{b}^{(k)}\right) = 0 \qquad \text{et} \qquad \Delta\left(\tilde{u}^{(k)}\right) \circ \Delta\left(\tilde{b}^{(k)}\right) = 0$$

avec, pour $z \in \mathbb{R}^n$, $\Delta(z) = (z_2 - z_1, \cdots, z_n - z_{n-1})$, et $\circ$ notant le produit terme à terme de deux vecteurs. Remarquons que $\circ$ a la propriété de distributivité suivante :

$$(\Delta(x) + \Delta(y)) \circ \Delta(z) = \Delta(x) \circ \Delta(z) + \Delta(y) \circ \Delta(z).$$

Concrètement, on cherche à vérifier qu'à chaque étape k, si l'estimation de la partie croissante est variable entre deux abscisses consécutives, l'estimation de la partie décroissante est elle constante et vice-versa comme illustré sur la figure suivante :

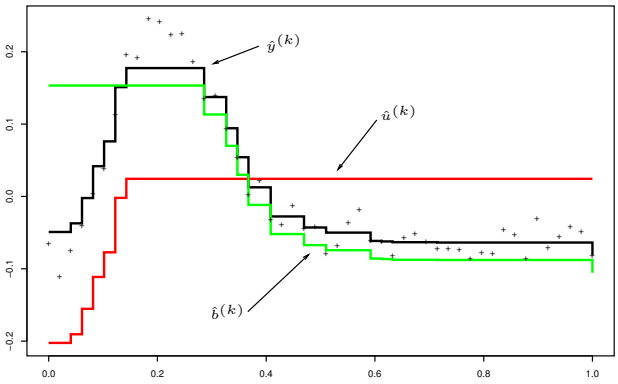


Fig. C.1 – Illustration des conditions d'identifiabilité.

Nous montrons tout d'abord la Proposition C.1 qui donne un résultat intermédiaire. Il concerne l'enchaînement de la régression isotonique sur un vecteur suivie de la régression antitonique sur les résidus (ou l'analogue en inversant les monotonies).

Proposition C.1
Soit $y \in \mathbb{R}^n$ *alors*

$$u = \text{iso}(y) \ \textit{et}\ b = \text{anti}(y-u) \Rightarrow \Delta(u) \circ \Delta(b) = 0 \tag{C.1}$$

et

$$b = \text{anti}(y) \ \textit{et}\ u = \text{iso}(y-b) \Rightarrow \Delta(u) \circ \Delta(b) = 0. \tag{C.2}$$

Preuve
Montrons par exemple

$$u_i < u_{i+1} \Rightarrow b_i = b_{i+1}.$$

Pour cela, nous montrons tout d'abord que lorsque l'on effectue la régression isotonique, la dernière observation d'un bloc est toujours inférieure ou égale à sa valeur ajustée. Pour cela, nous raisonnons par l'absurde en montrant qu'une situation comme celle représentée en figure C.2 aboutit à une contradiction.

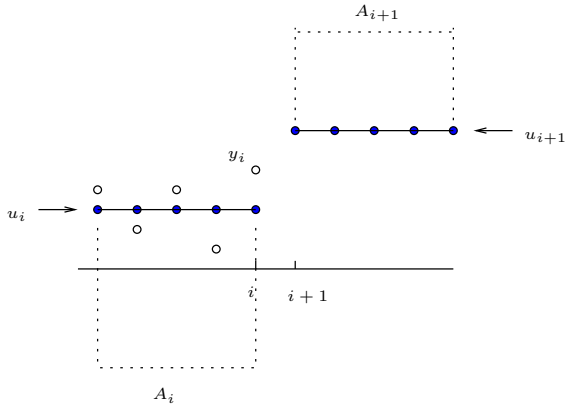


Fig. C.2 – Les deux blocs.

A_i est le bloc dont le dernier élément est y_i. Nous notons pour simplifier u_i la valeur commune attribuée à tous ses éléments par la régression isotonique. u_i est la moyenne des $y_j \in A_i$:

$$u_i = \frac{1}{|A_i|} \sum_{y_j \in A_i} y_j.$$

De même nous considérons le bloc suivant A_{i+1} et notons la valeur ajustée u_{i+1}.

Nous raisonnons donc par l'absurde en supposant $u_i < y_i$. Soit c_{i-1} la valeur moyenne des observations de $A_i - \{y_i\}$. On a clairement $c_{i-1} < u_i$. Par ailleurs, on a nécessairement $y_i < u_{i+1}$ sinon y_i aurait été mis dans A_{i+1} lors de l'application de l'algorithme PAVA. c_{i-1} étant la moyenne des y_j dans $A_i - \{y_i\}$, c'est la constante la proche au sens quadratique de ces y_j. On a donc

$$\sum_{j \in A_i - \{y_i\}} (y_j - c_{i-1})^2 < \sum_{j \in A_i - \{y_i\}} (y_j - u_i)^2.$$

Utiliser les 3 valeurs c_{i-1}, y_i et u_{i+1} pour l'ajustement conduirait ainsi à l'erreur suivante que

l'on peut majorer :

$$\begin{aligned}
&\sum_{j\in A_i-\{y_i\}} (y_j - c_{i-1})^2 + (y_i - y_i)^2 + \sum_{j\in A_{i+1}} (y_j - u_{i+1})^2 \\
&< \sum_{j\in A_i-\{y_i\}} (y_j - u_i)^2 + (y_i - u_i)^2 + \sum_{j\in A_{i+1}} (y_j - u_{i+1})^2 \\
&< \sum_{j\in A_i} (y_j - u_i)^2 + \sum_{j\in A_{i+1}} (y_j - u_{i+1})^2.
\end{aligned}$$

Cet ajustement serait donc meilleur que celui donné par la régression isotonique ce qui est contradictoire avec la définition. On a donc bien $y_i \leq u_i$. On montre de même que la première valeur d'un bloc est nécessairement supérieure à la moyenne du bloc où elle se trouve :

$$y_{i+1} \geq u_{i+1}.$$

Ainsi, il vient pour les résidus de la régression isotonique

$$r_i = y_i - u_i \leq 0 \qquad \text{et} \qquad r_{i+1} = y_{i+1} - u_{i+1} \geq 0$$

donc $r_i \leq r_{i+1}$. Par conséquent, les valeurs r_i et r_{i+1} sont regroupées lors de la régression antitonique, ce qui donne bien $b_i = b_{i+1}$ qui est le résultat voulu. □

Les conditions d'identifiabilité pour le second algorithme sont une conséquence immédiate de cette proposition. On a donc

$$\forall k \geq 1 \qquad \Delta(\tilde{u}^{(k)}) \circ \Delta(\tilde{b}^{(k)}) = 0. \tag{C.3}$$

Montrons maintenant par récurrence

$$\forall k \geq 1 : \qquad \Delta(\hat{u}^{(k)}) \circ \Delta(\hat{b}^{(k)}) = 0.$$

L'équation (C.3) prise avec $k = 1$ assure l'initialisation. Supposons que pour un $k \geq 1$, on a

$$\Delta(\hat{u}^{(k)}) \circ \Delta(\hat{b}^{(k)}) = 0.$$

Alors comme $\hat{u}^{(k+1)} = \hat{u}^{(k)} + \tilde{u}^{(k+1)}$ et $\hat{b}^{(k+1)} = \hat{b}^{(k)} + \tilde{b}^{(k+1)}$, on a par distributivité de $\circ$ et grâce à l'hypothèse de récurrence

$$\Delta(\hat{u}^{(k+1)}) \circ \Delta(\hat{b}^{(k+1)}) = \Delta(\hat{u}^{(k)}) \circ \Delta(\tilde{b}^{(k+1)}) + \Delta(\tilde{u}^{(k+1)}) \circ \Delta(\hat{b}^{(k)}). \tag{C.4}$$

Montrons la nullité des termes à droite cette égalité.
On a

$$\tilde{b}^{(k+1)} = \text{anti}(y - \hat{y}^{(k)} - \tilde{u}^{(k+1)}) = \text{anti}(y - \hat{b}^{(k)} - \hat{u}^{(k+1)})$$

et par définition

$$\hat{u}^{(k+1)} = \text{iso}(y - \hat{b}^{(k)}).$$

Ainsi, d'après (C.1)

$$\Delta(\hat{u}^{(k+1)}) \circ \Delta(\tilde{b}^{(k+1)}) = 0$$

soit

$$\left\{\Delta(\hat{u}^{(k)}) + \Delta(\tilde{u}^{(k+1)})\right\} \circ \Delta(\tilde{b}^{(k+1)}) = 0$$

et finalement

$$\Delta(\hat{u}^{(k)}) \circ \Delta(\tilde{b}^{(k+1)}) = 0$$

qui montre la nullité du premier terme. Pour l'autre terme, on a

$$\tilde{u}^{(k+1)} = \text{iso}(y - \hat{y}^{(k)}) = \text{iso}(y - \hat{u}^{(k)} - \hat{b}^{(k)})$$

et

$$\hat{b}^{(k)} = \text{anti}(y - \hat{u}^{(k)}).$$

On en déduit d'après (C.2)

$$\Delta(\tilde{u}^{(k+1)}) \circ \Delta(\hat{b}^{(k)}) = 0$$

et le terme à droite dans (C.4) est bien nul.

Annexe D

Propriétés des estimateurs à n fixé

Dans cette annexe, nous donnons certaines propriétés de l'estimateur de régression isotonique itérée, lorsque, n étant fixé, le nombre d'itérations k tend vers l'infini. En raison notamment des conditions d'identifiabilité, la norme V(.) définie par

$$\begin{array}{rccl} \mathrm{V}: & \mathcal{E} & \to & \mathbb{R}^+ \\ & y & \mapsto & \mathrm{V}(y) = \sum_{i=1}^{n-1} |y_{i+1} - y_i| \end{array}$$

est d'un usage plus naturel. Rappelons que

$$\mathcal{E} = \left\{ u \in \mathbb{R}^n : \sum_{i=1}^{n} u_i = 0 \right\}$$

ce qui fait bien de V une norme sur $\mathcal{E}$. Nous commençons par établir certaines propriétés pour cette norme. Nous retrouvons ensuite l'analogue de certaines d'entre elles avec la norme $\|.\|_n$ associée à la définition de la régression isotonique.

Utilisation de la norme V(.)

Avant d'établir les propriétés en question, donnons quelques égalités générales concernant la norme V(.), égalités dont les preuves ne présentent pas de difficulté :

- Soit $u = (u_1, \cdots, u_n)' \in \mathcal{C}_n^+$ et $b = (b_1, \cdots, b_n)' \in \mathcal{C}_n^-$. Alors,

$$\mathrm{V}(u) = u_n - u_1 \qquad \text{et} \qquad \mathrm{V}(b) = b_1 - b_n. \tag{D.1}$$

- Soit u et $\tilde{u}$ dans $\mathcal{C}_n^+$, b et $\tilde{b}$ dans $\mathcal{C}_n^-$. Alors,

$$\mathrm{V}(u + \tilde{u}) = \mathrm{V}(u) + \mathrm{V}(\tilde{u}) \qquad \text{et } \mathrm{V}(b + \tilde{b}) = \mathrm{V}(b) + \mathrm{V}(\tilde{b}). \tag{D.2}$$

- Soit u et u' dans $\mathcal{C}_n^+$ (resp. b et b' dans $\mathcal{C}_n^-$) tels que $u'' = u - u' \in \mathcal{C}_n^+$ (resp. $b'' = b - b' \in \mathcal{C}_n^-$). Alors,

$$\mathrm{V}(u'') = \mathrm{V}(u) - \mathrm{V}(u') \qquad \text{et} \qquad \mathrm{V}(b'') = \mathrm{V}(b) - \mathrm{V}(b'). \tag{D.3}$$

La proposition suivante montre que, localement, c'est-à-dire entre la $i^{\text{ème}}$ et la $(i+1)^{\text{ème}}$ coordonnées, $\hat{u}^{(k+1)}$ (resp. $\hat{b}^{(k+1)}$) présente une variation plus grande que $\hat{u}^{(k)}$ (resp. $\hat{b}^{(k)}$), ce que l'on peut observer sur la figure D.1.

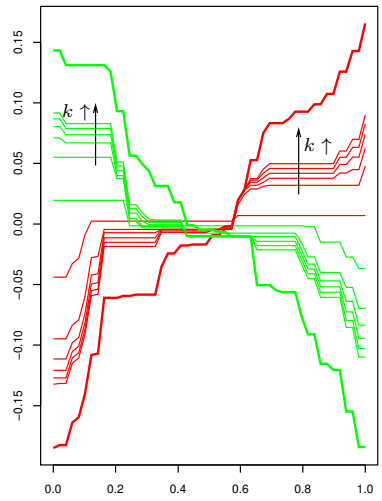


Fig. D.1 – Illustration des propriétés.

Proposition D.1
Soit $\hat{u}^{(k+1)}$ *et* $\hat{u}^{(k)}$ *les parties isotoniques obtenues à deux étapes successives, on a*

$$\forall i \in \{1, \cdots, n-1\} : \hat{u}_{i+1}^{(k+1)} - \hat{u}_i^{(k+1)} \geq \hat{u}_{i+1}^{(k)} - \hat{u}_i^{(k)} \geq 0.$$

De même pour $\hat{b}^{(k+1)}$ *et* $\hat{b}^{(k)}$,

$$\forall i \in \{1, \cdots, n-1\} : \hat{b}_{i+1}^{(k+1)} - \hat{b}_i^{(k+1)} \leq \hat{b}_{i+1}^{(k)} - \hat{b}_i^{(k)} \leq 0.$$

On a comme corollaires immédiats que les suites $\mathrm{V}(\hat{u}^{(k)})$, $\mathrm{V}(\hat{b}^{(k)})$ *et* $\mathrm{V}(\hat{y}^{(k)})$, *indéxées par* k, *sont croissantes.*

Preuve
Utilisons $\hat{u}^{(k+1)} = \hat{u}^{(k)} + \tilde{u}^{(k+1)}$. Pour deux abscisses successives i et $i+1$, on a

$$\hat{u}_{i+1}^{(k+1)} - \hat{u}_i^{(k+1)} = \left(\hat{u}_{i+1}^{(k)} - \hat{u}_i^{(k)}\right) + \left(\tilde{u}_{i+1}^{(k+1)} - \tilde{u}_i^{(k+1)}\right).$$

Or la séquence $\tilde{u}^{(k+1)}$ est croissante donc $\tilde{u}_{i+1}^{(k+1)} - \tilde{u}_i^{(k+1)} \geq 0$ et $\hat{u}_{i+1}^{(k+1)} - \hat{u}_i^{(k+1)} \geq \hat{u}_{i+1}^{(k)} - \hat{u}_i^{(k)}$. Par conséquent, la séquence $\hat{u}^{(k+1)}$ présente entre le $i$ème et la $(i+1)$ème coordonnées, une variation plus grande que $\hat{u}^{(k+1)}$. La variation globale $\mathrm{V}(\hat{u}^{(k+1)})$ de $\hat{u}^{(k+1)}$ est donc bien entendu supérieure à la variation $\mathrm{V}(\hat{u}^{(k)})$ ce qui montre que la suite $(\mathrm{V}(\hat{u}^{(k)}))_k$ est croissante. On étudie de même les parties décroissantes. La suite $(\mathrm{V}(\hat{y}^{(k)}))_k$ est alors croissante comme somme de deux suites croissantes puisque pour tout k, $\mathrm{V}(\hat{y}^{(k)}) = \mathrm{V}(\hat{u}^{(k)}) + \mathrm{V}(\hat{b}^{(k)})$. □

Proposition D.2
Pour tout k, *on a*

$$\begin{aligned}
\mathrm{V}(y - \hat{y}^{(k)}) &= \mathrm{V}(u_y^\star - \hat{u}^{(k)}) + \mathrm{V}(b_y^\star - \hat{b}^{(k)}) \\
&= \mathrm{V}(y) - \mathrm{V}(\hat{y}^{(k)}) \\
&= \sum_{j=k+1}^{\infty} \mathrm{V}(\tilde{u}^{(k)}) + \sum_{j=k+1}^{\infty} \mathrm{V}(\tilde{b}^{(k)}).
\end{aligned}$$

Preuve
On a $y - \hat{y}^{(k)} = (u^* - \hat{u}^{(k)}) + (b^* - \hat{b}^{(k)})$. Montrer le premier résultat revient à montrer que

$$\mathrm{V}\left((u^\star - \hat{u}^{(k)}) + (b^\star - \hat{b}^{(k)})\right) = \mathrm{V}(u^\star - \hat{u}^{(k)}) + \mathrm{V}(b^\star - \hat{b}^{(k)}).$$

Remarquons tout d'abord que $u^\star - \hat{u}^{(k)} = \sum_{j=k+1}^{\infty} \tilde{u}^{(j)}$ est une séquence croissante et que $b^\star - \hat{b}^{(k)} = \sum_{j=k+1}^{\infty} \tilde{b}^{(j)}$ est une séquence décroissante. Ainsi, l'écriture $y - \hat{y}^{(k)} = (u^\star - \hat{u}^{(k)}) + (b^\star - \hat{b}^{(k)})$ est une décomposition de Jordan. Il nous faut vérifier que les deux termes de la somme ont des mesures étrangères. Autrement dit, il faut vérifier que pour tout $i \in \{1, \cdots, (n-1)\}$,

$$\{(u^\star_{i+1} - \hat{u}^{(k)}_{i+1}) - (u^\star_i - \hat{u}^{(k)}_i)\} \times \{(b^\star_{i+1} - \hat{b}^{(k)}_{i+1}) - (b^\star_i - \hat{b}^{(k)}_i)\} = 0$$

ce qui revient à montrer

$$\{(u^\star_{i+1} - u^\star_i) - (\hat{u}^{(k)}_{i+1} - \hat{u}^{(k)}_i)\} \times \{(b^\star_{i+1} - b^\star_i) - (\hat{b}^{(k)}_{i+1} - \hat{b}^{(k)}_i)\} = 0.$$

On sait que $(u^\star_{i+1} - u^\star_i) \times (b^\star_{i+1} - b^\star_i) = 0$ et que $(\hat{u}^{(k)}_{i+1} - \hat{u}^{(k)}_i) \times (\hat{b}^{(k)}_{i+1} - \hat{b}^{(k)}_i) = 0$, donc en développant le membre de gauche, on est ramené à montrer

$$(u^\star_{i+1} - u^\star_i) \times (\hat{b}^{(k)}_{i+1} - \hat{b}^{(k)}_i) + (\hat{u}^{(k)}_{i+1} - \hat{u}^{(k)}_i) \times (b^\star_{i+1} - b^\star_i) = 0.$$

Si $u^\star_{i+1} - u^\star_i = 0$ alors, grâce à la proposition précédente, $\hat{u}^{(k)}_{i+1} - \hat{u}^{(k)}_i = 0$ et c'est fini. Si $u^\star_{i+1} - u^\star_i \neq 0$ alors c'est $b^\star_{i+1} - b^\star_i$ qui est nul et donc $\hat{b}^{(k)}_{i+1} - \hat{b}^{(k)}_i = 0$, toujours en raison de la proposition précédente. La décomposition $y - \hat{y}^{(k)} = (u^\star - \hat{u}^{(k)}) + (b^\star - \hat{b}^{(k)})$ est bien la décomposition de Jordan à variation minimale

$$\mathrm{V}(y - \hat{y}^{(k)}) = \mathrm{V}(u^\star - \hat{u}^{(k)}) + \mathrm{V}(b^\star - \hat{b}^{(k)})$$

et le premier point est vérifié.

Le second point est alors direct :

$$\begin{aligned}
\mathrm{V}(y - \hat{y}^{(k)}) &= \mathrm{V}(u^\star - \hat{u}^{(k)}) + \mathrm{V}(b^\star - \hat{b}^{(k)}) \\
&= \mathrm{V}(u^\star) - \mathrm{V}(\hat{u}^{(k)}) + \mathrm{V}(b^\star) - \mathrm{V}(\hat{b}^{(k)}) \text{ d'après l'équation D.3} \\
&= \{\mathrm{V}(u^\star) + \mathrm{V}(b^\star)\} - \{\mathrm{V}(\hat{u}^{(k)}) + \mathrm{V}(\hat{b}^{(k)})\} \\
&= \mathrm{V}(y) - \mathrm{V}(\hat{y}^{(k)}).
\end{aligned}$$

Enfin,

$$\begin{aligned}
\mathrm{V}(y - \hat{y}^{(k)}) &= \mathrm{V}(u^\star - \hat{u}^{(k)}) + \mathrm{V}(b^\star - \hat{b}^{(k)}) \\
&= \mathrm{V}\left(\sum_{j=k+1}^{\infty} \tilde{u}^{(k)}\right) + \mathrm{V}\left(\sum_{j=k+1}^{\infty} \tilde{b}^{(k)}\right) \\
&= \sum_{j=k+1}^{\infty} \mathrm{V}(\tilde{u}^{(k)}) + \sum_{j=k+1}^{\infty} \mathrm{V}(\tilde{b}^{(k)}) \text{ par continuité de la norme}
\end{aligned}$$

ce qui montre la dernière égalité. □

Proposition D.3
Les suites $\left(\mathrm{V}(y-\hat{y}^{(k)})\right)_k$, $\left(\mathrm{V}(u^\star-\hat{u}^{(k)})\right)_k$ *et* $\left(\mathrm{V}(b^\star-\hat{b}^{(k)})\right)_k$ *tendent vers 0 en décroissant. Les suites* $\left(\mathrm{V}(\tilde{u}^{(k)})\right)_k$ *et* $\left(\mathrm{V}(\tilde{b}^{(k)})\right)_k$ *sont convergentes de limite nulle.*

Preuve
On sait déjà que la suite $\left(\mathrm{V}(y-\hat{y}^{(k)})\right)_k$ est de limite nulle. Elle est de plus décroissante. En effet :

$$\begin{aligned}
\mathrm{V}(y-\hat{y}^{(k+1)}) &= \mathrm{V}(u^\star-\hat{u}^{(k+1)}) + \mathrm{V}(b^\star-\hat{b}^{(k+1)}) \\
&= \mathrm{V}(u^\star-\hat{u}^{(k)}-\tilde{u}^{(k+1)}) + \mathrm{V}(b^\star-\hat{b}^{(k)}-\tilde{b}^{(k+1)}) \\
&= \mathrm{V}(u^\star-\hat{u}^{(k)}) - \mathrm{V}(\tilde{u}^{(k+1)}) + \mathrm{V}(b^\star-\hat{b}^{(k)}) - \mathrm{V}(\tilde{b}^{(k+1)}) \\
&= \mathrm{V}(y-\hat{y}^{(k)}) - \mathrm{V}(\tilde{u}^{(k+1)}+\tilde{b}^{(k+1)}) \\
&\leq \mathrm{V}(y-\hat{y}^{(k)}).
\end{aligned}$$

On sait également que la suite $\left(\mathrm{V}(u^\star-\hat{u}^{(k)})\right)_k$ est de limite nulle. De plus

$$\mathrm{V}(u^\star-\hat{u}^{(k+1)}) = \mathrm{V}(u^\star-\hat{u}^{(k)}-\tilde{u}^{(k+1)}) = \mathrm{V}(u^\star-\hat{u}^{(k)}) - \mathrm{V}(\tilde{u}^{(k+1)}) \leq \mathrm{V}(u^\star-\hat{u}^{(k)})$$

donc elle décroît. C'est la même chose pour $\left(\mathrm{V}(b^\star-\hat{b}^{(k)}\right)_k$.
Les séries de terme général $\mathrm{V}(\tilde{u}^{(k)})$ et $\mathrm{V}(\tilde{b}^{(k)})$ sont convergentes : leur terme général tend vers 0 d'où le point 2. □

Utilisation de la norme $\|.\|_n$

La régression isotonique est un opérateur lipschitzien car elle correspond à la projection sur un cône. Nous montrons tout d'abord que la régression isotonique itérée est également un opérateur lipschitzien.

Proposition D.4 (Lipschitzianité de la régression isotonique itérée)
Soit y *et* z *deux vecteurs de* $\mathbb{R}^n$. *Soit, pour* $k \geq 1$, $\hat{y}^{(k)}$ *et* $\hat{z}^{(k)}$ *les vecteurs ajustés correspondants par régression isotonique itérée. On note les décompositions résultant de l'application de la méthode*

$$\hat{y}^{(k)} = \hat{u}_y^{(k)} + \hat{b}_y^{(k)} \qquad et \qquad \hat{z}^{(k)} = \hat{u}_z^{(k)} + \hat{b}_z^{(k)}.$$

Alors $\forall k \geq 1$, $\forall n \geq 1$,

$$\exists C_{n,k} \leq 2 : \qquad \|\hat{y}^{(k)} - \hat{z}^{(k)}\|_n \leq C_{n,k}\|y-z\|_n.$$

Preuve
Pour la première demi-étape de l'algorithme, écrivons

$$y - z = \{(y-\hat{u}_y^{(1)}) - (z-\hat{u}_z^{(1)})\} + (\hat{u}_y^{(1)} - \hat{u}_z^{(1)})$$

En reprenant la caractérisation de la régression isotonique comme projeté sur un cône donnée en Proposition 1.1 page 9, on obtient, en développant $\|y-z\|_n^2$

$$\|y-z\|_n^2 = \|(y-\hat{u}_y^{(1)}) - (z-\hat{u}_z^{(1)})\|_n^2 + \|\hat{u}_y^{(1)} - \hat{u}_z^{(1)}\|_n^2 + Q_1 \tag{D.4}$$

où

$$Q_1 = -2\{\langle y-\hat{u}_y^{(1)}, \hat{u}_z^{(1)}\rangle_n + \langle z-\hat{u}_z^{(1)}, \hat{u}_y^{(1)}\rangle_n\} \geq 0.$$

Selon le même principe, pour la fin de la première étape, en écrivant

$$(y - \hat{u}_y^{(1)}) - (z - \hat{u}_z^{(1)}) = \{(y - \hat{u}_y^{(1)} - \hat{b}_y^{(1)}) - (z - \hat{u}_z^{(1)} - \hat{b}_z^{(1)})\} + (\hat{b}_y^{(1)} - \hat{b}_z^{(1)}),$$

on obtient

$$\|(y - \hat{u}_y^{(1)}) - (z - \hat{u}_z^{(1)})\|_n^2 = \|(y - \hat{u}_y^{(1)} - \hat{b}_y^{(1)}) - (z - \hat{u}_z^{(1)} - \hat{b}_z^{(1)})\|_n^2 + \|\hat{b}_y^{(1)} - \hat{b}_z^{(1)}\|_n^2 + Q_1' \quad \text{(D.5)}$$

avec

$$Q_1' = -2\{\langle y - \hat{u}_y^{(1)} - \hat{b}_y^{(1)}, \hat{b}_z^{(1)} \rangle_n + \langle z - \hat{u}_z^{(1)} - \hat{b}_z^{(1)}, \hat{b}_y^{(1)} \rangle_n\} \geq 0.$$

Les équations (D.4) et (D.5) conduisent à

$$\|(y - z) - (\hat{y}^{(1)} - \hat{z}^{(1)})\|_n^2 \leq \|y - z\|_n^2 - \|\hat{u}_y^{(1)} - \hat{u}_z^{(1)}\|_n^2 - \|\hat{b}_y^{(1)} - \hat{b}_z^{(1)}\|_n^2. \quad \text{(D.6)}$$

Le principe se généralise facilement et l'on obtient, pour $k \geq 1$

$$\|(y-z)-(\hat{y}^{(k+1)}-\hat{z}^{(k+1)})\|_n^2 \leq \|(y-z)-(\hat{y}^{(k)}-\hat{z}^{(k)})\|_n^2 - \|\hat{u}_y^{(k+1)}-\hat{u}_z^{(k+1)}\|_n^2 - \|\hat{b}_y^{(k+1)}-\hat{b}_z^{(k+1)}\|_n^2.$$

Ainsi, en sommant sur k,

$$\forall k \geq 1 : \qquad \|(y - z) - (\hat{y}^{(k)} - \hat{z}^{(k)})\|_n^2 \leq \|y - z\|_n^2 - \left(\sum_{j=1}^{(k)} \|\hat{u}_y^{(j)} - \hat{u}_z^{(j)}\|_n^2 + \|\hat{b}_y^{(j)} - \hat{b}_z^{(j)}\|_n^2 \right)$$

Comme par ailleurs, $\|(y - z) - (\hat{y}^{(k)} - \hat{z}^{(k)})\|_n \geq \left| \|y - z\|_n - \|\hat{y}^{(k)} - \hat{z}^{(k)}\|_n \right|$, on déduit

$$\forall k \geq 1 : \qquad \|\hat{y}^{(k)} - \hat{z}^{(k)}\|_n \leq \|y - z\|_n + \left\{ \|y - z\|_n^2 - \left(\sum_{j=1}^{(k)} \|\hat{u}_y^{(j)} - \hat{u}_z^{(j)}\|_n^2 + \|\hat{b}_y^{(j)} - \hat{b}_z^{(j)}\|_n^2 \right) \right\}^{\frac{1}{2}}.$$

Ainsi, pour tout n, pour tout k, il existe une constante $C_{n,k}$ telle que$\|\hat{y}^{(k)} - \hat{z}^{(k)}\|_n \leq C_{n,k}\|y - z\|_n$. L'inéquation précédente montre que les constantes $C_{n,k}$ sont uniformément bornées par 2, d'où le résultat. □

Remarque
Il est tentant de penser $y \mapsto \hat{y}^{(k)}$ est 1-lipschitzienne puisque c'est en effet le cas à la limite : $\|\hat{y}^{\infty} - \hat{z}^{\infty}\|_n = \|y - z\|_n$. Cependant le petit exemple $y = (-1, 1, 0)$ et $z := u^\star = (-4/3, 2/3, 2/3)$ montre que ce n'est pas le cas puisque

$$\|\hat{y}^{(1)} - \hat{z}^{(1)}\|_n = \|\hat{y}^{(1)} - u^\star\|_n > \|y - u^\star\|_n.$$

On peut d'ailleurs penser que d'une manière générale, prendre $z := u^\star$ contredit cette intuition comme semble le montrer la figure suivante où la longueur $\|y - z\|_n := \|y - u^\star\|_n$ est représentée en vert et la longueur $\|\hat{y}^{(1)} - \hat{z}^{(1)}\|_n = \|\hat{y}^{(1)} - u^\star\|_n$ est représentée en rouge :

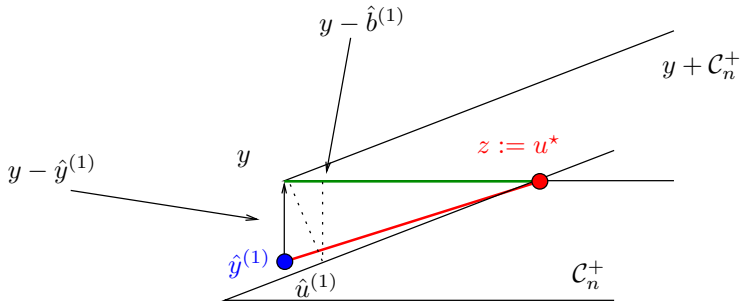


Fig. D.2 – Illustration que $y \mapsto \hat{y}^{(k)}$ n'est pas 1-lipschitzienne.

Les travaux de Bauschke & Borwein (1994) donnent certains résultats complémentaires comme celui de la convergence de la série

$$\sum_k \|y - \hat{y}^{(k)}\|_n^2.$$

En reprenant la représentation des termes $\hat{y}^{(k)}$ et $y - \hat{b}^{(k)}$ associée à l'interprétation de l'algorithme comme une méthode de Von Neumann (cf. figure 2.4 page 41), les résultats de la proposition qui suit sont naturels.

Proposition D.5
Les suites
- $\left(\|y - \hat{y}^{(k)}\|_n\right)_{k\geq 1}$
- $\left(\|u^\star - \hat{u}^{(k)}\|_n\right)_{k\geq 1}$
- $\left(\|b^\star - \hat{b}^{(k)}\|_n\right)_{k\geq 1}$
- $\left(\|\hat{u}^{(k+1)} - \hat{u}^{(k)}\|_n\right)_{k\geq 1}$
- $\left(\|\hat{b}^{(k+1)} - \hat{b}^{(k)}\|_n\right)_{k\geq 1}$

sont décroissantes vers 0.

Preuve
Toutes tendent vers 0 d'après les résultats antérieurs. On peut montrer qu'elles sont décroissantes par récurrence. Pour la première suite, prenons $\hat{y}^{(0)} = 0$ et utilisons la définition :

$$\hat{u}^{(1)} = \underset{u\in\mathcal{C}_n^+}{\operatorname{argmin}} \|y - u\|_n \Rightarrow \|y - \hat{u}^{(1)}\|_n \leq \|y\|_n = \|y - \hat{y}^{(0)}\|_n.$$

De même

$$\hat{b}^{(1)} = \underset{b\in\mathcal{C}_n^-}{\operatorname{argmin}} \|y - \hat{u}^{(1)} - b\|_n \Rightarrow \|y - \hat{y}^{(1)}\|_n = \|y - \hat{u}^{(1)} - \hat{b}^{(1)}\|_n \leq \|y - \hat{u}^{(1)}\|_n$$

donc

$$\|y - \hat{y}^{(1)}\|_n \leq \|y - \hat{y}^{(0)}\|_n.$$

C'est le même principe à un rang k quelconque. Par définition de $\hat{u}^{(k)}$:

$$\|y - \hat{b}^{(k-1)} - \hat{u}^{(k)}\|_n \leq \|y - \hat{b}^{(k-1)} - \hat{u}^{(k-1)}\|_n = \|y - \hat{y}^{(k-1)}\|_n$$

et par définition de $\hat{b}^{(k)}$:

$$\|y - \hat{u}^{(k)} - \hat{b}^{(k)}\|_n \leq \|y - \hat{u}^{(k)} - \hat{b}^{(k-1)}\|_n$$

donc

$$\|y - \hat{y}^{(k)}\|_n \leq \|y - \hat{y}^{(k-1)}\|_n.$$

On peut montrer les deux points suivants en utilisant le fait que la projection sur un cône est 1-lipschitzienne. Par exemple :

$$\begin{aligned}\|\hat{u}^{(k+1)} - \hat{u}^{(k)}\|_n &= \|\operatorname{iso}(y - \hat{b}^{(k)}) - \operatorname{iso}(y - \hat{b}^{(k-1)}\|_n \\ &\leq \|\hat{b}^{(k)} - \hat{b}^{(k-1)}\|_n \\ &\leq \|\operatorname{anti}(y - \hat{u}^{(k)}) - \operatorname{anti}(y - \hat{u}^{(k-1)})\|_n \\ &\leq \|\hat{u}^{(k)} - \hat{u}^{(k-1)}\|_n.\end{aligned}$$

C'est le même principe pour étudier $\left(\|\hat{b}^{(k+1)} - \hat{b}^{(k)}\|_n\right)$.

Pour les deux derniers cas, remarquons tout d'abord que

$$\operatorname{anti}(y - u^\star) = \operatorname{anti}(b^\star) = b^\star \qquad \text{et} \qquad \operatorname{iso}(y - b^\star) = \operatorname{iso}(u^\star) = u^\star.$$

En utilisant en plus la propriété de réduction des distances de la régression isotonique, on montre

$$\|b^\star - \hat{b}^{(1)}\|_n = \|\operatorname{anti}(y - u^\star) - \operatorname{anti}(y - \hat{u}^{(1)})\|_n \leq \|u^\star - \hat{u}^{(1)}\|_n$$

puis

$$\|u^\star - \hat{u}^{(2)}\|_n = \|\operatorname{iso}(y - b^\star) - \operatorname{iso}(y - \hat{b}^{(1)})\|_n \leq \|b^\star - \hat{b}^{(1)}\|_n$$

donc

$$\|u^\star - \hat{u}^{(2)}\|_n \leq \|u^\star - \hat{u}^{(1)}\|_n$$

et ainsi de suite de proche en proche. □

Proposition D.6
Soit pour $k \geq 1$, les estimations $\left(\hat{u}^{(k)}\right)_{k\geq 1}$ et $\left(\hat{b}^{(k)}\right)_{k\geq 1}$ obtenues par régression isotonique itérée. Alors, quelle que soit la décomposition $y = u + b$ avec $u \in \mathcal{C}_n^+$ et $b \in \mathcal{C}_n^-$, les suites

$$\left(\|u - \hat{u}^{(k)}\|_n\right)_{k\geq 1} \qquad \text{et} \qquad \left(\|b - \hat{b}^{(k)}\|_n\right)_{k\geq 1}$$

sont décroissantes.

Preuve
Il suffit de procéder comme en fin de démonstration précédente en remarquant que l'on a, pour toute décomposition (u, b) de y,

$$\operatorname{anti}(y - u) = \operatorname{anti}(b) = b \qquad \text{et} \qquad \operatorname{iso}(y - b) = \operatorname{iso}(u) = u.$$

Dès lors, par exemple,

$$\|b - \hat{b}^{(1)}\|_n \leq \|\operatorname{anti}(y - u) - \operatorname{anti}(y - \hat{u}^{(1)})\|_n \leq \|u - \hat{u}^{(1)}\|_n$$

puis

$$\|u - \hat{u}^{(2)}\|_n \leq \|\operatorname{iso}(y - b) - \operatorname{iso}(y - \hat{b}^{(1)})\|_n \leq \|b - \hat{b}^{(1)}\|_n$$

donc

$$\|u - \hat{u}^{(2)}\|_n \leq \|u - \hat{u}^{(1)}\|_n.$$

Le principe se généralité à tout k et aux séquences décroissantes. □

Annexe E

Fermeture de $\mathcal{C}^+$ et $\mathcal{C}^-$ pour $\|.\|$

On note μ la loi de X sur $[0,1]$ et on considère l'espace $L_2(\mu)$ des fonctions de carré intégrable pour la mesure μ sur l'intervalle $[0,1]$:

$$L_2(\mu) = \left\{ f : [0,1] \to \mathbb{R},\ \int_{[0,1]} f^2(x)\mu(dx) < \infty \right\} = \{ f : [0,1] \to \mathbb{R},\ \mathbb{E}[f^2(X)] < \infty \},$$

que l'on munit de la norme associée :

$$\|f\|^2 = \|f\|_\mu^2 = \mathbb{E}[f^2(X)] = \int_{[0,1]} f^2(x)\mu(dx).$$

En ce sens, deux fonctions sont égales si et seulement si elles coïncident μ-presque sûrement. De la même façon, une fonction de $L_2(\mu)$ est croissante (ou isotonique) si pour μ presque tout x

$$\mu\left(]x,1] \cap [f < f(x)]\right) = 0,$$

où l'on adopte la notation classique en théorie de la mesure :

$$[f < f(x)] = \{x' \in [0,1],\ f(x') < f(x)\}.$$

On note $\mathcal{C}^+$ l'ensemble des fonctions croissantes de $L_2(\mu)$. $\mathcal{C}^+$ est clairement un cône de $L_2(\mu)$. Le résultat suivant montre qu'il est de plus fermé, ce qui permettra d'assurer que la projection orthogonale de toute fonction de $L_2(\mu)$ sur $\mathcal{C}^+$ existe et est unique.

Lemme E.1 (Fermeture de $\mathcal{C}^+$)
Avec les notations précédentes, $\mathcal{C}^+$ est fermé dans $(L_2(\mu), \|.\|_\mu)$.

Preuve
Considérons une suite de fonctions (f_n) croissantes au sens ci-dessus et convergeant dans $L_2(\mu)$ vers une fonction f. Le but est de montrer que f est elle-même croissante. En notant

$$A = \left\{x \in [0,1],\ \mu\left(]x,1] \cap [f < f(x)]\right) > 0\right\},$$

il nous faut donc prouver que $\mu(A) = 0$.

Notons dans un premier temps que l'ensemble A est bien mesurable pour μ. En effet, soit H la fonction définie par

$$H(x) = \mu\left(]x,1] \cap [f < f(x)]\right) = \int_{[0,1]} \mathbb{1}_{]x,1]}(u)\mathbb{1}_{]-\infty,f(x)[}(f(u))\mu(du) = \int_{[0,1]} h(u,x)\mu(du).$$

La μ-mesurabilité de h entraîne celle de H (voir par exemple Billingsley, Probability and Measure, Théorème 18.1) et par là celle de $A = [H > 0]$.

Venons-en maintenant à la μ-négligeabilité de A. Pour ce faire, raisonnons par l'absurde en supposant que $\mu(A) > 0$, et notons pour tout entier naturel j non nul

$$A_j = \left\{ x \in [0,1],\ \mu\left( ]x,1] \cap \left[f < f(x) - \frac{2}{j} \right] \right) > 0 \right\},$$

si bien que (A_j) est une suite d'ensembles μ-mesurables croissante de limite A. En particulier, il existe un indice j_0 tel que $\mu(A_{j_0}) \geq \mu(A)/2$.

La convergence de (f_n) vers f implique qu'il existe un indice n_0 tel que $\|f_{n_0} - f\|_\mu \leq \sqrt{\mu(A)}/(2j_0)$. Nous allons aboutir à une contradiction en montrant que f_{n_0} ne peut être croissante. De l'inégalité de Markov il découle tout d'abord que

$$\mu\left(\left[|f_{n_0} - f| \geq \frac{1}{j_0} \right] \right) = \mu\left(x \in [0,1],\ |f_{n_0}(x) - f(x)| \geq 1/j_0 \right) \leq \frac{\mu(A)}{4}.$$

Pour tout x de $[0,1]$ tel que $|f_{n_0}(x) - f(x)| < 1/j_0$, nous avons l'inclusion suivante

$$\{x' \in]x,1],\ f_{n_0}(x') < f_{n_0}(x) - 1/j_0\} \supseteq \{x' \in]x,1],\ f(x') < f(x) - 2/j_0\},$$

ce qui s'écrit encore

$$]x,1] \cap [f_{n_0} < f_{n_0}(x) - 1/j_0] \supseteq]x,1] \cap [f < f(x) - 2/j_0].$$

Cela implique en particulier

$$\begin{aligned} &\{x \in [0,1],\ \mu(]x,1] \cap [f_{n_0} < f_{n_0}(x) - 1/j_0]) > 0\} \\ &\quad \supseteq \{x \in [0,1],\ \mu(]x,1] \cap [f < f(x) - 2/j_0]) > 0 \ \&\ |f_{n_0}(x) - f(x)| < 1/j_0\}, \end{aligned}$$

de sorte que

$$\begin{aligned} &\mu\left(x \in [0,1],\ \mu(]x,1] \cap [f_{n_0} < f_{n_0}(x) - 1/j_0]) > 0 \right) \\ &\quad \geq \mu\left(\{x \in [0,1],\ \mu(]x,1] \cap [f < f(x) - 2/j_0]) > 0\} \cap \left[|f_{n_0} - f| < \frac{1}{j_0} \right] \right) \\ &\quad \geq \mu\left(x \in [0,1],\ \mu(]x,1] \cap [f < f(x) - 2/j_0]) > 0 \right) - \mu\left(\left[|f_{n_0} - f| \geq \frac{1}{j_0} \right] \right), \end{aligned}$$

ce qui donne finalement

$$\mu\left(x \in [0,1],\ \mu(]x,1] \cap [f_{n_0} < f_{n_0}(x) - 1/j_0]) > 0 \right) \geq \frac{\mu(A)}{2} - \frac{\mu(A)}{4} = \frac{\mu(A)}{4} > 0.$$

Il s'ensuit que

$$\mu\left(x \in [0,1],\ \mu(]x,1] \cap [f_{n_0} < f_{n_0}(x)]) > 0 \right) > 0,$$

ce qui contredit la croissance de f_{n_0} et achève la preuve. □

Annexe F

Terme de biais : $\|u_n - u_+\|$

Rappelons que par terme de biais, nous entendons écart entre le résultat u_n de l'application de la régression isotonique sur les n points $(X_i, r(X_i))$ et la fonction u_+ qui est la meilleure approximation isotonique de la fonction r. La définition de u_n est ainsi associée à la norme $\|.\|_n$ alors que celle de u_+ fait intervenir la norme $\|.\|$. Pour deux fonctions g et h définies sur $I = [0, 1]$, on introduit la variable aléatoire

$$\Delta_n(g - h) = \|g - h\|^2 - \|g - h\|_n^2$$

qui, pour g une fonction donnée et une approximation h de cette fonction par exemple, fournit la différence entre la mesure de l'écart donné par la norme empirique et la mesure donnée par la norme L_2. L'obtention du Lemme F.1 est en grande partie basée sur les propriétés de concentration de cette variable aléatoire. Les propriétés utiles utiles pour la démonstration sont données dans la Section H

Lemme F.1 (Terme de biais)
Soit $(X_1, Y_1), \cdots, (X_n, Y_n)$ *un échantillon i.i.d. avec*

$$Y_i = r(X_i) + \varepsilon_i$$

où $r : [0, 1] \to [0, 1]$ *est une fonction à variation bornée. On note*

$$u_n = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|r - u\|_n^2 \qquad et \qquad u_+ = \underset{u \in \mathcal{C}^+}{\operatorname{argmin}} \|r - u\|^2,$$

alors u_n *converge preque sûrement vers* u_+ *:*

$$\lim_{n \to \infty} \|u_n - u_+\| = 0 \qquad p.s.$$

Preuve
On montre tout d'abord que

$$\|r - u_n\|_n \to \|r - u_+\| \qquad p.s. \tag{F.1}$$

On peut utiliser le Lemme H.1 page 135 en Annexe H pour montrer

$$\limsup \|r - u_n\|_n \leq \|r - u_+\| \qquad p.s. \tag{F.2}$$

Pour cela, on note

$$A_n := A_n(g_+) = \left\{ |\Delta_n(r-u_+)| > n^{-1/3} \right\} = \left\{ |\|r-u_+\|_n^2 - \|r-u_+\|^2| > n^{-1/3} \right\}.$$

Par définition de u_n,

$$\|r-u_n\|_n \leq \|r-u_+\|_n$$

donc sur $\overline{A_n}$

$$\|r-u_n\|_n^2 \leq \|r-u_+\|_n^2 \leq \|r-u_+\|^2 + n^{-1/3}.$$

Ainsi

$$B_n = \left\{ \|r-u_n\|_n^2 \leq \|r-u_+\|^2 + n^{-1/3} \right\} \supset \overline{A_n}.$$

Pour n fixé, l'ensemble des ω n'appartenant pas à B_n peut être assez important. Ce qui nous intéresse, c'est de nous assurer que l'ensemble des ω qui n'appartiennent pas à B_n pour une infinité d'entiers a un poids nul c'est-à-dire $\mathbb{P}(\liminf B_n) = 1$.
Considérons l'ensemble des ω pour lesquels il existe un entier $n_0(\omega)$ tel que l'inégalité $\|r-u_n\|_n^2 \leq \|r-u_+\|^2 + n^{-1/3}$ est vraie pour tout $n \geq n_0$ c'est-à-dire :

$$\liminf B_n = \left\{ \omega \in \Omega : \exists n_0(\omega) \qquad \forall n \geq n_0(\omega) \qquad \|r-u_n\|_n^2 \leq \|r-u_+\|^2 + n^{-1/3} \right\}.$$

Nous avons $\liminf B_n \supset \liminf \overline{A_n}$ donc

$$\mathbb{P}(\liminf B_n) \geq \mathbb{P}\left(\liminf \overline{A_n}\right).$$

Or

$$\mathbb{P}\left(\liminf \overline{A_n}\right) = 1 - \mathbb{P}(\limsup A_n)$$

et on sait d'après le Lemme H.1 que $\sum \mathbb{P}(A_n) < \infty$, donc d'après le lemme de Borel-Cantelli, $\mathbb{P}(\limsup A_n) = 0$. Finalement

$$\mathbb{P}(\liminf B_n) = 1. \tag{F.3}$$

Par ailleurs, pour tout $\omega \in \liminf B_n$, pour $n \geq n_0(\omega)$, on a

$$\|r-u_n\|_n^2 \leq \|r-u_+\|^2 + n^{-1/3},$$

donc

$$\limsup \|r-u_n\|_n^2 \leq \limsup\left(\|r-u_+\|^2\right) + \limsup\left(n^{-1/3}\right) = \|r-u_+\|^2.$$

Ce résultat ajouté à l'équation (F.3) donne (F.2).

Selon un principe comparable, on utilise le Lemme H.2 de l'Annexe H page 136 pour montrer

$$\liminf \|r-u_n\|_n \geq \|r-u_+\| \qquad p.s. \tag{F.4}$$

Par définition de u_+, on a

$$\forall n \qquad \|r-u_+\| \leq \|r-u_n\|.$$

On considère cette fois

$$A_n = \left\{ \sup_{h \in \mathcal{C}^+_{[a,b]}} |\Delta_n(r-h)| > n^{-1/3} \right\}, \qquad \tilde{A}_n = \left\{ |\Delta_n(r-u_n)| > n^{-1/3} \right\}$$

et

$$B_n = \left\{ \|r - u_n\|_n^2 \geq \|r - u_+\| - n^{-1/3} \right\}.$$

On vérifie que $\overline{A}_n \subset \overline{\tilde{A}}_n \subset B_n$. Ainsi

$$\liminf B_n \supset \liminf \overline{A}_n$$

et

$$\mathbb{P}\left(\liminf B_n\right) \geq \mathbb{P}\left(\liminf \overline{A}_n\right) = 1 - \mathbb{P}\left(\limsup A_n\right).$$

D'après le Lemme H.2 page 136, $\sum \mathbb{P}(A_n) < \infty$ donc par le lemme de Borel-Cantelli, $P(\limsup A_n) = 0$. Finalement

$$\mathbb{P}\left(\liminf B_n\right) = 0. \tag{F.5}$$

Pour tout $\omega \in \liminf B_n$, pour tout $n \geq n_0(\omega)$, on a

$$\|r - u_n\|_n^2 \geq \|r - u_+\|^2 - n^{-1/3}$$

donc $\liminf \|r - u_n\|_n^2 \geq \|r - u_+\|^2$ et grâce à (F.5), on obtient (F.4). Finalement, les équations (F.2) et (F.4) donnent le résultat voulu (F.1) :

$$\|r - u_n\|_n^2 \to \|r - u_+\|^2 \qquad p.s.$$

Comme conséquences du Lemme H.2 et du lemme de Borel-Cantelli, on a également

$$\lim_{n\to\infty} \|r - u_n\|_n - \|r - u_n\| = 0 \qquad p.s. \tag{F.6}$$

et résultant des deux dernières égalités

$$\|r - u_n\| \to \|r - u_+\| \qquad p.s. \tag{F.7}$$

Il reste à prouver la convergence presque sûre de u_n vers u_+. On s'appuie pour cela sur l'identité du parallélogramme (cf. figure F.1)

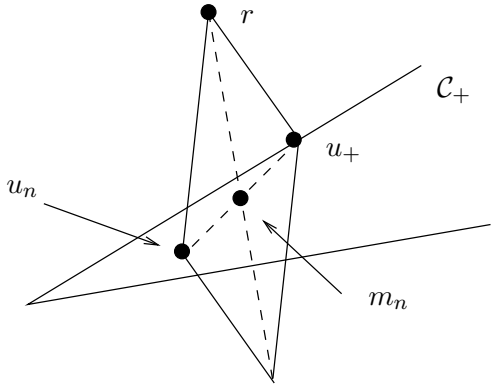


Fig. F.1 – Identité du parallélogramme.

En posant $m_n = (u_n + u_+)/2$, on a

$$2\left(\|r - u_+\|^2 + \|u_n - r\|^2\right) = \|u_n - u_+\|^2 + 4\|m_n - r\|^2,$$

soit

$$\|u_n - u_+\|^2 = 2\left(\|r - u_+\|^2 + \|u_n - r\|^2\right) - 4\|m_n - r\|^2.$$

Comme u_+ et u_n appartiennent à $\mathcal{C}^+$ qui est convexe, $m_n \in \mathcal{C}^+$. Ainsi $\|r - u_+\|^2 \leq \|r - m_n\|^2$ et

$$\|u_n - u_+\|^2 \leq 2\left(\|u_n - r\|^2 - \|r - u_+\|^2\right).$$

Avec l'équation (F.7), on conclut finalement

$$\lim_{n\to\infty} \|u_n - u_+\| = 0 \qquad p.s. \tag{F.8}$$

ce qui était le résultat voulu. □

Annexe G

Terme de variance : $\|\hat{u}_n - u_n\|$

Le terme de variance concerne l'écart entre l'estimateur de régression isotonique appliqué sur les données $(X_i, Y_.)$ et l'estimateur de régression isotonique appliqué aux points $(X_i, r(X_i))$ situés sur la courbe de régression. Contrairement à ce que nous avons fait pour le terme de biais, nous ne montrons pas ici une convergence presque sûre mais seulement une convergence en espérance. Le résultat, qui s'obtient sous l'hypothèse supplémentaire que le bruit est borné, s'énonce de la manière suivante :

Lemme G.1 (Terme de variance)
Soit $(X_1, Y_1), \cdots, (X_n, Y_n)$ *un échantillon i.i.d. avec*

$$Y_i = r(X_i) + \varepsilon_i$$

où $r : [0,1] \to [0,1]$ *est une fonction à variation bornée. On suppose de plus le bruit borné. On note*

$$u_n = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|r - u\|_n^2 \qquad et \qquad \hat{u}_n = \underset{u \in \mathcal{C}_n^+}{\operatorname{argmin}} \|Y - u\|_n^2,$$

alors $\hat{u}_n$ *converge en espérance vers* u_n *:*

$$\lim_{n \to \infty} \mathbb{E}\left[\|\hat{u}_n - u_n\|^2\right] = 0$$

Preuve
La régression isotonique correspondant à une projection sur le cône convexe fermé $\mathcal{C}_n^+$, les éléments $\hat{u}_n$ et u_n sont caractérisés par

$$\langle Y - \hat{u}_n, u - \hat{u}_n \rangle_n \leq 0 \tag{G.1}$$
$$\langle r - u_n, u - u_n \rangle_n \leq 0 \tag{G.2}$$

pour tout $u \in \mathcal{C}_n^+$ avec ici la notation vectorielle

$$\begin{aligned} Y &= \left(Y_{(1)}, \cdots, Y_{(n)}\right)' \\ r &= \left(r(X_{(1)}), \cdots, r(X_{(n)})\right)' \\ \hat{u}_n &= \left(\hat{u}_n(X_{(1)}), \cdots, \hat{u}_n(X_{(n)})\right)' \\ u_n &= \left(u_n(X_{(1)}), \cdots, u_n(X_{(n)})\right)'. \end{aligned}$$

En prenant $u = u_n$ dans l'équation (G.1) et $u = \hat{u}_n$ dans l'équation (G.2), on obtient

$$\langle Y - \hat{u}_n, u_n - \hat{u}_n \rangle_n \leq 0 \qquad \langle r - u_n, \hat{u}_n - u_n \rangle_n \leq 0.$$

On en déduit, en sommant ces deux inégalités

$$\|\hat{u}_n - u_n\|_n^2 \;\leq\; \langle \varepsilon, \hat{u}_n - u_n \rangle_n \tag{G.3}$$

avec $\varepsilon = \left(\varepsilon_{(1)}, \cdots, \varepsilon_{(n)}\right)'$.

On utilise maintenant le Lemme I.1 d'approximation des séquences monotones bornées. Sa démonstration est reportée en annexe I page 141. Ce Lemme montre que l'on peut approcher à une distance inférieure à δ toute séquence croissante majorée par un élément d'un sous-espace vectoriel H^+. Pour des séquences majorées en valeur aboslue par un réel K, H^+ est de dimension $N+1 < n$ avec $N = (2K)^2/\delta^2$.

Le fait de supposer le bruit borné justifie l'existence d'un réel K qui majore en valeur absolue les composantes de Y et donc également celles de $\hat{u}_n$ et de u_n.

Ainsi, on introduit $\hat{h}_n$ et h_n définis par

$$\hat{h}_n = \inf_{h \in H_+} \|\hat{u}_n - h\|_n \qquad h_n = \inf_{h \in H_+} \|u_n - h\|_n$$

et on a

$$\|\hat{u}_n - \hat{h}_n\|_n \leq \delta \qquad \|u_n - h_n\|_n \leq \delta.$$

On déduit alors de l'équation (G.3) :

$$\begin{aligned}
\langle \varepsilon, \hat{u}_n - u_n \rangle_n &= \langle \varepsilon, \hat{u}_n - \hat{h}_n \rangle_n + \langle \varepsilon, \hat{h}_n - h_n \rangle_n \langle \varepsilon, h_n - u_n \rangle_n \\
&\leq \langle \varepsilon, \hat{h}_n - h_n \rangle_n + 2\delta \|\varepsilon\|_n \\
&\leq \|\hat{h}_n - h_n\|_n \left\langle \varepsilon, \frac{\hat{h}_n - h_n}{\|\hat{h}_n - h_n\|_n} \right\rangle_n + 2\delta \|\varepsilon\|_n \\
&\leq \left\{ \|\hat{h}_n - \hat{u}_n\|_n + \|\hat{u}_n - u_n\|_n + \|u_n - h_n\|_n \right\} \sup_{v \in H_+, \|v\|_n = 1} \langle \varepsilon, v \rangle_n \\
&\quad + 2\delta \|\varepsilon\|_n.
\end{aligned}$$

Finalement, on obtient

$$\langle \varepsilon, \hat{u}_n - u_n \rangle_n \leq \{ \|\hat{u}_n - u_n\|_n + 2\delta \} \sup_{v \in H_+, \|v\|_n = 1} \langle \varepsilon, v \rangle_n + 2\delta \|\varepsilon\|_n.$$

Avec l'équation (G.3), cela donne

$$\|\hat{u}_n - u_n\|_n^2 \leq \{ \|\hat{u}_n - u_n\|_n + 2\delta \} \sup_{v \in H_+, \|v\|_n = 1} \langle \varepsilon, v \rangle_n + 2\delta \|\varepsilon\|_n$$

donc

$$\|\hat{u}_n - u_n\|_n^2 \leq \{ \|\hat{u}_n - u_n\|_n + 2\delta \} \|\pi_{H^+}(\varepsilon)\|_n + 2\delta \|\varepsilon\|_n \tag{G.4}$$

où $\pi_{H^+}(\varepsilon)$ est le projeté orthogonal de ε sur H_+.
Cette équation peut être écrite

$$\|\hat{u}_n - u_n\|_n^2 \leq \|\hat{u}_n - u_n\|_n \times \|\pi_{H^+}(\varepsilon)\|_n + 2\delta \left\{ \|\pi_{H^+}(\varepsilon)\|_n + \|\varepsilon\|_n \right\}.$$

On passe à l'espérance et on obtient :

$$\mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2 \right] \leq \mathbb{E}\left[\|\hat{u}_n - u_n\|_n \times \|\pi_{H^+}(\varepsilon)\|_n \right] + 2\delta \left\{ \mathbb{E}\left[\|\pi_{H^+}(\varepsilon)\|_n \right] + \mathbb{E}\left[\|\varepsilon\|_n \right] \right\}.$$

En appliquant l'inégalité de Cauchy-Schwarz et en notant

$$\begin{aligned} x^2 &= \mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2\right] \\ \alpha_n &= \sqrt{\mathbb{E}\left[\|\pi_{H^+}(\varepsilon)\|_n^2\right]} \\ \beta_n &= 2\delta\left\{\mathbb{E}\left[\|\pi_{H^+}(\varepsilon)\|_n\right] + \mathbb{E}\left[\|\varepsilon\|_n\right]\right\} \end{aligned}$$

il vient

$$x^2 - \alpha_n x - \beta_n \leq 0.$$

On fait ainsi apparaître à gauche un second degré et le signe négatif implique $x \leq \frac{\alpha_n + \sqrt{\alpha_n^2 + 4\beta_n}}{2}$ soit

$$\mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2\right] \leq \left(\frac{\alpha_n + \sqrt{\alpha_n^2 + 4\beta_n}}{2}\right)^2.$$

Si on suppose les ε_i i.i.d. centrés de variance commune σ^2 alors, d'après la Proposition J.1 donnée en Annexe J page 145, $\mathbb{E}\left[\|\pi_{H^+}(\varepsilon)\|_n^2\right] = \sigma^2 \frac{N+1}{n}$. On a donc

$$\alpha_n = \sigma\sqrt{\frac{N+1}{n}} = \sigma\frac{\sqrt{4K^2 + \delta^2}}{\sqrt{n}\delta}.$$

Pour $\delta = n^{-\alpha}$ avec $\alpha < 1/2$, on a $\alpha_n \to 0$ quand $n \to \infty$. Par ailleurs, l'inégalité de Jensen implique

$$\beta_n \leq 2\delta(\alpha_n + \sigma).$$

Comme $\alpha_n\delta = \sigma\sqrt{\frac{4K^2+\delta^2}{n}} \to 0$, on en déduit

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2\right] = 0. \tag{G.5}$$

Le fait de supposer le bruit borné permet d'utiliser le Lemme H.3 de l'Annexe H page 138. Ce lemme ajouté au fait que

$$\forall X \geq 0 \qquad \mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}\left(X \geq t\right) dt$$

donne l'existence de $c_1'', c_2'', \beta > 0$ tels que

$$\mathbb{E}\left[\left|\|\hat{u}_n - u_n\|_n^2 - \|\hat{u}_n - u_n\|^2\right|\right] \leq \int_0^{+\infty} c_1'' e^{-c_2'' n^\beta t^2} dt.$$

En posant $f_n(t) = c_1'' e^{-c_2'' n^\beta t^2}$, on a $\forall t \geq 0$, $\lim_{n\to\infty} f_n(t) = 0$ et $\forall t \geq 0$, $\forall n \geq 0$,

$$|f_n(t)| \leq c_1'' e^{-c_2'' t^2} \text{ avec } \int_0^{+\infty} c_1'' e^{-c_2'' t^2} dt < \infty.$$

Par convergence dominée, on a donc, lorsque $n \to \infty$,

$$\mathbb{E}\left[\|\hat{u}_n - u_n\|_n^2\right] - \mathbb{E}\left[\|\hat{u}_n - u_n\|^2\right] \to 0.$$

Du fait de l'équation (G.5), on déduit de ce dernier résultat

$$\lim_{n\to\infty} \mathbb{E}\left[\|\hat{u}_n - u_n\|\right]^2 = 0 \tag{G.6}$$

qui achève la preuve. □

Annexe H

Inégalités de concentration

Dans cette annexe, nous montrons trois lemmes qui permettent de contrôler l'écart entre la norme empirique $\|.\|_n$ et la norme $\|.\|$ de la différence de deux fonctions.

Pour deux fonctions g et h définies sur $I = [0, 1]$, à valeurs dans $[-C, C]$, on considère

$$\Delta_n(g-h) = \frac{1}{n}\sum_{i=1}^{n}\left\{(g(X_i)-h(X_i))^2 - \mathbb{E}\left[(g(X)-h(X))^2\right]\right\}. \tag{H.1}$$

Avec les notations habituelles, $\Delta_n(g-h)$ s'écrit

$$\Delta_n(g-h) = \|g-h\|_n^2 - \|g-h\|^2,$$

c'est-à-dire que pour g une fonction donnée et h une approximation de cette fonction par exemple, $\Delta_n(g-h)$ fournit la différence entre la mesure de l'écart donné par la norme empirique et la mesure donnée par la norme L_2.

L'écriture développée (H.1) montre que c'est une moyenne de variables aléatoires centrées. On a ainsi

– par la loi faible des grands nombres,

$$\forall \epsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}\left(|\Delta_n(g-h)| > \epsilon\right) = 0,$$

– par la loi forte,

$$\mathbb{P}\left(\lim_{n\to\infty} \Delta_n(g-h) = 0\right) = 1$$

– et par l'inégalité de Chebytchev,

$$\forall t > 0, \qquad \mathbb{P}\left(|\Delta_n(g-h)| > t\right) \leq \frac{1}{t^2}\operatorname{var}\left(\Delta_n(g-h)\right).$$

Dans les lemmes qui suivent, nous précisons, moyennant certaines hypothèses, la concentration de cette variable aléatoire autour de 0.

Lemme H.1
Pour toutes fonctions g et h définies sur $[0, 1]$ et à valeurs dans $[-C, C]$, il existe des réels $\alpha > 0$, $\beta > 0$, $c_1 > 0$ et $c_2(C) > 0$ tels que

$$\mathbb{P}\left(|\Delta_n(g-h)| > n^{-\alpha}\right) \leq c_1 \exp\left(-c_2 n^{\beta}\right). \tag{H.2}$$

Preuve
Dans le cas où les $g(X_i) - h(X_i)$ sont bornées en valeur absolue par la constante $2C$, l'inégalité de Hoeffding (cf. Hoeffding (1963)) donne directement

$$\forall n \qquad \forall t > 0 \qquad \mathbb{P}\left(|\Delta_n(g-h)| > t\right) \leq 2\exp\left(-\frac{t^2 n}{8C^2}\right) \tag{H.3}$$

Si l'on prend $t = n^{-\alpha}$ avec $\alpha \in (0, 1/2)$, il vient

$$\forall n \qquad \mathbb{P}\left(|\Delta_n(h)| > n^{-\alpha}\right) \leq 2\exp\left(-\frac{n^{1-2\alpha}}{8C^2}\right) \tag{H.4}$$

et le résultat est démontré avec $c_1 = 2$, $c_2 = 1/(8C^2)$ et $\beta = 1 - 2\alpha > 0$. □

Toujours pour une fonction g fixée, le lemme qui vient est plus fort puisqu'il assure la concentration de $\Delta_n(g-h)$ autour de 0 lorsque h décrit l'ensemble des fonctions croissantes à valeurs dans $[-C, C]$. Il s'appuie sur le fait que l'on peut approcher toute fonction croissante à l'aide d'un nombre fini de fonctions croissantes spécifiques.

Lemme H.2
Soit g une fonction définie sur $[0,1]$ et à valeurs dans $[-C, C]$. On note $\mathcal{C}^+_{[0,1]}$ l'ensemble des fonctions croissantes définies sur $[0,1]$ et à valeurs dans $[-C, C]$. Il existe $\alpha' > 0$, $\beta' > 0$, $c'_1 > 0$ et $c'_2 > 0$ tels que

$$\mathbb{P}\left(\sup_{h \in \mathcal{C}^+_{[0,1]}} |\Delta_n(g-h)| > n^{-\alpha'}\right) \leq c'_1 \exp\left(-c'_2 n^{\beta'}\right). \tag{H.5}$$

Preuve
On commence par montrer que l'application $h \mapsto \Delta_n(g-h)$ est lipschtzienne. On montre facilement que pour deux fonctions h et $\tilde{h}$ quelconques, on a

$$\begin{aligned}\Delta_n(g-h) - \Delta_n(g-\tilde{h}) = \frac{1}{n}\sum_{i=1}^{n} \left\{g(X_i) - h(X_i) + g(X_i) - \tilde{h}(X_i)\right\}\left(h(X_i) - \tilde{h}(X_i)\right) \\ - \mathbb{E}\left[\left\{g(X) - h(X) + g(X) - \tilde{h}(X)\right\}\left(h(X) - \tilde{h}(X)\right)\right].\end{aligned}$$

Pour des fonctions à valeurs dans $[-C, C]$, il vient

$$|\Delta_n(g-h) - \Delta_n(g-\tilde{h})| \leq 4C \times \left\{\frac{1}{n}\sum_{i=1}^{n} |h(X_i) - \tilde{h}(X_i)| + \mathbb{E}\left[|h(X) - \tilde{h}(X)|\right]\right\}$$

et par l'inégalité de Jensen,

$$|\Delta_n(g-h) - \Delta_n(g-\tilde{h})| \leq 4C \times \left\{\|h - \tilde{h}\|_n + \|h - \tilde{h}\|\right\}.$$

Comme $\|h - \tilde{h}\| = \mathbb{E}\left[\|h - \tilde{h}\|_n\right]$, si $\|h - \tilde{h}\|_n \leq \delta$ alors $\|h - \tilde{h}\| \leq \delta$. Ainsi,

$$\forall \delta > 0, \qquad \|h - \tilde{h}\|_n \leq \delta \Rightarrow |\Delta_n(g-h) - \Delta_n(g-\tilde{h})| \leq 8C\delta$$

et $h \mapsto \Delta_n(g-h)$ est bien lipschtzienne pour la norme empirique $\|\cdot\|_n$.

Dès lors, considérons un ensemble de δ-recouvrement de $\mathcal{C}^+_{[0,1]}$ pour la norme $\|.\|_n$: $\{e_j^*\}_{j=1\cdots M}$. Il convient de noter que l'ensemble $\{e_j^*\}_{j=1\cdots M}$ est aléatoire puisqu'il dépend des points X_i, mais que son cardinal M peut être choisi déterministe et majoré comme suit (voir Lemme H.4), en notant [1] $N = \lceil \frac{2C}{\delta} \rceil$

$$M = \binom{n+N}{N} \leq n^N,$$

la dernière inégalité étant vérifiée pour tout n supérieur ou égal à 2 dès que N est supérieur ou égal à 3.

On sait maintenant que pour toute fonction $h \in \mathcal{C}^+_{[0,1]}$, il existe $e^* \in \{e_j^*\}_{j=1\cdots M}$ telle que $\|h - e^*\|_n \leq \delta$. On a alors d'après le résultat précédent

$$|\Delta_n(g-h) - \Delta_n(g-e^*)| \leq 8C\delta.$$

Soit désormais un réel $t > 0$ et considérons $\delta = t/(16C)$. Cherchons à contrôler

$$\mathbb{P}\left(\sup_{h \in \mathcal{C}^+_{[0,1]}} |\Delta_n(g-h)| > t \right).$$

En vertu de l'inégalité triangulaire, on a pour tout h et tout $e^* \in \{e_j\}_{j=1\cdots M}$

$$|\Delta_n(g-h)| \leq |\Delta_n(g-h) - \Delta_n(g-e^*)| + |\Delta_n(g-e^*)|.$$

Pour tout h tel que $|\Delta_n(g-h)| > t$, puisqu'il existe $e^* \in \{e_j^*\}_{j=1\cdots M}$ telle que

$$|\Delta_n(g-h) - \Delta_n(g-e^*)| \leq t/2,$$

on a nécessairement $|\Delta_n(g-e^*)| > t/2$. Ainsi, pour tout h croissante

$$\mathbb{P}\left(|\Delta_n(g-h)| > t\right) \leq \mathbb{P}\left(\max_{j=1\cdots M} |\Delta_n(g-e_j^*)| > t/2 \right).$$

En d'autres termes

$$\mathbb{P}\left(\sup_{h \in \mathcal{C}^+_{[0,1]}} |\Delta_n(g-h)| > t \right) \leq \mathbb{P}\left(\max_{j=1\cdots M} |\Delta_n(g-e_j^*)| > t/2 \right).$$

On applique alors la borne de l'union

$$\begin{aligned}\mathbb{P}\left(\sup_{h \in \mathcal{C}^+_{[0,1]}} |\Delta_n(g-h)| > t \right) &\leq \mathbb{P}\left(\bigcup_{j=1}^{M} |\Delta_n(g-e_j^*)| > t/2 \right) \\ &\leq \sum_{j=1}^{M} \mathbb{P}\left(|\Delta_n(g-e_j^*)| > t/2\right).\end{aligned}$$

D'après l'équation (H.3) et la majoration $M \leq n^N = n^{\lceil \frac{2C}{\delta} \rceil}$ avec ici $\delta = t/(16C)$,

$$\mathbb{P}\left(\sup_{h \in \mathcal{C}^+_{[0,1]}} |\Delta_n(g-h)| > t \right) \leq 2M \exp\left(-\frac{t^2 n}{8C^2}\right) \leq 2\exp\left(\left\lceil \frac{32C^2}{t} \right\rceil \log n - \frac{t^2 n}{8C^2} \right).$$

1. $\lceil x \rceil$ désigne le plus petit entier supérieur ou égal à x.

Pour tout $\alpha' \in (0, 1/3)$, il existe une constante $c_2' = c_2'(\alpha')$ telle que pour tout n

$$\left\lceil \frac{32C^2}{n^{-\alpha'}} \right\rceil \log n - \frac{n^{-2\alpha'} n}{8C^2} \leq -c_2' n^{1-2\alpha'},$$

d'où le résultat annoncé en posant $t = n^{-\alpha'}$ et $\beta' = 1 - 2\alpha'$. □

Le lemme que nous énonçons maintenant est un raffinement du précédent, puisque la borne supérieur est prise sur les deux fonctions g et h. Sa preuve s'appuie cependant sur les mêmes outils.

Lemme H.3
On note $\mathcal{C}^+_{[0,1]}$ l'ensemble des fonctions croissantes définies sur $[0,1]$ et à valeurs dans $[-C, C]$. Il existe $\alpha'' > 0$, $\beta'' > 0$, $c_1'' > 0$ et $c_2'' > 0$ tels que

$$\mathbb{P}\left(\sup_{h_1 \in \mathcal{C}^+_{[0,1]}, h_2 \in \mathcal{C}^+_{[0,1]}} |\Delta_n(h_1 - h_2)| > n^{\alpha''} \right) \leq c_1'' \exp\left(-c_2'' n^{\beta''} \right). \tag{H.6}$$

Preuve
Considérons à nouveau un δ-recouvrement de l'ensemble $\mathcal{C}^+_{[0,1]}$ pour la norme $\|.\|_n$: $\{e_j^*\}_{j=1\cdots M}$. On sait que pour toute fonction $h_1 \in \mathcal{C}^+_{[0,1]}$ (resp. h_2), il existe h_1^* (resp. h_2^*) dans $\{e_j^*\}_{j=1\cdots M}$ telles que

$$\|h_1 - h_1^*\|_n \leq \delta \qquad \text{et} \qquad \|h_2 - h_2^*\|_n \leq \delta.$$

D'après ce qu'on a vu dans la démonstration du lemme précédent, on a dès lors, pour toute fonction g à valeurs dans $[-C, C]$,

$$|\Delta_n(g - h_1) - \Delta_n(g - h_1^*)| \leq 8C\delta \qquad \text{et} \qquad |\Delta_n(g - h_2) - \Delta_n(g - h_2^*)| \leq 8C\delta.$$

En particulier

$$|\Delta_n(h_2 - h_1) - \Delta_n(h_2 - h_1^*)| \leq 8C\delta \qquad \text{et} \qquad |\Delta_n(h_1^* - h_2) - \Delta_n(h_1^* - h_2^*)| \leq 8C\delta.$$

Considérons $\delta = t/(32C)$. D'après l'inégalité triangulaire

$$|\Delta_n(h_1 - h_2)| \leq |\Delta_n(h_2 - h_1) - \Delta_n(h_2 - h_1^*)| + |\Delta_n(h_2 - h_1^*)|,$$

ainsi

$$|\Delta_n(h_1 - h_2)| > t \Rightarrow |\Delta_n(h_2 - h_1^*)| > 3t/4.$$

Toujours d'après l'inégalité triangulaire

$$|\Delta_n(h_2 - h_1^*)| \leq |\Delta_n(h_1^* - h_2) - \Delta_n(h_1^* - h_2^*)| + |\Delta_n(h_1^* - h_2^*)|,$$

donc

$$|\Delta_n(h_2 - h_1^*)| > 3t/4 \Rightarrow |\Delta_n(h_1^* - h_2^*)| > t/2.$$

Ainsi, pour toutes fonctions h_1 et h_2 de $\mathcal{C}^+_{[0,1]}$

$$\mathbb{P}\left(|\Delta_n(h_1 - h_2)| > t \right) \leq \mathbb{P}\left(\max_{h_1^*, h_2^* \in \{e_j^*\}_{j=1\cdots M}} |\Delta_n(h_1^* - h_2^*)| > t/2 \right)$$

donc

$$\mathbb{P}\left(\sup_{h_1\in\mathcal{C}^+_{[0,1]},h_2\in\mathcal{C}^+_{[0,1]}}|\Delta_n(h_1-h_2)|>t\right)\leq\mathbb{P}\left(\max_{h_1^*,h_2^*\in\{e_j^*\}_{j=1\cdots M}}|\Delta_n(h_j^*-h_{j'}^*)|>t/2\right).$$

Comme précédemment, par la borne de l'union et en prenant $\delta=t/(32C)$,

$$\begin{aligned}\mathbb{P}\left(\sup_{h_1\in\mathcal{C}^+_{[0,1]},h_2\in\mathcal{C}^+_{[0,1]}}|\Delta_n(h_1-h_2)|>t\right)&\leq\sum_{1\leq j_1\neq j_2\leq M}\mathbb{P}\left(|\Delta_n(e_j^*-e_{j'}^*)|>t/2\right)\\&\leq M^2\exp\left(-\frac{t^2n}{8C^2}\right)\\&\leq\exp\left(2\left\lceil\frac{64C^2}{t}\right\rceil\log n-\frac{t^2n}{8C^2}\right).\end{aligned}$$

Pour tout $\alpha''\in(0,1/3)$, il existe une constante $c_2''=c_2''(\alpha'')$ telle que pour tout n

$$2\left\lceil\frac{64C^2}{n^{-\alpha''}}\right\rceil\log n-\frac{n^{-2\alpha''}n}{8C^2}\leq-c_2''n^{1-2\alpha''},$$

d'où le résultat annoncé en posant $t=n^{-\alpha''}$ et $\beta''=1-2\alpha''$. □

Nous concluons cette partie par l'estimation des covering numbers utilisée dans les preuves des lemmes précédents.

Lemme H.4
On note $\mathcal{C}^+_{[0,1]}$ l'ensemble des fonctions croissantes définies sur $[0,1]$ et à valeurs dans $[-C,C]$, et $\|.\|_n$ la norme empirique relative à l'échantillon $(X_1,\ldots,X_n)$. Pour tout $\delta>0$, il existe un δ-recouvrement de $(\mathcal{C}^+_{[0,1]},\|.\|_n)$ de cardinal inférieur ou égal à $M=\binom{n+N}{N}$, où $N=\lceil\frac{2C}{\delta}\rceil$ et $\lceil x\rceil$ désigne la partie entière supérieure.

Preuve
Notons $X_{(1)}\leq\cdots\leq X_{(n)}$ l'échantillon $(X_1,\ldots,X_n)$ réordonné. Puisque la distance pour la norme empirique est définie pour tout couple de fonctions g et h de $\mathcal{C}^+_{[0,1]}$ par

$$\|g-h\|_n=\sqrt{\frac{1}{n}\sum_{i=1}^n(g(X_{(i)})-h(X_{(i)}))^2},$$

il suffit que la distance entre g et h en chaque point $X_{(i)}$ soit inférieure à δ pour que ceci reste vrai en norme empirique. Pour simplifier, supposons $N_0=C/\delta$ entier et considérons la subdivision suivante de l'intervalle $[-C,C]$

$$\mathcal{S}=\{-C=-N_0\delta<-(N_0-1)\delta<\cdots<-\delta<0<\delta<\cdots<(N_0-1)\delta<N_0\delta=C\}.$$

Notons alors $\mathcal{E}^+_{[0,1]}$ l'ensemble des fonctions croissantes sur $[0,1]$, à valeurs dans $\mathcal{S}$ et constantes sur les intervalles $(X_{(i)},X_{(i+1)})$, les valeurs sur $[0,X_{(1)}]$ et $[X_{(n)},1]$ étant constantes et respectivement fixées par $X_{(1)}$ et $X_{(n)}$.

Tout d'abord, il est clair que toute fonction g de $\mathcal{C}^+_{[0,1]}$ peut être approchée en norme $\|.\|_n$ par une fonction de $\mathcal{E}^+_{[0,1]}$: il suffit de considérer en chaque point $X_{(i)}$ la valeur de $\mathcal{S}$ la plus proche de $g(X_{(i)})$ (cf. figure H.1). D'autre part, pour se ramener à un problème de combinatoire classique, on peut voir le cardinal de $\mathcal{E}^+_{[0,1]}$ comme le nombre de façons de distribuer n euros parmi $N+1 = 2N_0+1$ personnes. En effet, dans cette correspondance, le nombre d'euros attribués à la personne 1 correspond au nombre de points prenant la valeur $-C = -N_0\delta$, etc., le nombre d'euros attribués à la personne $N+1$ correspond au nombre de points prenant la valeur $C = N_0\delta$. Tous ces nombres sont entiers, compris entre 0 et n, et leur somme vaut n. Il s'ensuit (voir par exemple le polycopié de L. Lovász et K. Vesztergombi, *Discrete Mathematics*, Theorem 4.5)

$$|\mathcal{E}^+_{[0,1]}| = \binom{n+N}{N}$$

et le lemme est démontré. □

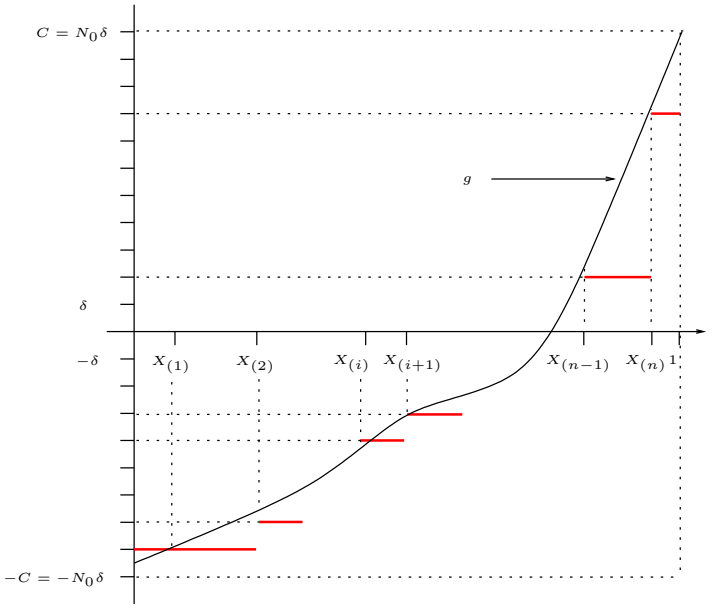


Fig. H.1 – Recouvrement de $\mathcal{C}^+_{[0,1]}$ par $\mathcal{E}^+_{[0,1]}$.

Annexe I

Lemme d'approximation des séquences bornées

On note $\mathcal{C}^+_{n,K}$ (resp. $\mathcal{C}^-_{n,K}$) le sous-ensemble de $\mathcal{C}^+_n$ (resp. $\mathcal{C}^-_n$) formé par les vecteurs dont les composantes sont bornées par un réel K en valeur absolue.
On considère un entier N tel que $N < n$. On pose $\Delta = 1/N$. On introduit les vecteurs de $\mathbb{R}^n$ suivants :

$$h^+_k(i) = \begin{cases} 0 & \text{si } \frac{i}{n} < k\Delta \\ 1 & \text{sinon} \end{cases} \quad , \quad k = 1, \ldots, N,$$

$$h^+_k(i) = \begin{cases} 1 & \text{si } \frac{i}{n} < k\Delta \\ 0 & \text{sinon} \end{cases} \quad , \quad k = 1, \ldots, N,$$

en notant $h^+_k = \left(h^+_k(1), \cdots, h^+_k(n)\right)'$ et $h^-_k = \left(h^-_k(1), \cdots, h^-_k(n)\right)'$. Le vecteur h^+_3 est représenté en figure I.1. On pose par ailleurs $h^+_0 = h^-_0 = (1, \cdots, 1)'$ et l'on définit

$$H_+ = \text{Vect}(h^+_0, \cdots, h^+_N) \quad H^-_+ = \text{Vect}(h^-_0, \cdots, h^-_N).$$

On introduit le réel $\delta > 0$ tel que

$$\Delta = \Delta(\delta) = \delta^2/(2K)^2, \quad N = N(\delta) = \Delta^{-1} = (2K)^2/\delta^2$$

ce qui suppose

$$n > \left(\frac{2K}{\delta}\right)^2 \qquad \text{soit} \qquad \delta > \frac{2K}{\sqrt{n}}.$$

On a le lemme suivant :

Lemme I.1
Pour tout vecteur $f \in \mathcal{C}^+_{n,K}$,

$$\inf_{h \in H^+} \|f - h\|_n \leq \delta$$

et pour tout vecteur $f \in \mathcal{C}^-_{n,K}$,

$$\inf_{h \in H^-} \|f - h\|_n \leq \delta.$$

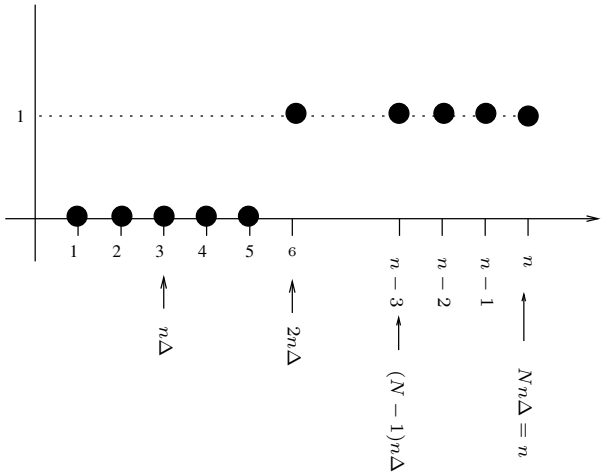


Fig. I.1 – Exemple : le vecteur h_3^+.

Preuve
Prenons le cas d'un vecteur croissant que nous notons : $f = (f(1/n), f(2/n), \cdots, f(1))'$. On introduit, pour $j = 1, \cdots, N$, les quantités $f_j = f(j\Delta)$ et on pose $f_0 = -K$, $f_{N+1} = K$. On introduit les vecteurs de H_+ :

$$h_- = f_0 h_0^+ + \sum_{j=1}^{N-1} (f_j - f_{j-1}) h_j^+ + f_N h_N^+$$

et

$$h_+ = f_1 h_0^+ + \sum_{j=1}^{N-1} (f_{j+1} - f_j) h_j^+ + f_{N+1} h_N^+.$$

Un exemple où $n = 27$ et $N = 9$ est représenté en figure I.2.
En transformant l'écriture de h_+ sous la forme

$$h_+ = f_1 h_0^+ + \sum_{j=2}^{N} (f_j - f_{j-1}) h_{j-1}^+ + f_{N+1} h_N^+,$$

on obtient

$$h_+ - h_- = \sum_{j=1}^{N-1} (f_j - f_{j-1})(h_{j-1}^+ - h_j^+) + (f_N - f_{N-1}) h_{N-1}^+ + (f_{N+1} - f_N) h_N^+$$

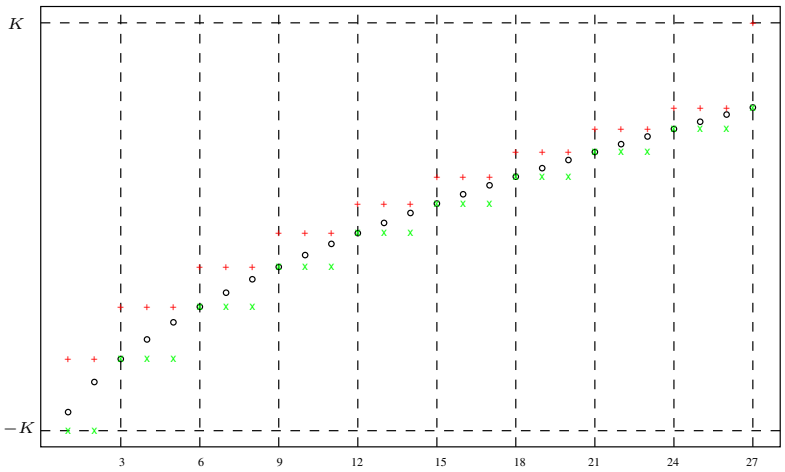


Fig. I.2 – h_+ en rouge et h_- en vert.

Ainsi,

$$\begin{aligned}\|h_+ - h_-\|_n^2 &\leq \frac{1}{n}\left(\sum_{j=1}^{N-1}(f_j - f_{j-1})^2 \times n\Delta + (f_N - f_{N-1}) \times n\Delta + (f_{N+1} - f_N)^2\right)\\ &\leq \Delta \sum_{j=1}^{N+1}(f_j - f_{j-1})^2\\ &\leq \Delta \times (2K)^2 \sum_{j=1}^{N+1}\left(\frac{f_j - f_{j-1}}{2K}\right)^2.\end{aligned}$$

Or $0 \leq \frac{f_j - f_{j-1}}{2K} \leq 1$ donc

$$\left(\frac{f_j - f_{j-1}}{2K}\right)^2 \leq \frac{f_j - f_{j-1}}{2K},$$

ainsi,

$$\|h_+ - h_+\|_n^2 \leq \Delta(2K)^2 \times \frac{1}{2K}\sum_{j=1}^{N} f_j - f_{j-1} = 4\Delta K^2.$$

Comme $\Delta = \frac{\delta^2}{(2K)^2}$, on a bien $\inf_{h\in H}\|f - h\|_n^2 \leq \delta^2$. Une démonstration similaire, en raisonnant avec les fonctions h_j^- aboutit au résultat souhaité pour f décroissante. La démonstration qui vient d'être faite s'applique dans le cas du design fixe $(0, 1/n, \cdots, (n-1)/n, n)$ mais le raisonnenment se généralise facilement. □

Annexe J

Projection d'un vecteur aléatoire sur un sous-espace

Proposition J.1
Soit $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)' \in \mathbb{R}^n$ *un vecteur aléatoire tel que les variables* ε_i *sont i.i.d. d'espérance* $\mathbb{E}[\varepsilon_i] = \mu$ *et de variance* $\mathrm{var}(\varepsilon_i) = \sigma^2 < \infty$. *Soit* π_H *la matrice dans la base canonique de la projection orthogonale sur un sous-espace vectoriel* H. *On suppose que* H *est de dimension* m *et qu'il contient la constante* $\mathbb{1}_n$, *alors*

$$\mathbb{E}\left[\|\pi_H \varepsilon\|_n^2\right] = \mu^2 + \frac{m}{n}\sigma^2.$$

Preuve
Comme $\|\pi_H \varepsilon\|_n^2 = \frac{1}{n}(\pi_H \varepsilon)'(\pi_H \varepsilon)$ est un réel, il est égal à sa trace. On obtient ainsi :

$$\begin{aligned}
\mathbb{E}\left[\|\pi_H \varepsilon\|_n^2\right] &= \frac{1}{n}\mathbb{E}\left[(\pi_H \varepsilon)'(\pi_H \varepsilon)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left((\pi_H \varepsilon)'(\pi_H \varepsilon)\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left(\varepsilon' \pi_H' \pi_H \varepsilon\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left(\varepsilon' \pi_H \varepsilon\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left(\varepsilon \varepsilon' \pi_H\right)\right] \\
&= \frac{1}{n}\,\mathrm{tr}\left(\mathbb{E}\left[\varepsilon \varepsilon'\right] \pi_h\right)
\end{aligned}$$

Avec les hypothèse faites sur les ε_i, on a $\mathbb{E}[\varepsilon\varepsilon'] = \sigma^2 \mathrm{Id} + \mu^2 U$, où U est la matrice $n \times n$ dont tous les éléments valent 1. Ainsi,

$$\mathbb{E}\left[\|\pi_H \varepsilon\|_n^2\right] = \frac{\sigma^2}{n}\,\mathrm{tr}(\pi_H) + \frac{\mu^2}{n}\,\mathrm{tr}(U\pi_H).$$

Comme π_H est la matrice de projection de H qui est de dimension m, $\mathrm{tr}(\pi_H) = m$. De plus, au facteur $1/n$ près, la matrice U est celle de la projection orthogonale sur le vecteur $\mathbb{1}_n$:

$$\frac{U}{n} = \mathbb{1}_n(\mathbb{1}_n'\mathbb{1}_n)^{-1}\mathbb{1}_n'.$$

On a $\mathrm{tr}(U\pi_H) = \mathrm{tr}(\pi_H U)$. Comme $\pi_H U$ est la matrice de la composition de la projection orthogonale sur $\mathrm{Vect}(\mathbb{1}_n)$ suivie de celle sur H qui contient $\mathbb{1}_n$, on a $\mathrm{tr}(\pi_H U) = \mathrm{tr}(U) = n$, d'où le résultat. □

Stat Comput
DOI 10.1007/s11222-012-9346-4

Iterative bias reduction: a comparative study

P.-A. Cornillon · N. Hengartner · N. Jegou · E. Matzner-Løber

Received: 28 April 2011 / Accepted: 27 July 2012

Abstract Multivariate nonparametric smoothers, such as kernel based smoothers and thin plate splines smoothers, are adversely impacted by the sparseness of data in high dimension, also known as the curse of dimensionality. Adaptive smoothers, that can exploit the underlying smoothness of the regression function, may partially mitigate this effect. This paper presents a comparative simulation study of a novel adaptive smoother (IBR) with competing multivariate smoothers available as package or function within the R language and environment for statistical computing. Comparison between the methods are made on simulated datasets of moderate size, from 50 to 200 observations, with two, five or 10 potential explanatory variables, and on a real dataset. The results show that the good asymptotic properties of IBR are complemented by a very good behavior on moderate sized datasets, results which are similar to those obtained with Duchon low rank splines.

Keywords Multivariate smoothing · Thin-plate splines · Duchon splines · Kernel regression · Iterative bias reduction

P.-A. Cornillon
IRMAR, Univ. Rennes 2, 35043 Rennes, France
e-mail: pac@uhb.fr

N. Jegou · E. Matzner-Løber (✉)
Univ. Rennes 2, 35043 Rennes, France
e-mail: eml@uhb.fr

N. Jegou
e-mail: nicolas.jegou@uhb.fr

N. Hengartner
Los Alamos National Laboratory, Los Alamos, NW 87545, USA
e-mail: nickh@lanl.gov

Published online: 30 August 2012

1 Introduction

In many applications, one seeks to explain a response variable by a set of potential explanatory variables. Regression, which is a fundamental data analysis tool, solves this problem by estimating the functional relationships between pairs of observations $(X_i, Y_i), i = 1, \ldots, n$. In its simplest form, one models the conditional expectation of the dependent variable Y given the independent variables $X \in \mathbb{R}^d$ by a linear combination of the covariates and estimates the parameters by minimizing a suitable cost function between the observed and the fitted values, usually the sum of squared errors. More generally, one may explicitly specify parametric families for regression functions that describe the conditional expectation of the dependent variable Y given X.

Nonparametric regression provides a more flexible model that does not require the specification of a particular parametric form from the conditional expectation. Instead, it only assumes that the conditional expectation of Y be a smooth function of the covariates X. Typically, nonparametric models are estimated locally, and the predicted values are smoother than the original observations. Hence nonparametric regression estimators are often called smoothers.

Over the past thirty years, numerous smoothers have been proposed: running-mean smoother, running-line smoother, bin smoother, kernel based smoother, splines regression smoother, smoothing splines smoother, locally weighted running-line smoother, just to mention a few. We refer to Buja et al. (1989), Eubank (1988), Fan and Gijbels (1996), and Hastie and Tibshirani (1995) for more in depth treatments of regression smoothers. Most of these smoothers behavior is closely related to a good choice for the smoothing parameter λ and much has been written on how to select an appropriate smoothing parameter (see for example Simonoff 1996). Classical smoothers have to face *the curse*

of dimensionality which could be summarized as follows: as the dimension of the data increases, so does the sparseness of the covariates and as a consequence, nonparametric smoothers must average over larger neighborhoods, which in turn produces more heavily biased smoothers. Optimally selecting the smoothing parameter does not alleviate this problem and as a remedy, the common wisdom is to avoid all together general nonparametric smoothing with moderate sample size in dimensions higher than three. In this cases, it is usual practice in the statistical community to fit structurally constrained regression models such as additive models (Hastie and Tibshirani 1995; Wood 2004), MARS (Friedman 1991), projection pursuit models (Friedman and Stuetzle 1981) or additive L_2-Boosting (Bühlmann and Yu 2003).

The popularity of additive models (or MARS models) stems from its interpretability and from the fact that the estimated regression function converges to the best additive approximation of the true regression function at the optimal univariate mean squared error rate of $n^{-2\nu/(2\nu+1)}$, where ν is the smoothing index (see for example Tsybakov 2009). While additive models do not estimate the true underlying regression function, one hopes that the approximation error will be small enough so that for moderate sample sizes, the prediction mean square error of the additive model is less than the prediction error of a fully nonparametric regression model.

The optimal mean square error rate of convergence depends on both the dimension d of the covariates and the smoothness of the unknown regression function, which is of course unknown. It is well-known that for regression function m from $\mathbb{R}^d$ to $\mathbb{R}$ known to belong to some smoothness functional classes (e.g Holder, Sobolev, Besov), the optimal mean squared error rate of convergence is $n^{-2\nu/(2\nu+d)}$. Thus, if the regression function m is of smoothness index $\nu = 2d$, then the optimal rate is $n^{-4/5}$, a value recognized as the optimal mean squared error of estimates for twice differentiable univariate regression functions. This suggests that nonparametric regression in higher dimensions is practical, provided that the true regression function is known to be sufficiently smooth, and that the smoothing methods exploits this knowledge.

While in practice, one rarely knows *a priori* the smoothness of the regression function, there exists smoothers achieving the optimal asymptotic mean square error without prior specification of the smoothness. Such methods are called adaptive, and we refer the interested reader to Lepski (1991), Gyorfi et al. (2002), Tsybakov (2009) for general discussions on adaptation in nonparametric estimation. Roughly speaking, adaptation can be achieved either by direct estimation (see for example Lepski's method, Lepski 1991, and related papers) or by aggregation of different procedures (see Yang 2000). Even if potential gain can be achieved by these nonparametric adaptive estimators, there is a lack of multivariate adaptive smoothers that work well in practice with moderate sample size n (ranging from a hundred to few thousands observations). Recently, Cornillon et al. (2011b) proposed an adaptive iterative smoothing method that is very promising for such datasets. The method, called *Iterative Bias Reduction* (abbreviated to IBR in this paper), starts out with a biased smoother that has a large smoothing parameter λ (ensuring that the data are over-smoothed) and then proceeds to estimate and correct the bias in an iterative fashion. This approach is attractive in that it uses existing smoothers, yet by iteratively estimating and correcting the bias, it achieves adaptation.

The aim of this paper is (1) to demonstrate, through simulations and applications to a real dataset, the good practical performance of IBR predicted by the asymptotic theory in Cornillon et al. (2011b) for moderate sample sizes and (2) to compare its performances to those obtained by various competitors. All these competitors must be usable for end-user and thus must be included in some R packages. This last consideration leads us to compare IBR to the following methods: additive models (R package **mgcv**), projection pursuit regression (R function **ppr**), MARS (R package **mda**), additive L_2-Boosting (R package **mboost**) and direct multivariate regression modeling such as low rank thin-plate splines or Duchon splines (R package **mgcv**, Wood 2003).

The paper is organized as follows. Section 2 briefly introduces the IBR smoother, discusses how to initiate and stop the iterative procedure, and reviews its theoretical properties. Section 3 presents IBR with thin plate splines and Duchon splines and discusses the choice of the initial values in order to obtain biased (pilot) smoothers. Section 4 assesses the finite sample properties of the IBR smoother by comparing in simulations its performances with other multivariate smoothing methods that have end-user implementation. Section 5 discusses variable selection for nonparametric smoothers, and show through simulations, improvements in the prediction mean squared error. Section 6 applies variable selection for the IBR smoother to the Los Angeles Ozone dataset and concluding remarks end the paper.

2 IBR: iterative bias reduction

2.1 Preliminaries: linear smoother

Suppose that the pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ are related through the nonparametric regression model

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

where $m(\cdot)$ is an unknown smooth function and the disturbances ε_i are independent mean zero and variance σ^2 random variables that are independent of all the covariates

$(X_1, \ldots, X_n)$. It is helpful to rewrite equation (1) in vector form by setting $Y = (Y_1, \ldots, Y_n)^t$, $m = (m(X_1), \ldots, m(X_n))^t$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$, to get

$$Y = m + \varepsilon. \tag{2}$$

Linear smoothers can be written in vector format as

$$\widehat{m}_1 = S_1 Y, \tag{3}$$

where S_1 is an $n \times n$ smoothing matrix depending on a smoothing parameter and $\widehat{m}_1 = \widehat{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^t$ denotes the vector of fitted values. The conditional bias of such a linear smoother is

$$B(\widehat{m}_1) = \mathbb{E}[\widehat{m}_1 | X] - m = (S_1 - I)m. \tag{4}$$

2.2 Bias reduction of linear smoothers

The expression (4) for the bias suggests that it can be estimated by smoothing the negative residuals $-R_1 = -(Y - \widehat{m}_1) = -(I - S_1)Y$. That is,

$$\hat{b}_1 := -S_2 R_1 = -S_2(I - S_1)Y \tag{5}$$

estimates the bias using a (possibly) different smoother S_2. Correcting the pilot smoother $\widehat{m}_1$ by subtracting $\hat{b}_1$ yields a *bias corrected* smoother

$$\widehat{m}_2 = S_1 Y + S_2(I - S_1)Y = \big(S_1 + S_2(I - S_1)\big)Y.$$

Since $\hat{m}_2$ is itself a linear smoother, it is possible to correct its bias as well. Repeating the bias reduction step $k-1$ times produces the linear smoother given in the following proposition. We have to keep in mind that in order to reduce the bias, we need a biased initial smoother. Moreover, at each iteration, reducing the bias is done at the cost of increasing the variance. A natural question is how to stop algorithm (c.f. Sect. 2.4).

Proposition 1 (Residual smoothing estimator) *After $k-1$ iterations, the bias corrected estimator can be explicitly written as*

$$\begin{aligned}\hat{m}_k &= S_1 Y + S_2(I - S_1)Y + \cdots \\ &\quad + S_k(I - S_{k-1}) \cdots (I - S_1)Y \\ &= \big[I - (I - S_k)(I - S_{k-1}) \cdots (I - S_1)\big]Y. \end{aligned} \tag{6}$$

An alternative approach is to estimate the bias by plugging in an estimator $\tilde{m} = S_2 Y$ for the regression function m into the expression of the bias (4). This produces the estimator

$$\tilde{b}_1 = (S_1 - I)S_2 Y$$

for the bias.

Proposition 2 (Plug-in estimator) *After $k-1$ iterations, plug-in bias estimator can be explicitly written as*

$$\begin{aligned}\hat{m}_k &= S_1 Y + (I - S_1)S_2 Y + \cdots + (I - S_1)(I - S_2) \cdots S_k Y \\ &= \big[I - (I - S_1)(I - S_2) \ldots (I - S_k)\big]Y. \end{aligned} \tag{7}$$

While in general, these two estimates for the bias lead to distinct bias corrected smoothers (6) and (7), they are identical when the same smoothing matrix is used at every step of the procedure.

Proposition 3 (Iterating the same smoothing matrix) *Taking $S = S_1 = S_2 = \cdots = S_k$, both the plug-in estimator and the residual smoothing estimator agree and the k^{th} iterated bias corrected smoother can be written as*

$$\hat{m}_k = \big[I - (I - S)^k\big]Y. \tag{8}$$

This closed form shows that the qualitative behavior of the sequence of iterative bias corrected smoothers $\widehat{m}_k$ is governed by the spectrum of $I - S$ (see Cornillon et al. 2011b). If the eigenvalues λ_j of $I - S$ are in $[0, 1]$ then as k tends to infinity, the bias converges to 0 and the variance increases to $n\sigma^2$.

In the univariate case, smoothers of the form (8) arise from the L_2-boosting algorithm with a symmetric base smoother S and a convergence factor μ_k equal to one (see Friedman 2001, for a definition of this factor). Thus we can interpret the L_2-boosting algorithm as an iterative bias reduction procedure in this special case. From a historical perspective, the idea of estimating the bias from residuals to correct a pilot estimator of a regression function goes back to the concept of *twicing* introduced by Tukey (1977) to estimate the bias of misspecified multivariate regression models. The idea of iterative debiasing regression smoothers is also present in Breiman (1999) in the context of the *bagging* algorithm. More recently, the interpretation of the L_2-boosting algorithm as an iterative bias correction scheme was alluded to in Ridgeway's discussion of the paper on the statistical interpretation of boosting of Friedman et al. (2000). Bühlmann and Yu (2003) present the statistical properties of the L_2-boosted univariate smoothing splines, while Di Marzio and Taylor (2008) describe the behavior of univariate kernel smoothers after a single bias-correction iteration.

2.3 Prediction with smoothers

The linear smoother defined by (3) predicts the conditional expectation of responses only at the design points. It is useful to extend regression smoothers to enable predictions at arbitrary locations $x \in \mathbb{R}^d$ of the covariates. Such an extension allows us to assess and compare the quality of various

smoothers by how well the smoother predicts new observations.

To this end, write the prediction of the linear smoother S at an arbitrary location x as

$$\hat{m}(x) = S(x)^t Y,$$

where $S(x)$ is a vector column of size n whose entries are the weights for predicting $m(x)$. The vector $S(x)$ is readily computed for many of the smoothers used in practice. For example, for a kernel smoother (with a bandwidth h), one readily obtains that

$$S(x) = \frac{1}{\sum_{l=1}^{n} K(\frac{x-X_l}{h})} \times \left(K\left(\frac{x-X_1}{h}\right), \ldots, K\left(\frac{x-X_n}{h}\right)\right)^t.$$

We want to find a similar equation for the IBR smoother $\hat{m}$. Writing the latter smoother as

$$\begin{aligned}\widehat{m}_k &= \widehat{m}_0 + \widehat{b}_1 + \cdots + \widehat{b}_k \\ &= S\big[I + (I-S) + (I-S)^2 + \cdots + (I-S)^{k-1}\big]Y \\ &= S\widehat{\beta}_k,\end{aligned}$$

it follows that we can predict $m(x)$ by

$$\widehat{m}_k(x) = S(x)^t \widehat{\beta}_k,$$
$$\text{with } \widehat{\beta}_k = \big[I + (I-S) + (I-S)^2 + \cdots + (I-S)^{k-1}\big]Y.$$

The sequence of parameters $\widehat{\beta}_k$ can be computed recursively by

$$\widehat{\beta}_k = Y + (I-S)\widehat{\beta}_{k-1}.$$

2.4 Stopping rules

As we can see from Eq. (8), the qualitative behavior of the iterated estimator is governed by the spectrum of $I - S$. For splines smoothers and kernel smoothers with a positive define kernel, the spectrum lies in the unit interval $[0, 1]$ (Cornillon et al. 2011b). The package **ibr** (Cornillon et al. 2011a) is implemented with these types of smoothers. It follows from Eq. (8) that as the number of iterations k goes to infinity, the sequence of iterated smoothers $\hat{m}_k$ tends to reproduce the raw data Y. Thus iterating the algorithm until convergence is not desirable. However, since each iteration reduces the bias and increases the variance, often a few iterations of the algorithm will produce a better smoother than the pilot smoother. This brings up the important question of how to decide when to stop the iterative bias correction process.

Viewing the latter question as a model selection problem suggests stopping rules based on Akaike Information Criterion, AIC (Akaike 1973), modified AIC criterion (Hurvich et al. 1998), Bayesian Information Criterion, BIC (Schwarz 1978), and Generalized Cross Validation, GCV (Craven and Wahba 1979). These selectors, all implemented in the **ibr** package, can be written in a common form

$$\underset{k\in\mathcal{K}}{\operatorname{argmin}}\big\{\log \widehat{\sigma_k}^2 + \Phi\big(\operatorname{tr}(S_k)\big)\big\},$$

where $\widehat{\sigma_k}^2 = \frac{1}{n}\|Y - \hat{m}_k\|^2$, ($\|\|$ is the usual Euclidean norm) and

$$\Phi_{\mathrm{AIC}}\big(\operatorname{tr}(S_k)\big) = 2\frac{\operatorname{tr}(S_k)}{n},$$
$$\Phi_{\mathrm{BIC}}\big(\operatorname{tr}(S_k)\big) = \log n\frac{\operatorname{tr}(S_k)}{n}$$
$$\Phi_{\mathrm{AIC_C}}\big(\operatorname{tr}(S_k)\big) = 1 + 2\frac{\operatorname{tr}(S_k)+1}{n-\operatorname{tr}(S_k)-2},$$
$$\Phi_{\mathrm{GCV}}\big(\operatorname{tr}(S_k)\big) = -2\log\left(1 - \frac{\operatorname{tr}(S_k)}{n}\right).$$

We are interested in choosing the best selector for the number of iterations k among the above listed procedures. It is instructive to observe how each of these criteria behave over the entire range of k, from zero to infinity. When the number of iterations k tends to infinity, $\operatorname{tr}(S_k)$ converges to n. This means that we are almost interpolating the data, which implies that the residual sum of square, and hence $\hat{\sigma}_k$, tends to zero. Thus for splines and kernel smoothers with positive definite kernels, the ratio $\operatorname{tr}(S_k)/n$ increases monotonically to one and the estimated variance decreases to zero with a growing number of iterations k.

Figure 1 shows the different qualitative behavior of the penalization term Φ. Both the AIC and BIC penalties are linear in $\operatorname{tr}(S_k)/n$, and reach 2 and $\log n$, respectively, at $\operatorname{tr}(S_k) = n$. For problem with large σ^2, the AIC and BIC criteria will select the large number of iterations k, producing smoothers that nearly interpolate the data, which defeats the purpose of smoothing. This behavior is consistent with the general experience in nonparametric smoothing, where it is well-known that AIC criterion has a noticeable tendency to select smoothing parameters that are smaller than needed. As this leads to undersmooth the data, Hurvich et al. (1998) introduced a corrected version of the AIC ($\mathrm{AIC_C}$) under the simplifying assumption that the nonparametric smoother $\hat{m}$ is unbiased. This assumption is problematic in our context, as IBR deliberately starts out with a very biased estimate. For these reasons, and because of the asymptotic results given in Theorem 2 in Cornillon et al. (2011b), we advocate using GCV as the default stopping rule and use it in our simulations.

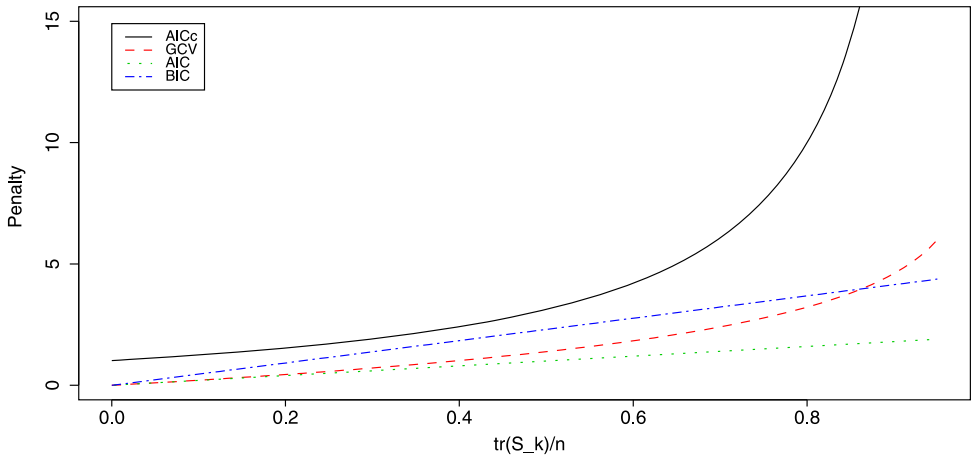


Fig. 1 Φ-penalties for various selectors as a function of $\mathrm{tr}(S_k)/n$

3 IBR with splines or kernels

Before discussing about initial values for IBR, let us present some IBR base smoothers S: thin-plate splines (TPS), Duchon splines and Gaussian kernel.

3.1 Thin plate splines

Suppose the unknown function m from $\mathbb{R}^d \to \mathbb{R}$ belongs to the Sobolev space $\mathcal{H}^{(\nu)}(\Omega) = \mathcal{H}^{(\nu)}$, where ν is an unknown integer such that $\nu > d/2$ and Ω is an open bounded subset of $\mathbb{R}^d$. Recall that thin plate splines (TPS) arise as the solution of the following minimization problem on $\mathcal{H}^{(\nu)}$ (see Gu 2002; Wood 2003)

$$\frac{1}{n}\big\|Y_i - f(X_i)\big\|^2 + \lambda J_\nu^d(f),$$

where

$$J_\nu^d(f) = \sum_{\alpha_1+\cdots+\alpha_d=\nu} \frac{\nu!}{\alpha_1!\cdots\alpha_d!} \times \int\cdots\int \left(\frac{\partial^\nu f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}\right)^2 dx_1\cdots dx_d.$$

The first part of the functional to be minimized controls the data fitting while the second part, $J_\nu^d(f)$, controls the smoothness. The trade-off between these two opposite goals is ensured by the choice of the smoothing parameter λ. The null space of $J_\nu^d(f)$ consists of polynomials with maximum degree of $(\nu-1)$. This subspace is of finite dimension $M = \binom{\nu+d-1}{\nu-1}$. Let us denote $\{\phi_1(.),\ldots,\phi_M(.)\}$ a basis of this subspace. If λ is known (and provided $\nu > d/2$ to ensure a continuous solution) the solution of the minimization problem is a TPS which has the following form:

$$g(x) = \sum_{j=1}^{M} \alpha_j \phi_j(x) + \sum_{i=1}^{n} \delta_i \eta_\nu^d\big(\|x - X_i\|\big)$$

where $\|.\|$ denotes the usual Euclidean norm. The vector $\delta \in \mathbb{R}^n$ of coefficients is subject to the constraint $T\delta = 0$ with the matrix T defined as $T_{ij} = \phi_j(X_i)$. Furthermore we have (where $\propto$ denotes proportional to):

$$\eta_\nu^d(r) \propto \begin{cases} r^{2\nu-d}\log(r) & d \text{ even}, \\ r^{2\nu-d} & d \text{ odd}. \end{cases}$$

To determine the vectors of coefficients α and δ, and thus the TPS solution, a closed form solution exists (see, for instance, Gu 2002). The TPS smoother can also be written as a linear smoother $S_\lambda Y$ where the dependency on d and ν is not written explicitly. Usually λ is unknown and has to be estimated from the data. Usual (classical) procedure is to minimize GCV criterion to determine an optimal $\hat{\lambda}$ that ensures the trade-off between smoothness and fitting. Moreover the order ν, which depends on unknown $m(.)$, is unknown and the classical approach is to choose an integer ν_0 without explicit statistical method to rely on. Usually it is chosen as the smallest integer value such as $\nu_0 > d/2$.

3.2 IBR with TPS

The approach proposed here is completely different: we deliberately choose a large λ (which is very easy) to ensure a very biased smoother. We choose ν_0 (as usual) the smallest integer value such as $\nu_0 > d/2$. But if the pilot estimator S_λ is a thin plate estimator of order $\nu_0 \le \nu$, then there exists an optimal number of iterations $k_n^\star$ such that the resulting smoother $\widehat{m}_k$ satisfies (Theorems 1 and 2 in Cornillon et al. 2011b)

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\big(\hat{m}_{k_n^\star}(X_j) - m(X_j)\big)\right)^2\right] = O\big(n^{-2\nu/(2\nu+d)}\big).$$

While this existence theorem does not provide any practical guidance for finding the optimal number of iterations $k_n^\star$, it can be used in conjunction with Li (1987) to prove optimality of GCV stopping rule (see Cornillon et al. 2011b). Thus, IBR ensures adaptivity: we do not know the true ν but if we choose a ν_0 as usual we are sure to get the optimal rate

of convergence and the optimal number of iterations with GCV. Recall that the classical TPS does not ensure adaptivity. Moreover the choice of starting point for λ and the minimization procedure is greatly simplified in IBR framework compared to the classical TPS. Currently, the optimality of GCV have only been proven for TPS smoothers, although our simulations strongly suggest that a similar result must hold for kernel based smoothers.

3.3 IBR with Duchon splines

It is well-known that beside computational problems, TPS suffer from the fact that the dimension M_0 of the null space of $J^d_{\nu_0}(.)$ increases exponentially with d due to the condition $\nu_0 > d/2$. In his seminal paper Duchon (1977) presents a mathematical framework that extends TPS. Noting that the Fourier transform is isometric the smoothness penalty $J^d_{\nu_0}(f)$ can be replaced by its squared norm in Fourier space, that is,

$$\int \left\| D^{\nu_0} f(t) \right\|^2 dt \quad \text{can be replaced by}$$
$$\int \left\| \mathcal{F}\big(D^{\nu_0} f\big)(\tau) \right\|^2 d\tau.$$

In order to solve the problem of exponential growth of the dimension of the null space of $J^d_{\nu_0}(.)$, and to get new interpolation methods, Duchon introduced a weighting function to define a new smoothness penalty:

$$J^d_{\nu_0,s}(f) = \int |\tau|^{2s} \left\| \mathcal{F}\big(D^{\nu_0} f\big)(\tau) \right\|^2 d\tau.$$

The solution of the new variational problem:

$$\frac{1}{n} \left\| Y_i - f(X_i) \right\|^2 + \lambda J^d_{\nu_0,s}(f),$$

is

$$g(x) = \sum_{j=1}^{M_0} \alpha_j \phi_j(x) + \sum_{i=1}^{n} \delta_i \eta^d_{\nu_0,s}\big(\|x - X_i\|\big),$$

provided that $\nu_0 + s > d/2$ and $s < d/2$. The $\{\phi_j(x)\}$ are still a basis of the subspace spanned by polynomial of degree $\nu_0 - 1$. We also have that:

$$\eta^d_{\nu_0,s}(r) \propto \begin{cases} r^{2\nu_0+2s-d} \log(r) & d \text{ if } 2\nu_0 + 2s - d \text{ is even,} \\ r^{2\nu_0+2s-d} & d \text{ otherwise} \end{cases}$$

still with the same constraint on coefficients: $T\delta = 0$.

For the special case $s = 0$, the Duchon splines reduces to the TPS. But if one wants to have a lower dimension for the null space of $J^d_{\nu_0,s}$, for instance a pseudo-cubic splines with an order $\nu_0 = 2$, one can choose (as suggested by Duchon 1977) $s = \frac{d-1}{2}$.

The same problem of the determination of λ exists for Duchon splines. It can be solved by using $\hat{\lambda}$ which optimizes the GCV criterion. This classical framework is implemented in the R package **mgcv**. In some circumstances (depending on the sample size n, on the dimension d, on the design, on m the unknown regression function and on the error distribution) **mgcv** optimization procedure fails and the user has to use low rank splines (see details in Wood 2003). To our knowledge, no data-driven method in **mgcv** is proposed to help the user in the choice of a sensible rank, but the **mgcv** methods are rather insensitive to this choice.

Obviously, as Duchon splines solution are of the same form as TPS, IBR method with Duchon splines base smoother can be used to circumvent the problem of TPS in high dimension. No choice of rank is needed and the optimization procedure to get an optimal number of iterations is straightforward.

3.4 Initial values of IBR

As discussed previously, the IBR method relies on the choice of a pilot smoother that over-smooths the data. In this section we discuss the choice of the smoothness of the pilot S. Our discussion in the section distinguishes splines (thin-plate or Duchon) based smoother and kernel based smoothers.

Splines smoothers depend on a regularization constant that pre-multiplies the roughness penalty. Qualitatively, "large" values of λ lead to over-smoothing the data whereas "small" values of λ produce under-smooth of the data. What value to take for large and small depend on the design, and it is difficult to define a range of value for λ that over-smooth every dataset without considering the data. Instead of focusing on selecting λ, every smoothing package (with splines smoother) defines and uses an equivalent degree of freedom (edf), taken to be the trace of the smoothing matrix, that is loosely interpreted as the number of independent parameters needed to represent the smoother.

Consider S the smoothing matrix of a splines smoother. The first M_0 eigen-values are equal to one (corresponding of the null space of $J^d_{\nu_0,s}(f)$) and the other eigen-values are all positive and depend on the value of the smoothing parameter λ. Hence

$$\mathrm{tr}(S) = \mathrm{edf} = M_0 + \mathrm{function}(\lambda).$$

As the end-user may not readily know the value of M_0, the requested argument `df` in **ibr** is not the edf itself, but a multiplicative coefficient applied to M_0 to get the edf, i.e., $\mathrm{edf} = M_0 \times \mathtt{df}$. Thus `df` should be chosen greater than 1 to ensure that $\mathrm{edf} > M_0$.

Let us give an example: suppose $d = 5$.

- If the user wants to use TPS ($s = 0$), to ensure continuity, the package requires that at least $\nu_0 = \lfloor d/2 \rfloor + 1 = 3$ and $M_0 = \binom{\nu_0+d-1}{\nu_0-1} = 21$. If `df` is chosen equal to 1.1, the initial TPS of order ν_0 will use a smoothing parameter λ whose trace equals 23.1.

- If the user wants to use Duchon splines, the package will set as default value the pseudo-cubic splines setting: $\nu_0 = 2$ and $s = (d-1)/2 = 2$. This setting leads to $M_0 = 6$. If `df` is chosen equal to 1.1, the initial Duchon splines of order 2 will use a smoothing parameter λ whose trace equals 6.6. Duchon base smoother is obviously more smooth than the TPS base smoother.

As an aside, starting with different `df` values leads to the same solution. Therefore we only report the results with $df = 1.001$.

For TPS, the dimension of the null space of the smoothness penalty M_0 grows exponentially with the number of covariates d for continuous thin plate splines. As a result, these smoothers may not be able to over-smooth the data even in moderate dimensions d. For example, if $d = 8$ then the minimal value for $M_0 = 792$, and it is obvious that in such a case, one needs to have a larger number of observations (at least $n = 792$) to simply be able to compute the smoother. Thus, to have the very smooth pilot required by our method, several thousands of data may be needed and this kind of large datasets is beyond the scope of our paper. This difficulty does not arise with the Duchon splines or kernel smoothers.

3.5 IBR with kernel smoother

Kernel smoothers do not suffer this kind of limitations and can be used with various (moderate) dimension d. In general, multivariate kernel smoothers are governed by a vector of bandwidth, one bandwidth for each explanatory variable. We recall the reader that we do not seek to select an "optimal" bandwidth, just some reasonable one that guarantees that the initial smoother over-smooths the data. As for smoothing splines, our implementation abstracts the particulars of the smoothing parameter (in this case the bandwidth) in favor of the edf. We can get a reasonable pilot smoother by using a single bandwidth on each variable if we standardize the data. Our experience suggests that we obtain better results by selecting a bandwidth that makes each one dimensional smoother (in each variable) having the same small effective degree of freedom, which is the `df` argument. Values of `df` we found to work well in our examples are 1.05 and 1.1.

4 Simulations

In this section, we present some of the results by applying our bias reduction procedure to simulated data sets and compare the results with various competing procedures implemented in R. Our comparisons vary the sample sizes ($n =$ 50, 100 and 200), the pilot smoothers, the noise over signal ratio, the type of errors and consider various functions in $\mathbb{R}^2$, $\mathbb{R}^5$ and $\mathbb{R}^{10}$. Let us expose the settings (all the codes are available on the authors' webpages).

4.1 Settings

4.1.1 Errors distribution

The distribution of errors ε is Gaussian and its variance is chosen such that the noise over signal ratio ($\mathrm{var}(\varepsilon)/\mathrm{var}(m)$) is 5 %, 10 % and 20 %. Each sample of the explanatory variables X_j ($1 \le j \le d$) is drawn uniformly and independently on $[0, 1]$.

4.1.2 Evaluations

Usually, in simulation studies, it is possible to evaluate the error between the true function and the estimator on a grid. Evaluating the error on such a grid in dimension 1 or 2 is easy but it becomes computationally intensive in dimension 5 or 10. For example in dimension 5, a regular grid with 10 points in each direction requires 10^5 points to evaluate the error. Therefore, we propose two measures of the error: the classical Mean Square Error (MSE) and the Mean Square Prediction Error (MSPE). We choose randomly 10 % of the data in the sample (excluding the extreme points in each direction) and denote this test set $\mathcal{T}$. The remaining 90 % of the data (denoted $\mathcal{L}$) is used to estimate m. The MSE is calculated as follows:

$$\mathrm{MSE} = \frac{1}{|\mathcal{L}|} \sum_{j \in \mathcal{L}} \big(\hat{m}(X_j) - m(X_j)\big)^2.$$

We compute the MSPE on the remaining 10 % (the test set denoted by $\mathcal{T}$):

$$\mathrm{MSPE} = \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} \big(\hat{m}(X_j) - m(X_j)\big)^2.$$

This measure gives an insight on the behavior of the smoothers between data points as the distance between data points increases with the dimension d.

4.1.3 Competitors

We use for kernel smoother different values of `df` argument, but only one `df` argument for TPS smoother and Duchon smoother (see Sect. 3.4). We compare with smoothers having an R package available: linear models, MARS algorithm of Friedman (1991) as implemented in the R package **mda**, projection pursuit regression using the R function **ppr** where the number of components is chosen by data splitting, additive models instantiated in the R package **mgcv**, low rank Duchon splines and classical TPS as implemented in **mgcv**, additive Boosting Bühlmann and Yu (2003) from **mboost**, and regression trees Breiman et al. (1984) found in the R package **rpart**.

4.1.4 Replicates

We replicate every setting 500 times. However, for iterative smoothing procedures such as IBR, GAM, or maximization procedures such as those based on the choice of an optimal $\hat{\lambda}$ by GCV (low rank TPS, Duchon splines in **mgcv**), it could happen (hopefully in a small number of cases) that the proposed estimator is not well conditioned and huge errors can occur. Such problem could easily be fixed by analyzing the results one by one. However, since we are computing hundreds and hundreds of runs, this is impossible so we decide to exclude for each method its 5 % poorest runs. Therefore all the results are computed with 475 replications.

4.1.5 Results

For each method, we calculate the mean and the standard deviation of the 475 mean square prediction errors. To help with the comparison of the different methods, we divide each value by the smallest one among all the methods which we interpret as a relative efficiency of a method against the best method: this gives the value one to the most efficient method. In almost all the simulations, the method having the smallest error considering the mean has also the smallest variance.

As the level of noise is increasing, the difference between smoothers are decreasing, so we only present the 10 % case. According to our simulations, ranking among MSE or MSPE are relatively the same, so we only present MSPE tables but will have discussions on the difference between MSE and MSPE for a limited number of smoothers.

4.2 Dimension two

In dimension two, we consider one additive function m_1 and three functions with interaction. All the results are given in Table 1 for MSPE.

$m_1(x_1, x_2) = 10(x_1 - 0.5)^2 + 5\exp\big(-(x_2 - 0.3)^2/0.09\big)$
(additive),

$m_2(x_1, x_2) = 10x_1^2 + \exp(2x_2)\{x_1 < 0.5\} + \exp(2x_2)$
(not smooth),

$m_3(x_1, x_2) = 10x_1x_2^2 + 2$
(pure interaction),

$$m_4(x_1, x_2) = 40\frac{\exp(8(x_1 - 0.5)^2 + (x_2 - 0.5)^2)}{\exp(8((x_1 - 0.2)^2 + (x_2 - 0.7)^2))}$$
(complex interaction).

As expected, GAM modeling (**mgcv**) and **gamboost** have the lowest MSPE for the additive function with a slight advantage for **mgcv** package. True nonparametric multivariate modeling such as MARS, IBR, low rank TPS or Duchon splines (**mgcv**) give similar results which are not that far from GAM.

For the other functions, IBR with TPS or Duchon splines are performing as good as low rank TPS or Duchon splines using **mgcv** package, and these 5 competitors are performing much better than structural ones. Notice that modifying the argument `df` can lead to very small improvement for kernel pilot smoother. This suggests that IBR is robust to the choice of `df`.

4.3 Dimension five

In dimension five, we decide to use an additive function with an interaction, a single index function, a three index function and a function with interactions:

$$m_1(x_1, \ldots, x_5) = 10\sin(\pi x_1x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5,$$

$$m_2(x_1, \ldots, x_5) = 10\exp\left(\frac{(2(x_1 - 0.5) - 2(x_2 - 0.5) + 2(x_3 - 0.5) + x_4 + x_5 - 1)^2}{5}\right),$$

$$m_3(x_1, \ldots, x_5) = 3\frac{(x_1 + 2x_4)}{\sqrt{5}} - 22\cos\left(\frac{\pi}{2}\frac{x_3 + 2x_5}{5}\right) + 10\exp\left(\frac{(2(x_1 - 0.5) - 2(x_2 - 0.5) + 2(x_3 - 0.5) + x_4 + x_5 - 1)^2}{5}\right),$$

$$m_4(x_1, \ldots, x_5) = 6\sin(\pi x_1x_2) + 10\cos\left(\sqrt{\left(x_3^2 + x_4^2\right)}\right) + 10x_4x_5.$$

Results of simulations are summarized in Table 2.

As in dimension 2, IBR (kernel and splines), **mgcv** low rank TPS and Duchon Splines globally outperform the other smoothers. As expected, it can be noticed that the relative difference is increasing with n but decreasing with d. But it comes as a relative surprise that even for moderate sample

Table 1 Dimension 2: ratio of the mean and variance of the MSPE over 475 simulations divided for all the competitors by the smallest value (mean or variance)

	R packages		**stats**	**rpart**	**stats**	**mda**	**mgcv**	**mboost**	**ibr**				**mgcv**	
$d=2$	n		par	tree	ppr	mars	gam	gamb	K1.05	K1.1	S	DS	ds	tps
m_1	50	me	19	12	4.6	1.9	1	1	2.3	2	2	1.8	1.8	2
		sd	16	11	5.7	1.8	1	1	2.7	2.4	2.3	2	2	2.3
	100	me	40	16	6.3	1.7	1	1.1	2.6	2.3	2	1.7	1.9	2.2
		sd	27	12	11	1.4	1	1	2.9	2.3	2	1.7	1.7	2
	200	me	73	23	5.7	1.8	1	1	2.4	2.5	1.9	1.7	1.9	2.5
		sd	44	16	17	1.6	1	1	2.4	2.5	1.7	1.5	1.7	2.1
m_2	50	me	3.2	5	2.4	1.9	1.8	1.9	1.4	1.5	1	1.1	1	1
		sd	2.4	3.5	2.2	1.4	1.4	1.4	1.3	1.5	1	1.1	1	1
	100	me	4.5	4.2	3.1	2.5	2.2	2.3	1.5	1.5	1	1.1	1.1	1
		sd	2.6	2.5	2.4	1.5	1.4	1.4	1.3	1.4	1	1.1	1.1	1
	200	me	5.8	4.2	3	2.9	2.6	2.7	1.4	1.4	1	1.1	1.1	1
		sd	3.1	2.4	2.1	1.6	1.3	1.4	1.2	1.3	1	1.1	1	1
m_3	50	me	9.7	21	1.9	10	9	9.2	1.1	1.1	1.2	1.1	1	1.2
		sd	7.8	22	1.6	8.5	7.9	8.8	1.2	1.2	1.3	1.1	1	1.2
	100	me	20	22	3.1	19	18	19	1	1	1.5	1.2	1.1	1.5
		sd	15	20	2.9	15	14	16	1.1	1.1	1.4	1.2	1	1.3
	200	me	43	26	3.7	39	37	38	1	1	1.7	1.3	1.3	1.9
		sd	25	20	3.2	24	22	24	1	1.1	1.3	1.1	1	1.4
m_4	50	me	7.2	9.6	2.6	5.8	5	4.9	1.2	1.2	1.1	1.1	1	1.1
		sd	6.8	11	2.9	5.7	5.1	5.2	1.2	1.4	1.1	1	1	1.2
	100	me	15	12	4.4	9.5	9	8.8	1	1.1	1.2	1.1	1	1.1
		sd	14	16	5.3	9.1	8.6	8.8	1	1.1	1.1	1	1	1.1
	200	me	29	11	6.9	19	18	18	1	1.1	1.4	1.2	1.2	1.4
		sd	25	12	8	16	16	16	1	1.1	1.3	1.1	1.2	1.3

size ($n=50$ that is $|\mathcal{L}|=45$ in learning set) the approximation error of GAM modeling can't be balanced by its low estimation error (compared to fully nonparametric modeling).

However, for the first function, which is nearly additive, GAM performs better than IBR for small sample size ($n=50$ and $n=100$). As n increases, IBR kernel performs better. One can think when n is sufficiently big, IBR kernel yields a better estimation of the slight interaction than the other smoothing procedures.

Roughly speaking the performances of IBR Duchon Splines or low rank **mgcv** Duchon Splines are similar. In Table 2, IBR appears to have a slight advantage but this slight advantage is due to the universal choice of the rank for the low rank Duchon Splines (arbitrarily chosen equal to $n/3$). A manual investigation shows that fine-tuning this choice of rank can lead to an advantage of low rank **mgcv** Duchon Splines over IBR Duchon Splines on MSE at the cost of increasing the MSPE. These two methods remain always in the same range and advantage will differ with the noise level, the rank or the type of function $m(.)$. Obviously, a good choice of rank in low rank smoothing splines can lead to improvement but this is beyond the scope of this paper. Again, the IBR procedure is robust to the choice of initial `df` in dimension five.

All these phenomena can also be observed with the MSE instead of MSPE. Moreover, when the noise level increases the differences between IBR and **mgcv** Duchon splines vanish. The complete results are available on the authors' webpage.

As a conclusion, in dimension two or five, without information on the structure of the regression function, one could advocate the use of IBR or low rank Duchon splines using **mgcv** package. Moreover, one can think that the distance between IBR (or low rank Duchon splines) and GAM gives an idea of the additivity of the function $m(.)$.

4.4 Dimension 10

Let us consider the same functions as in dimension five and just add five superfluous explanatory variables which are pure noise. Results are summarized in Table 3. The

Table 2 Dimension 5: ratio of the mean and variance of the MSPE over 475 simulations divided for all the competitors by the smallest value (mean or variance)

	R packages		**stats**	**rpart**	**stats**	**mda**	**mgcv**	**mboost**	**ibr**				**mgcv**	
$d=5$	n		par	tree	ppr	mars	gam	gamb	K1.05	K1.1	S	DS	ds	tps
m_1	50	me	2.1	6.3	2.8	1.6	1	1.6	1.4	1.3	1.7	1.3	1.8	–
		sd	1.9	4.9	2.7	1.6	1	1.4	1.6	1.4	2.2	1.3	1.8	–
	100	me	2.8	6.2	3.4	1.4	1	1.3	1	1	1.1	1	1.2	2.9
		sd	2.3	4.6	3.1	1.3	1	1.1	1.2	1.2	1.2	1.1	1.3	4
	200	me	5.4	9.9	4.9	2.3	1.9	2.1	1	1	1.3	1.1	1.2	1.3
		sd	4.5	7.3	5.1	2	1.7	1.7	1	1	1.3	1.1	1.2	1.3
m_2	50	me	5.4	6.1	2.2	5.5	5.9	4.6	1	1	1.1	1.1	1.8	–
		sd	5.2	5.5	2.8	5.2	5.3	4.7	1	1	1.1	1	1.7	–
	100	me	8.4	9.2	1.9	8.9	8.2	7.7	1.2	1.1	1	1.2	1.2	4.8
		sd	8.3	7.7	2.7	7.9	7.4	7.5	1.3	1.2	1	1.3	1.4	4.4
	200	me	12	12	1.5	12	11	11	1.4	1.4	1	1.3	1.3	2.1
		sd	11	9	2.2	9.6	9.3	9.6	1.3	1.3	1	1.4	1.5	1.7
m_3	50	me	5.1	6	2.6	5.7	5.5	4.8	1	1	1.1	1	1.7	–
		sd	5.1	5.2	2.8	5	5.3	4.7	1	1	1.1	1	1.6	–
	100	me	8.2	9.6	2.2	9.2	8	7.9	1.2	1.1	1	1.2	1.2	5
		sd	8.3	7.9	2.5	7.9	7.4	7.7	1.3	1.2	1	1.3	1.4	4.6
	200	me	12	13	1.6	12	11	11	1.4	1.4	1	1.3	1.3	2.1
		sd	11	9.5	2.2	9.5	9.3	9.6	1.3	1.3	1	1.4	1.5	1.6
m_4	50	me	2	5.9	2.5	2.1	1.6	2.1	1.2	1.2	1.5	1	1.1	–
		sd	1.7	4.7	2.2	1.9	1.4	1.9	1.3	1.2	1.9	1	1.2	–
	100	me	3	6.9	3.6	2.5	2.2	2.4	1.1	1.1	1.1	1	1.1	3
		sd	2.3	4.9	3.2	2	1.7	1.7	1.1	1.1	1.1	1	1.1	3.8
	200	me	5.6	10	5	4.3	3.9	4	1	1	1.3	1.1	1.1	1.3
		sd	4.2	7.7	5	2.9	2.7	2.8	1	1	1.3	1.1	1.2	1.4

minimum effective degree of freedom for thin plate splines smoother will be $M_0 = 6188$ which is far greater than the number of observations n. Thus thin plate splines smoother cannot be used in dimension 10.

Recall that we have 10 variables with only five active variables and five vacuous variables. This fact is unknown to the users. We construct an initial smoothing matrix using the 10 variables and iterate. So at each step of the algorithm all the variables (even the superfluous ones) are used. The results are not that different than those obtained in the previous section.

The main conclusion of that section is that the nonparametric methods (IBR and **mgcv** Duchon Splines) gives (i) similar results as GAM modeling (**mgcv** package) for nearly additive function (ii) much better results than GAM for non-additive function, even for very small sample in moderate dimension. This fact comes as a surprise as the common wisdom is to advocate structural modeling with a small sample sizes and moderate dimension.

Nonparametric methods appear relatively robust to the possible addition of pure noise variables to the set of explanatory variables. However, potential gains could be obtained if the initial smoothing matrix only contains the variables of interest or at least the variables of interest plus a limited number of unrelated variables.

5 Nonparametric smoothing with variable selection

The IBR method starts with a pilot smoother S for all the explanatory variables $X_1, \ldots, X_d$, and then iterates that smoother. But if some explanatory variables are not related to Y, it seems intuitively clear that excluding them should improve the predictive capability of the smoother. This suggests that variable selection may be beneficial. For computational reasons, we advocate using ascendent variable selection to construct more parsimonious multivariate smoothers.

To proceed with this variable selection procedure, we need to choose a criterion. Classical criterion for variable selection in linear models are AIC or BIC. The GCV criterion, which is well suited for splines smoothing, is also available. For example, let us assume that the selected criterion is BIC. Our forward variable selection procedure starts by building

Table 3 Dimension 10: ratio of the mean and variance of the MSPE over 475 simulations divided for all the competitors by the smallest value (mean or variance)

	R packages		**stats**	**rpart**	**stats**	**mda**	**mgcv**	**mboost**	**ibr**			**mgcv**
$d = 10$	n		par	tree	ppr	mars	gam	gamb	K1.05	K1.1	DS	ds
m_1	50	me	1.7	4.3	2.5	1.3	1	1.3	1.6	1.6	1.6	1.7
		sd	1.5	3.6	2.2	1.3	1	1.2	1.5	1.5	1.5	1.5
	100	me	2.5	5.3	3.3	1.2	1	1.2	2.2	2.2	2.2	2.5
		sd	2.4	4.7	3.6	1.3	1	1.1	2.4	2.3	2.4	2.5
	200	me	2.7	4.9	3.2	1.1	1	1.1	1.7	1.7	1.7	1.9
		sd	2.7	4.1	3.5	1.1	1	1	1.9	1.9	1.8	2.1
m_2	50	me	1.8	1.6	1.8	1.7	2	1.3	1	1	1.1	1.6
		sd	1.6	1.4	2.2	1.6	1.6	1.4	1	1	1.1	1.5
	100	me	2.9	3.2	1.5	3.1	3	2.4	1	1.1	1.1	1.9
		sd	3.2	3.3	2.3	3.2	3	2.8	1	1.2	1.1	1.9
	200	me	5.2	5.4	1.3	5.4	5.1	4.6	1	1.1	1.1	1.1
		sd	4.6	4.1	1.7	4.2	4.1	4.1	1	1.2	1.1	1.2
m_3	50	me	1.7	1.6	1.9	1.8	1.9	1.3	1	1	1.1	1.6
		sd	1.6	1.4	2.1	1.7	1.6	1.3	1	1	1	1.5
	100	me	2.8	3.3	1.6	3.2	2.9	2.6	1	1.1	1.1	1.8
		sd	3.1	3.3	2.3	3.2	2.9	2.9	1	1.2	1.1	1.9
	200	me	5.1	5.4	1.1	5.5	5	4.6	1	1.1	1.1	1.1
		sd	4.6	4	1.3	4.2	4.1	4	1	1.2	1.1	1.2
m_4	50	me	1.2	3.1	1.7	1.1	1.1	1.2	1.1	1.1	1	1.1
		sd	1.1	2.7	1.8	1.2	1.1	1.2	1	1	1	1.1
	100	me	1.4	3.3	2	1.2	1.1	1.1	1.1	1.1	1	1.1
		sd	1.3	2.9	2.2	1.3	1	1	1.1	1.1	1	1.1
	200	me	1.9	3.5	2.2	1.4	1.3	1.4	1.1	1.1	1	1.1
		sd	1.6	2.9	2.2	1.2	1.1	1.1	1.1	1.1	1	1.1

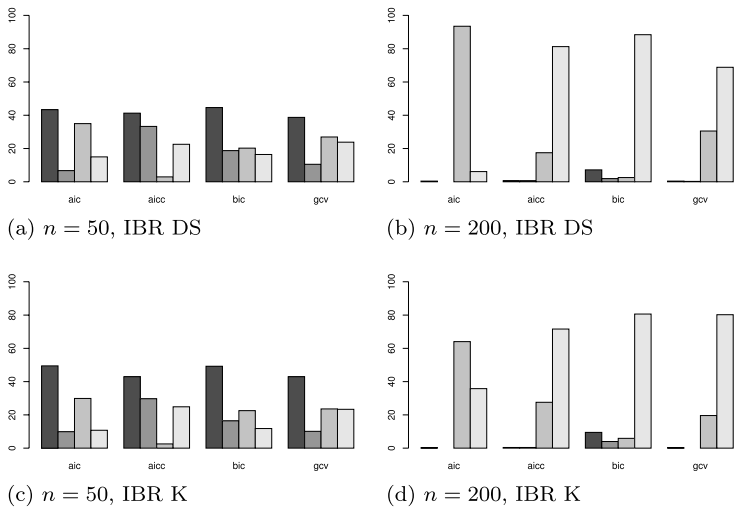
d univariate smoothers, one for each of the explanatory variables, and each smoother with the same equivalent degree of freedom. We apply the IBR algorithm to each of these univariate smoothers and select their respective optimal number of iterations, using GCV. Of these d smoothers, we select the one with the smallest BIC, and fix that variable. Next, we consider all $d-1$ bivariate smoothers that include the previously selected variable. Again, we apply the IBR algorithm and consider the bivariate model with the smallest BIC. If the latter BIC value is larger than the smallest BIC value from the univariate fit, we stop the forward fitting selection and return the univariate smoother. If not, then we consider all $d-2$ trivariate smoothers that extend the "best" bivariate smoother. We proceed with the forward selection until no improvement in the BIC is observed. This forward selection procedure has been implemented within the **ibr** package (function **forward**).

5.1 Criteria for variable selection

Here, we report on the performance of our variable selection method using the simulated data in Sect. 4.4. Again, we consider kernel based smoothers and Duchon splines smoother as our model may contain up to 10 variables, which makes TPS not practical. To help understand the results and provide further insights into the qualitative behavior of variable selection for IBR smoothing, we analyze the selected model as follows: We roughly divide the selection results into four categories:

- First category: the variable selection criterion leads to a selected model which misses some of the true variables and include some other which are pure noise ("wrong" category).
- Second category: the variable selection criterion leads to a selected model which misses only some true variables ("not enough" category).
- Third category: the variable selection criterion leads to a selected model which includes all the true variables and some other ("too many" category).
- Fourth category: the variable selection criterion leads to a selected model which is the good one ("exact" category).

As similar results were obtained for the other functions we only present the results for function 2.



(a) $n = 50$, IBR DS (b) $n = 200$, IBR DS

(c) $n = 50$, IBR K (d) $n = 200$, IBR K

Fig. 2 Variable selection features for the function 2. The barplot shows the percentage of occurrences of each category: category one ("wrong") *dark*, two ("not enough") *dark grey*, three ("too many") *grey* and four (exact) *light grey*

As shown in Fig. 2, the percentages of the first category are roughly the same for all the variable selection criteria when n is small. However when n is bigger, that percentage is bigger for BIC. The second category leads to models with poor prediction accuracy. It can be seen that the percentage of this category for BIC is greater than those obtained by the other criteria. That can partly explain the poor prediction accuracy of models selected with BIC because that criterion tends to select more parsimonious models. The percentages of the third category reveals that AIC (especially) and GCV (to a fewer extend) tend to select too many variables (compared to BIC). But since IBR is somewhat robust to the inclusion of a few vacuous variables (see Sect. 4.4), this does not appear to overly degrade the predictive capability of the resulting IBR smoother. The fourth category has a higher percentage for GCV (if using kernel pilot smoother) compared to AIC (and BIC) and explains why GCV is better at selecting good models for prediction. In conclusion, we advocate to use again GCV criteria in our function **forward**.

5.2 Simulation results for variable selection

In Table 4, we compare IBR with variable selection (using GCV) to its competitors available in R: the **leaps** package for classical multivariate regression, **mars**, **gam**, **gamboost** with their built-in selection procedure and ppr.

The conclusion are about the same as in dimension 5: IBR gives the best results except for the first function (nearly additive) with $n = 50$. Again, the argument `df` seems unimportant for the (kernel) pilot smoother: IBR is robust to the choice of `df`. Compared to dimension 5, it can be noticed that the variable selection slightly reduces the differences.

6 Real data example: Los Angeles ozone data

As a real world example, consider the classical data set of ozone concentration in the Los Angeles basin. This is as a standard dataset for comparing the performance of multivariate smoothers (Breiman 1996; Bühlmann and Yu 2003). The sample size of the data is $n = 330$ and the number of explanatory variables $d = 8$ (Pressure, Wind speed, Humidity, Temperature measured at Sandburg, Temperature measured at El Monte, Inversion base height, Pressure gradient, Inversion base temperature and Visibility). We compare our iterative bias procedure with existing methods.

We estimate mean squared prediction error $\mathbb{E}[(Y - \hat{m}(X))^2]$ by randomly splitting the data into 297 training observations and 33 test observations and averaging 50 times over such random partitions.

For the IBR smoother, we use a multivariate Gaussian kernel and select the bandwidth so that the univariate smoother in each of the variables has the same trace, i.e., the same effective degree of freedom, $df = 1.1$. We do so at each iteration.

For Duchon or IBR Duchon, since all the variables are not of the same range, we decide to scale the variables, again this is done at each iteration. The training part is scaled, the smoothers are evaluated and new values (centered by the mean and divided by the standard error evaluated on the training set) are predicted. Figure 3 summarizes the results.

Low rank Duchon Splines and ibrDS perform better than the others methods and lead to a reduction of the mean prediction error of more than 15 % over competing multivariate methods. Recall that L_2 boosting is a component-wise procedure (Friedman 2001) ideally suited to fitting constrained models such as additive models (Bühlmann and Yu 2003). In order to deal with possible interactions, Bühlmann and Yu (2006) include second order and quadratic interaction

Table 4 Ratio of the mean and variance of the errors over 475 simulations divided for all the competitors by the smallest value (mean or variance)

	R packages		**stats**	**stats**	**mda**	**mgcv**	**mboost**	**ibr**			**mgcv**
$d = 10$	n		par	ppr	mars	gam	gamb	K1.1	K1.3	DS	ds
m_1	50	me	1.7	3	1.5	1	1.6	1.6	1.5	1.7	2
		sd	1.8	2.7	1.6	1	1.5	1.9	1.8	1.7	1.9
	100	me	2.5	3.5	1.3	1	1.3	1.1	1	2.3	2.6
		sd	2.4	3.9	1.4	1	1.2	1.5	1.4	2.3	2.7
	200	me	5.2	6.3	2.2	1.9	2.2	1	1.1	4.7	3.8
		sd	4.1	5.4	1.7	1.5	1.5	1	1.1	3.8	3.2
m_2	50	me	1.8	2.4	2.2	2	1.6	1	1.4	1.4	2.1
		sd	1.5	2.3	1.7	1.6	1.5	1	1.5	1.3	1.6
	100	me	4.8	2.9	5.9	5	4.6	1	1.3	3.4	3.6
		sd	4.1	3.2	4.6	3.9	4	1	1.5	3.3	2.7
	200	me	8.2	2.1	9	7.9	7.6	1.1	1	5.7	1.9
		sd	6.9	2.6	6.4	6.2	6.2	1.1	1	5.4	1.8
m_3	50	me	2	2.4	2.3	2	1.7	1	1.3	1.5	2.1
		sd	1.7	2.4	1.9	1.6	1.5	1	1.3	1.4	1.7
	100	me	6.2	3.6	7	5.7	5.6	1	1.3	4	4
		sd	5	3.9	5.4	4.5	4.7	1	1.5	3.7	3.1
	200	me	9.1	1.8	9.4	8.1	8	1.1	1	5.9	1.9
		sd	7.1	2.1	6.7	6.5	6.3	1.1	1	5.5	1.8
m_4	50	me	1.8	2.3	1.5	1.2	1.6	1.1	1	1.1	1.5
		sd	1.6	2.1	1.3	1	1.4	1.2	1	1	1.3
	100	me	2.8	3.6	2.3	1.9	2.1	1	1	2	2.1
		sd	2.2	3.1	1.8	1.3	1.4	1	1	1.5	1.5
	200	me	5.6	6.6	4.2	3.8	4	1	1.1	4.3	3.1
		sd	4	5.4	3.1	2.6	2.6	1	1.1	3.2	2.6

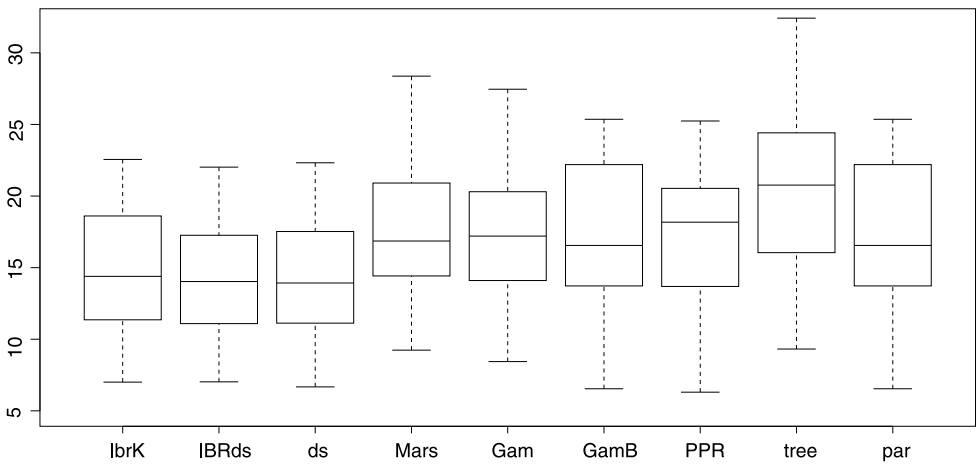


Fig. 3 Boxplot of the MPE for the different competing methods

terms within the L_2 boosting framework. The inclusion of higher order interaction terms increases the number of explanatory variables from 8 to 45. With the interaction terms included, the L_2 boosting proposed by Bühlmann and Yu (2003) yields an out of sample prediction MSE of 15.60. This result is obtained by discarding the two gross outliers in the 50 random partitions. Without removing these outliers, their results remain consistent with the results published in Bühlmann and Yu (2006). Our iterative bias reduction procedure is fully multivariate and finds directly an estimation of $m(X_1, \ldots, X_d)$ (where $d = 8$). As its results are better than additive models or models with low order interactions,

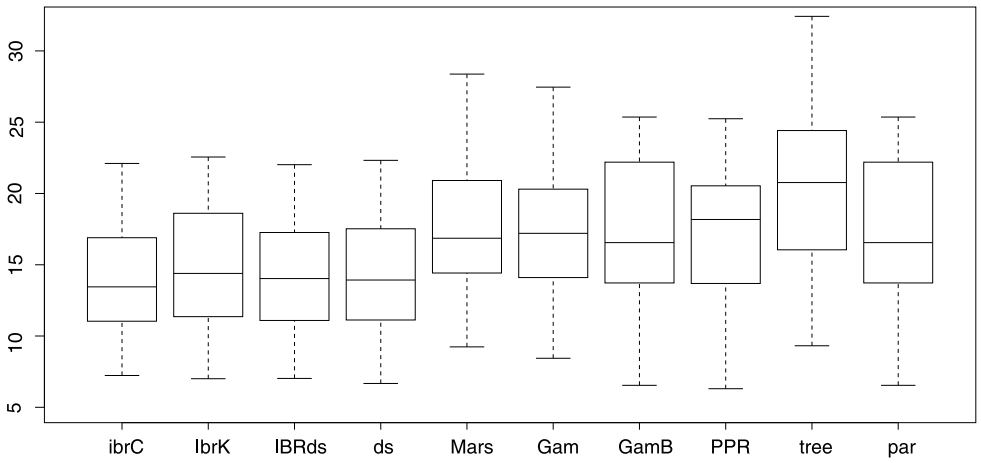


Fig. 4 Boxplot of the MPE for the different competing methods and the variable selection procedure

we can conclude that interaction of high order is significant for this dataset.

The previous results were obtained using all the 8 covariates. Further improvements are possible using variable selection. We apply our forward variable selection method to each of the 50 randomly split data (into 297 observations to fit the model and 33 observations to validate the predictions) to select the predictive variables using the GCV criterion. With high consistency, the procedure selects the 5 variables Wind, Humidity, Temp_Sand, Inv_Base_height, Pressure_Grad. Furthermore, with only these five variables, the mean predicted error drops to 13.8 (Fig. 4), which is a small improvement over the prediction error we had when using all the 8 variables.

7 Conclusion

Cornillon et al. (2011b) propose a new smoothing method IBR that has the desirable property of being simple and yet capable of adaptation, which suggests that it may be used to perform fully nonparametric smoothing in moderate dimensions. This paper compares this new method with classical and non-classical multivariate smoothing methods.

This simulation study shows that even for very moderate learning sample size (such as $n = 45$ or $n = 90$) in moderate dimension (up to $d = 10$) nonparametrics smoothers such as IBR (kernel or splines, package **ibr**) or low rank splines (Duchon or TPS, package **mgcv**) can lead to significant improvement over structurally constrained modeling such as GAM. These two kinds of modeling are very close in performances and can be thought as leading more or less the same results.

One can think that in the light of the results, the classical idea of quantifying the amount of non-additivity in the regression function $m(.)$ by measuring the distance between GAM modeling and a fully nonparametric modeling can be investigated in a practical manner.

Acknowledgements We would like to thank the associate editor and the referees for very valuable remarks and for pointing out to us the work of Duchon (1977).

References

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, B.F. (eds.) Second International Symposium on Information Theory, pp. 267–281. Academiai Kiado, Budapest (1973)

Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996)

Breiman, L.: Using adaptive bagging to Debais regressions. Tech. Rep. 547, Department of Statistics, UC Berkeley (1999)

Breiman, L., Freiman, J., Olshen, R., Stone, C.: Classification and Regression Trees, 4th edn. CRC Press, Boca Raton (1984)

Bühlmann, P., Yu, B.: Boosting with the l_2 loss: regression and classification. J. Am. Stat. Assoc. **98**, 324–339 (2003)

Bühlmann, P., Yu, B.: Sparse boosting. J. Mach. Learn. Res. **7**, 1001–1024 (2006)

Buja, A., Hastie, T., Tibshirani, R.: Linear smoothers and additive models. Ann. Stat. **17**, 453–510 (1989)

Cornillon, P.A., Hengartner, N., Matzner-Løber, E.: Iterative bias reduction multivariate smoothing in R: the IBR package (2011a). arXiv:1105.3605v1

Cornillon, P.A., Hengartner, N., Matzner-Løber, E.: Recursive bias estimation for multivariate regression (2011b). arXiv:1105.3430v2

Craven, P., Wahba, G.: Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math. **31**, 377–403 (1979)

Di Marzio, M., Taylor, C.: On boosting kernel regression. J. Stat. Plan. Inference **138**, 2483–2498 (2008)

Duchon, J.: Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: Shemp, W., Zeller, K. (eds.) Construction Theory of Functions of Several Variables, pp. 85–100. Springer, Berlin (1977)

Eubank, R.: Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York (1988)

Fan, J., Gijbels, I.: Local Polynomial Modeling and Its Application, Theory and Methodologies. Chapman & Hall, New York (1996)

Friedman, J.: Multivariate adaptive regression splines. Ann. Stat. **19**, 337–407 (1991)

Friedman, J.: Greedy function approximation: A gradient boosting machine. Ann. Stat. **28**, 1189–1232 (2001)

Friedman, J., Stuetzle, W.: Projection pursuit regression. J. Am. Stat. Assoc. **76**, 817–823 (1981)

Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Stat. **28**, 337–407 (2000)
Gu, C.: Smoothing Spline ANOVA Models. Springer, Berlin (2002)
Gyorfi, L., Kohler, M., Krzyzak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer, Berlin (2002)
Hastie, T.J., Tibshirani, R.J.: Generalized Additive Models. Chapman & Hall, New York (1995)
Hurvich, C., Simonoff, G., Tsai, C.L.: Smoothing parameter selection in nonparametric regression using and improved Akaike information criterion. J. R. Stat. Soc. B **60**, 271–294 (1998)
Lepski, O.: Asymptotically minimax adaptive estimation. I: Upper bounds. Opitmally adaptive estimates. Theory Probab. Appl. **37**, 682–697 (1991)
Li, K.C.: Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Ann. Stat. **15**, 958–975 (1987)
Ridgeway, G.: Additive logistic regression: a statistical view of boosting: discussion. Ann. Stat. **28**, 393–400 (2000)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
Simonoff, J.S.: Smoothing Methods in Statistics. Springer, New York (1996)
Tsybakov, A.: Introduction to Nonparametric Estimation. Springer, Berlin (2009)
Tukey, J.W.: Explanatory Data Analysis. Addison-Wesley, Reading (1977)
Wood, S.N.: Thin plate regression splines. J. R. Stat. Soc. B **65**, 95–114 (2003)
Wood, S.N.: Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Am. Stat. Assoc. **99**, 673–686 (2004)
Yang, Y.: Combining different procedures for adaptive regression. J. Multivar. Anal. **74**, 135–161 (2000)

Bibliographie

Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. Dans *Second international symposium on information theory*, réd. B.N. Petrov & B.F. Csaki, pp. 267–281. Academiai Kiado, Budapest.

Anevski D. & Soulier P. (2011). Monotone spectral density estimation. *The Annals of Statistics*, **39**(1), 418–438.

Angelov S., Harb B., Kannan S. & Wang L.S. (2006). Weighted isotonic regression under the l 1 norm. Dans *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 783–791. ACM.

Ansley C.F. & Kohn R. (1994). Convergence of the backfitting algorithm for additive models. *Journal of the Australian Mathematical Society (Series A)*, **57**(03), 316–329.

Ayer M., Brunk H.D., Ewing G.M., Reid W.T. & Silverman E. (1955). An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, pp. 641–647.

Barlow R.E., Bartholomew D.J., Bremner J.M. & Brunk H.D. (1972). *Statistical inference under order restrictions : The theory and application of isotonic regression.* J. Wiley.

Bartholomew D.J. (1959). A test of homogeneity for ordered alternatives. *Biometrika*, **46**(1/2), 36–48.

Bauschke H.H. & Borwein J.M. (1994). Dykstra's alternating projection algorithm for two sets. *Journal of Approximation Theory*, **79**(3), 418–443.

Bellman R.E. (1961). *Adaptive control processes : A guided tour.* Princeton University Press.

Best M.J. (1984). Equivalence of some quadratic programming algorithms. *Mathematical Programming*, **30**(1), 71–87.

Best M.J. & Chakravarti N. (1990). Active set algorithms for isotonic regression ; a unifying framework. *Mathematical Programming*, **47**(1), 425–439.

Boyle J.P. & Dykstra R.L. (1986). A method for finding projections onto the intersection of convex sets in hilbert spaces. *Lecture Notes in Statistics*, **37**(28-47), 4.

Breiman L., Freiman J., Olshen R. & Stone C. (1984). *Classification and regression trees.* CRC Press, 4 ed.

Breiman L. & Friedman J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, pp. 580–598.

Breiman L. & Spector P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review/Revue Internationale de Statistique*, pp. 291–319.

Briane M. & Pagès G. (2006). *Théorie de l'intégration*. Vuibert, 4 ed.

Brunk H.D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, **26**(4), 607–616.

Brunk H.D. (1956). On an inequality for convex functions. Dans *Proc. Amer. Math. Soc*, vol. 7, pp. 817–824.

Brunk H.D. (1958). On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, pp. 437–454.

Brunk H.D. (1965). Conditional expectation given a σ-lattice and applications. *The Annals of Mathematical Statistics*, **36**(5), 1339–1350.

Brunk H.D. (1970). Estimation of isotonic regression. *Nonparametric Techniques in Statistical Inference*, pp. 177–195.

Brunk H.D., Ewing G.M. & Utz W.R. (1957). Minimizing integrals in certain classes of monotone functions. *Pacific journal of mathematics*, **7**(1), 833–847.

Bühlmann P. & Yu B. (2003). Boosting with the l2 loss. *Journal of the American Statistical Association*, **98**(462), 324–339.

Buja A., Hastie T. & Tibshirani R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pp. 453–510.

Chakravarti N. (1989). Isotonic median regression : a linear programming approach. *Mathematics of operations research*, pp. 303–308.

Cheney W. & Goldstein A.A. (1959). Proximity maps for convex sets. Dans *Proc. Amer. Math. Soc*, vol. 10, pp. 448–450.

Cornillon P.A., Hengartner N. & Matzner-Løber E. (2011). Recursive bias estimation for multivariate regression smoothers. Rap. Tech., arXiv.

Craven P. & Wahba G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**(4), 377–403.

Cryer J.D., Robertson T., Wright F.T. & Casady R.J. (1972). Monotone median regression. *The Annals of Mathematical Statistics*, **43**(5), 1459–1469.

Di Marzio M. & Taylor C.C. (2008). On boosting kernel regression. *Journal of Statistical Planning and Inference*, **138**(8), 2483–2498.

Diggle P., Morris S. & Morton-Jones T. (1999). Case-control isotonic regression for investigation of elevation in risk around a point source. *Statistics in medicine*, **18**(13), 1605–1613.

Durot C. (2007). On the lp-error of monotonicity constrained estimators. *The Annals of Statistics*, **35**(3), 1080–1104.

Dykstra R.L. (1981). An isotonic regression algorithm. *Journal of Statistical Planning and Inference*, **5**(4), 355–363.

Dykstra R.L. (1983). An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, pp. 837–842.

Dykstra R.L. & Robertson T. (1982). An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, **10**(3), 708–716.

Efron B. (1983). Estimating the error rate of a prediction rule : improvement on cross-validation. *Journal of the American Statistical Association*, pp. 316–331.

Efron B. & Tibshirani R. (1997). Improvements on cross-validation : The. 632+ bootstrap method. *Journal of the American Statistical Association*, pp. 548–560.

Fan J. & Gijbels I. (1996). *Local polynomial modelling and its applications*, vol. 66. Chapman & Hall/CRC.

Fridlyand J., Snijders A.M., Pinkel D., Albertson D.G. & Jain A.N. (2004). Hidden markov models approach to the analysis of array cgh data. *Journal of multivariate analysis*, **90**(1), 132–153.

Friedman J.H. & Fisher N.I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, **9**(2), 123–143.

Friedman J.H. & Stuetzle W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, pp. 817–823.

Frisén M. (1986). Unimodal regression. *The Statistician*, pp. 479–485.

Gebhardt F. (1970). An algorithm for monotone regression with one or more independent variables. *Biometrika*, **57**(2), 263–271.

Geng Z. & Shi N.Z. (1990). Algorithm as 257 : isotonic regression for umbrella orderings. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **39**(3), 397–402.

Goldstein J. & Kruskal J.B. (1976). Least square fitting for monotonic functions having integer values. *Journal of the American Statistical Association*, **71**, 370–373.

Good I. & Gaskins R. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, pp. 42–56.

Goulden C.H. (1952). *Methods of statistical analysis*. Wiley, 2nd edition ed.

Green P.J. & Silverman B.W. (1994). *Nonparametric regression and generalized linear models : a roughness penalty approach*, vol. 58. Chapman & Hall/CRC.

Grenander U. (1956). On the theory of mortality measurement. part ii. *Skand. Akt.*, **39**, 125–153.

Grotzinger S.J. & Witzgall C. (1984). Projections onto order simplexes. *Applied mathematics & optimization*, **12**(1), 247–270.

Gu C. (2002). *Smoothing spline ANOVA models*. Springer Verlag.

Guyader A. (2011). Contributions à l'estimation non paramétrique et à la simulation d'événements rares. *Habilitation à diriger des recherches*.

Györfi L., Kohler M., Krzyżac A. & Walk H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York.

Hageman L.A. & Young D.M. (1981). *Applied iterative methods*. Academic Press.

Hall P. & Huang L.S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics*, **29**(3), 624–647.

Halperin I. (1962). The product of projection operators. *The Annals of Statistics*, **23**, 96–99.

Hanson D.L., Pledger G. & Wright F.T. (1973). On consistency in monotonic regression. *The Annals of Statistics*, pp. 401–421.

Härdle W. & Hall P. (1993). On the backfitting algorithm for additive regression models. *Statistica neerlandica*, **47**(1), 43–57.

Harezlak J. (1998). *Bump hunting in regression revisited*. Thèse de doctorat, University of British Columbia.

Hastie T.J. & Tibshirani R.J. (1990). *Generalized additive models*. Chapman & Hall/CRC.

Heckman N. (1992). Bump hunting in regression analysis. *Statistics & probability letters*, **14**(2), 141–152.

Hoeffding W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pp. 13–30.

Hörmander L. (2007). *Notions of Convexity*. Birkhäuser.

Horowitz J., Klemelä J. & Mammen E. (2006). Optimal estimation in additive regression models. *Bernoulli*, **12**(2), 271–298.

Hotelling H. (1941). Experimental determination of the maximum of a function. *The Annals of mathematical statistics*, **12**(1), 20–45.

Hupé P., Stransky N., Thiery J.P., Radvanyi F. & Barillot E. (2004). Analysis of array cgh data : from signal ration to gain and loss to dna regions. *Bioinformatics*, **20**(18), 3413–3422.

Hurvich C.M., Simonoff J.S. & Tsai C.L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **60**(2), 271–293.

Isac G. & Németh A.B. (1986). Monotonicity of metric projections onto positive cones of ordered euclidean spaces. *Archiv der Mathematik*, **46**(6), 568–576.

Isac G. & Németh A.B. (1987). Corrigendum to "monotonicity of metric projections onto positive cones of ordered euclidean spaces". *Archiv der Mathematik*, **49**(4), 367–368.

Ishkanian A.S., Malloff C.A., Watson S.K., DeLeew R.J., Chi B., Coe B.P., Snijders A., Albertson D.G., Pinkel D. & Marra M.A. (2004). A tiling resolution dna microarray with complete coverage of the human genome. *Nature genetics*, **36**(3), 299–303.

Jaillard S., Drunat S., Bendavid C., Aboura A., Etcheverry A., Journel H., Delahaye A., Pasquier L., Bonneau D., Toutain A., Burglen L., Guichet A., Pipiras E., Gilbert-Dussardier B., Benzacken B., Martin-Coignard D., Henry C., David A., Lucas J., Mosser J., David V., Odent S., Verloes A. & Dubourg C. (2010). Identification of gene copy number variations in patients with mental retardation using array-cgh : Novel syndromes in a large french series. *European journal of medical genetics*, **53**(2), 66–75.

Jong K., Marchiori E., Van Der Vaart A., Ylstra B., Weiss M. & Meijer G. (2003). Chromosomal breakpoint detection in human cancer. *Applications of Evolutionary Computing*, pp. 107–116.

Kohavi R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Dans *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2*, pp. 1137–1143. Morgan Kaufmann Publishers Inc.

Kruskal J.B. (1964). Nonmetric multidimensional scaling : a numerical method. *Psychometrika*, **29**, 115–129.

Lee C.I.C. (1983). The min-max algorithm and isotonic regression. *The Annals of Statistics*, **11**(2), 467–477.

Liu M.H. & Ubhaya V.A. (2009). An o (n) algorithm for weighted least squares regression by integer quasi-convex and unimodal or umbrella functions. *Computers & Mathematics with Applications*, **58**(4), 776–783.

Lombard P.B. & Brunk H.D. (1963). Evaluating the relation of juice composition of madarin oranges to percent acceptance of a taste panel. *Fd. Technol.*, **17**, 113–115.

Luss R., Rosset S. & Shahar M. (2011). Isotonic recursive partitioning. *Arxiv preprint arXiv :1102.5496*.

Mammen E., Linton O. & Nielsen J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, **27**(5), 1443–1490.

Mammen E., Marron J.S., Turlach B.A. & Wand M.P. (2001). A general projection framework for constrained smoothing. *Statistical Science*, pp. 232–248.

Mammen E. & Thomas-Agnan C. (1999). Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics*, **26**(2), 239–252.

Mammen E. & Yu K. (2007). Additive isotone regression. *Asymptotics : Particles, Processes and Inverse Problems, IMS Lecture Notes-Monograph Series*, **55**, 179–195.

Maxwell W.L. & Muckstadt J.A. (1985). Establishing consistent and realistic reorder intervals in production-distribution systems. *Operations Research*, pp. 1316–1341.

Menendez J.A. & Salvador B. (1987). An algorithm for isotonic median regression. *Computational Statistics & Data Analysis*, **5**(4), 399–406.

Meyer M. & Woodroofe M. (2000). On the degrees of freedom in shape-restricted regression. *The annals of Statistics*, **28**(4), 1083–1104.

Moreau J.J. (1962). Décomposition orthogonale d'un espace préhilbertien selon deux cônes mutuellement polaires. *C. R. Acad. Sci.*, **255**, 238–240.

Moscato P. (1989). On evolution, search, optimization, genetic algorithms and martial arts : Towards memetic algorithms. *Caltech concurrent computation program, C3P Report*, **826**, 1989.

Mukerjee H. (1988). Monotone nonparametric regression. *The Annals of Statistics*, **16**(2), 741–750.

Nadaraya E.A. (1964). On estimators regression. *Theory of Probability and its Applications*, p. 9.

Olshen A.B., Venkatraman E.S., Lucito R. & Wigler M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, **5**(4), 557–572.

Opsomer J.D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **73**(2), 166–179.

Opsomer J.D. & Ruppert D. (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, **25**(1), 186–211.

Pardalos P.M. & Xue G. (1999). Algorithms for a class of isotonic regression problems. *Algorithmica*, **23**(3), 211–222.

Pardalos P.M., Xue G.L. & Yong L. (1995). Efficient computation of an isotonic median regression. *Applied Mathematics Letters*, **8**(2), 67–70.

Picard F., Robin S., Lavielle M., Vaisse C. & Daudin J.J. (2005). A statistical approach for array cgh data analysis. *BMC bioinformatics*, **6**(1), 27.

Picard R.R. & Cook R.D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, pp. 575–583.

Pinkel D., Seagraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W.L., Chen C. & Zhai Y. (1998). High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, **20**, 207–211.

Polzehl J. & Spokoiny V.G. (2000). Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **62**(2), 335–354.

Reboul L. (2005). Estimation of a function under shape restrictions. applications to reliability. *The Annals of Statistics*, **33**(3), 1330–1356.

Restrepo Palacios A. & Bovik A.C. (1994). On the statistical optimality of locally monotonic regression. *Signal Processing, IEEE Transactions on*, **42**(6), 1548–1550.

Revuz D. & Yor M. (2005). *Continuous Martingales and Brownian Motion*, vol. 293 de *Grundlehren der matematishen Wissenshaften*. Springer Verlag, 3 ed.

Robertson T. & Waltman P. (1968). On estimating monotone parameters. *The Annals of Mathematical Statistics*, **39**(3), 1030–1039.

Robertson T. & Wright F.T. (1980). Algorithms in order restricted statistical inference and the cauchy mean value property. *The Annals of Statistics*, pp. 645–651.

Rosenzweig B.A., Pine P.S., Domon O.E., Morris S.M., Chen J.J. & Sistare F.D. (2004). Dye bias correction in dual-labeled cdna microarray gene expression measurements. *Environmental health perspectives*, **112**(4), 480.

Roundy R. (1986). A 98%-effective lot-sizing rule for a multi-product, multi-stage production/inventory system. *Mathematics of Operations Research*, pp. 699–727.

Rudin W. (1975). *Analyse réelle et complexe*. Masson, Paris.

Schmidt K.D. (1993). *Developpement of a precommercial thinning guide for black spruce*. Thèse de doctorat, University of New Brunswick, Faculty of Forestry.

Schrijver A. (1986). *Theory of Linear and Integer Programming*. John Wiley.

Schwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Silverman B.W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 97–99.

Silverman B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–52.

Snee R.D. (1977). Validation of regression models : methods and examples. *Technometrics*, pp. 415–428.

Snijders A.M., Nowak N., Segraves R., Blackwood S., Brown N., Conroy J., Hamilton G., Hindle A.K., Huey B., Kimura K., Myambo K., Palmer J., Ylstra B., Yue J.P., Gray J.W., Jain A.N., Pinkel D. & Albertson D.G. (2001). Assembly of microarrays for genome-wide measurement of dna copy number by cgh. *Nature Genetics*, **29**, 263–264.

Solinas-Toldo S., Lampel S., Stilgenbauer S., Nickolenko J., Benner A., Döhner H., Cremer T. & Lichter P. (1997). Matrix-based comparative genomic hybridization : biochips to screen for genomic imbalances. *Genes, chromosomes and cancer*, **20**(4), 399–407.

Stone C.J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, pp. 689–705.

Stout Q.F. (2008). Unimodal regression via prefix isotonic regression. *Computational Statistics & Data Analysis*, **53**(2), 289–297.

Thompson W.A. (1962). The problem of negative estimates of variance components. *The Annals of Mathematical Statistics*, **33**(1), 273–289.

Tibshirani R.J., Hoefling H. & Tibshirani R. (2011). Nearly-isotonic regression. *Technometrics*, **53**(1), 54–61.

Turner T.R. & Wollan P.C. (1997). Locating a maximum using isotonic regression. *Computational statistics & data analysis*, **25**(3), 305–320.

Ubhaya V. (1987). An o (n) algorithm for least squares quasi-convex approximation. *Computers & Mathematics with Applications*, **14**(8), 583–590.

Van Eeden C. (1956). Maximum likelihood estimation of ordered probabilities. *Indag. Math*, **18**, 444–455.

Van Eeden C. (1957). Maximum likelihood estimation of partially or completely ordered parameters, i. *Indag. Math*, **19**, 128–136.

Von Neumann J. (1950). *Functional Operators. II. The Geometry of Orthogonal Spaces*. Annals of Mathematics Studies, no. 22. Princeton University Press, Princeton, N. J.

Wahba G. (1990). *Spline models for observational data*, vol. 59. Society for Industrial Mathematics.

Wang Y. & Huang J. (2002). Limiting distribution for monotone median regression. *Journal of Statistical Planning and Inference*, **107**, 281–287.

Watson G.S. (1964). Smooth regression analysis. *Sankhyā : The Indian Journal of Statistics, Series A*, pp. 359–372.

Wong W.H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *The Annals of Statistics*, **11**(4), 1136–1141.

Wright F.T. (1978). Estimating strictly increasing regression functions. *Journal of the American Statistical Association*, pp. 636–639.

Wright F.T. (1981). The asymptotic behavior of monotone regression estimates. *The Annals of Statistics*, **9**(2), 443–448.

Wu W.B., Woodroofe M. & Mentz G. (2001). Isotonic regression : Another look at the changepoint problem. *Biometrika*, **88**(3), 793–804.

Zarantonello E.H. (1971). *Contributions to nonlinear functional analysis : proceedings*. N° 27. Academic Press.