Richesse et complexité des données fonctionnelles

Frédéric Ferraty[1], Philippe Vieu[2]

[1] Auteur correspondant : Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse Cedex, France, ferraty@cict.fr

[2] Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse Cedex, France

Résumé Les progrés récents en matière de stockage et de traitement des données se traduisent de plus en plus fréquemment dans de nombreux domaines scientifiques par la présence de données de type fonctionnel (courbes, images, ...). Les défis proposés aux statisticiens pour appréhender ce type de données ont abouti depuis quelques années à la construction de nombreuses méthodes statistiques. Il se trouve que la complexité de ce type de données amène une richesse d'information qu'une méthode statistique (aussi sophistiquée soit elle) arrive difficilement à capter, tandis que des techniques de boosting capables d'utiliser les complémentarités de différentes méthodes se révèlent souvent plus performantes. L'objectif de ce travail est d'illustrer ce point de vue au travers d'un problème couramment rencontré en pratique : celui de la prévision d'une variable réponse réelle à partir d'une variable explicative fonctionnelle. Un rapide tour d'horizon des méthodes habituellement utilisées sera effectué, et leur complémentarité sera mise en évidence au travers d'un jeu de données issu d'un problème de chimie quantitative.

Keywords : Analyse de données fonctionnelles, Boosting, Méthodes de sélection, Modèles fonctionnels, Régression, Spectrométrie, Statistique non-paramétrique.

1 Introduction

La plupart des domaines scientifiques font face à des situations où les données recueillies sont de nature continue (courbes, images, ...). On pourrait citer par exemple, sans chercher à être exhaustif, la biologie, la climatologie, l'économétrie, la chimie quantitative, ... Bien évidemment, ces données continues ne sont en réalité observées que sur une grille formée par un ensemble fini de points de discrétisation, de telle sorte que l'on peut penser à les traiter au moyen des outils usuels (paramétriques ou non) de la statistique multidimensionnelle. Cette approche naïve des choses peut s'avérer intéressante dans des situations où les points de discrétisation sont peu nombreux et relativement distants les uns des autres. Depuis quelques années les moyens technologiques en matière de recueil (et de stockage) de données se sont considérablement accrus, amenant des grilles d'observations de données fonctionnelles de plus en plus fines et rendant inadéquates les méthodes statistiques multivariées usuelles, non seulement du fait des problèmes de grande dimension liés au nombre important de variables mais aussi du fait de la corrélation importante existant entre deux observations proches d'un même phénomène continu. Cette situation paradoxale conduirait à considérer que, loin d'être un avantage pour la connaissance des phénomènes, l'abondance de données aboutirait à une détérioration des résultats statistiques. Ce paradoxe était tellement fort et présent chez les statisticiens dans les années

80, que nos collègues anglosaxons allèrent jusqu'à invoquer la main du diable en imposant à la communauté statistique le terme de malédiction de la dimension (littéralement, "curse of dimensionality"). Dans une vision plus scientifique des choses, la communauté statistique s'est d'avantage attachée à relever ce qui en réalité constituait plus un défi qu'une malédiction. Ainsi se sont développés des outils/modèles propres aux problèmes multivariés (tels ceux de la statistique semi-paramétrique) qui sont hors de propos ici, mais aussi des outils/modèles propres aux problèmes fonctionnels capables de prendre en compte l'aspect continu des données et de tirer toute la richesse de leurs observations en des points de mesure très rapprochés.

La communauté statistique française a souvent joué un rôle moteur sur cette thématique. S'il fallait ne retenir que quelques exemples, on pourrait citer à cet égard le papier [13] qui fût un des premiers (le premier peut-être) à envisager des variables statistiques continues/fonctionnelles, l'article [11] qui est un des piliers théoriques abondamment utilisé encore de nos jours lors de l'étude des propriétés mathématiques des estimateurs issus de nombreux modèles pour variables fonctionnelles, celui de [5] concernant l'Analyse en Composante Principales de courbes, celui de [7] présentant les fondements théoriques pour l'analyse de séries temporelles par le biais de l'étude statistique de leurs trajectoires découpées en morceaux continus, ou bien les travaux récents de [21] posant les bases de l'analyse nonparamétrique de données fonctionnelles. La statistique pour variables fonctionnelles a été largement popularisée depuis une quinzaine d'années, grâce en particulier à l'impulsion de Jim Ramsay (voir les ouvrages généraux [28], [29] et [30]). L'engouement provoqué dans la communauté scientifique par les problèmes (autant théoriques qu'appliqués) inhérents à ce nouveau type de données s'est traduit par un nombre considérable de travaux récents qu'il serait vain de vouloir décrire exhaustivement ici, mais dont le lecteur pourra avoir une idée en consultant les ouvrages généraux [28], [29], [30], [21] et [20] ou bien en se rapportant aux divers numéros spéciaux que de nombreuses revues internationales de premier plan ont récemment consacré à ce thème (voir [12], [25], [33] et [17]). Une fois encore il faut souligner l'impact de la communauté statistique française dans ces développements récents, notamment sous l'impulsion du groupe STAPH de Toulouse (http ://www.math.univ-toulouse.fr/staph/) dont les ouvrages généraux [21], [17] et [20] sont les émanations directes.

Cette bibliographie, nécessairement restreinte (et donc arbitraire ...), met en évidence le fait que le statisticien dispose à l'heure actuelle d'un éventail de méthodes relativement important. Plus encore que pour des données multidimensionnelles classiques, la richesse et la complexité des données fonctionnelles font que chacune de ces méthodes a ses propres avantages (et donc ses propres limites) ; avantages souvent directement hérités du modèle sous jacent pour lequel la méthode a été construite. Ainsi, il va devenir intéressant de s'orienter vers des techniques permettant de combiner les avantages respectifs de chaque méthode (techniques habituellement connues sous le terme de "boosting"). L'objet de ce travail est d'illustrer cette idée à partir d'un problème classique de régression et au travers d'un jeu de données très abondamment utilisé dans la littérature et qui est issu d'un problème de chimie quantitative en industrie agroalimentaire.

Notre travail est structuré de la manière suivante. Nous partirons d'un problème concret de chimie quantitative (voir Paragraphe 2.1) concernant la prédiction du taux d'un certain composant chimique à partir de l'analyse spectrométrique d'une substance donnée. Nous évoquerons dans le Paragraphe 2.2 les limites des méthodes statistiques usuelles pour faire face à ce genre de données, puis nous formaliserons le problème dans le

Paragraphe 2.5 au travers d'un modèle de régression avec variable réponse réelle (le taux à prédire) et variable explicative fonctionnelle (la courbe spectrométrique). Nous évoquerons ensuite dans les Paragraphes 3 et 4 les principaux modèles statistiques récemment introduits dans ce cadre fonctionnel, et nous verrons comment les méthodes de prédiction qui leur sont associées se comportent sur le problème de spectrométrie décrit précédemment. Plus précisément, le Paragraphe 3 sera centré sur les modèles dits sélectifs ("sparse models" en anglais) dont la philosophie générale est basée sur l'extraction d'un petit nombre de points de discrétisation de la variable fonctionnelle qui permettront de transformer le problème fonctionnel en une étude de régression multidimensionnelle, tandis que le Paragraphe 4 sera centré sur des modèles que nous qualifierons de "purement fonctionnels" en ce sens qu'il chercheront à utiliser l'intégralité du prédicteur fonctionnel continu. Ces deux Paragraphes 3 et 4 seront l'occasion de présenter les diverses modélisations du problème de régression pour données fonctionnelles qui existent dans la littérature et d'évoquer les différentes méthodes d'estimation que chacun de ces modèles peut amener à construire. Le Paragraphe 5 permettra d'illustrer, sur ce problème particulier de spectrométrie, la richesse des jeux de données fonctionnelles et la nécessaire complémentarité des diverses approches statistiques que l'on doit mettre en œuvre pour les analyser.

Le jeu de données que nous utiliserons pour illustrer notre propos (voir Paragraphes 2.1 et 2.5) est très couramment utilisé dans la littérature fonctionnelle comme base d'illustration de toute nouvelle méthode statistique. Nous profiterons de ce travail pour dresser en appendice une liste exhaustive de toutes les méthodes déjà testées sur ces données et des résultats qu'elles ont donnés.

2 Un problème concret de régresssion sur variable fonctionnelle

2.1 Un jeu de courbes spectrométriques

L'abondance des jeux de données fonctionnelles actuellement disponibles offre un grand choix de possibilités d'applications au statisticien désireux d'illustrer le comportement d'une nouvelle méthode. Étant donné que notre propos est d'illustrer la complémentarité des diverses approches proposées dans la littérature il est naturel de concentrer notre travail sur l'exemple qui a été le plus souvent étudié. Cet exemple concerne l'industrie agroalimentaire et plus particulièrement un problème de contrôle de qualité sur de la viande hachée. La variable fonctionnelle est donnée par la courbe d'absorbance de la lumière en fonction de la longueur d'onde, courbe obtenue au moyen d'une technique classique (et peu onéreuse) de spectrométrie dans le proche infrarouge. Le jeu de données dont nous disposons est constitué de 215 courbes d'absorbance (voir Figure 1).
Depuis l'article [6] ces données ont été abondament étudiées dans la littérature et nous renvoyons à [21] pour une présentation plus détaillée. En présence d'un tel échantillon, les problèmes statistiques qui peuvent se poser sont les mêmes qu'en statistique multivariée classique. Il peut s'agir de problèmes de type descriptif du type *"Peut-on classer ces courbes (et comment) en plusieurs catégories ?"*; et dans ce cas des techniques de classification devront être développées. Il peut s'agir de problèmes de prédiction *"Peut-on prédire à partir de la courbe χ une autre caratéristique (fonctionnelle ou non) Y ?"*; et dans ce cas des techniques de régression (resp. de discrimination) devront être développées

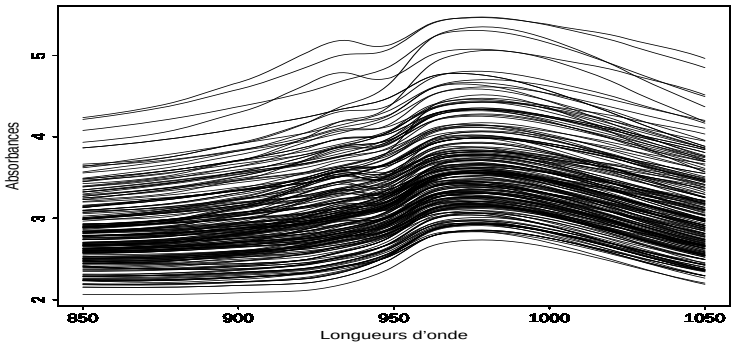

FIG. 1 – Les 215 courbes spectrométriques

si Y est quantitative (resp. si Y est qualitative). Face à d'autres questions du type *"Certaines parties de la variable χ, par exemple certaines régions du spectre ou simplement quelques points de discrétisations bien choisis, seraient-ils plus informatifs que d'autres ?"*, les techniques statistiques à développer pourront s'inspirer des méthodes de sélection de variables.

2.2 Les limites de la statistique classique

Il est clair que, bien qu'il s'agisse d'un phénomène continu la courbe l'absorbance n'est en réalité observée que sur une grille finie. Sur cet exemple, la grille est formée de 100 longueurs d'onde $\lambda_1, \dots \lambda_{100}$ équidistantes sur l'intervalle $[850, 1050]$ (intervalle correspondant au proche infra-rouge). Ainsi, les valeurs réellement observées ne sont que les 100 discrétrisations de chacune de ces 215 courbes que nous noterons désormais

$$X_i^j = \chi_i(\lambda_j), i = 1, \dots 215, j = 1, \dots 100.$$

Une approche naïve conduirait à considérer ces données comme un échantillon classique multivarié de dimension p et à utiliser les diverses approches statistiques classiques pour les analyser. Une telle approche pourrait s'avérer utile en petite dimension (c'est à dire pour un faible nombre de points de discrétisation), mais une des premières caractéristiques des jeux de données fonctionnelles est d'être de très grande dimension. Ici on a déjà $p = 100$, mais les spectromètres modernes actuellement disponibles sur le marché peuvent fournir une grille avec plusieurs milliers de points. La deuxième caractéristique importante de ce type de données, liée directement à la nature continue des phénomènes étudiés, est la très forte colinéarité existant entre les variables discrétisées X^j qui constitue un obstacle de taille à l'utilisation de la plupart des méthodes usuelles de la statistique multivariée.

2.3 La modélisation fonctionnelle

Au vu de ce qui précède il apparaît naturel de considérer les jeux de données fonctionnelles pour ce qu'ils sont réellement : c'est à dire comme des échantillons de n observations $(\chi_i, i = 1, \dots n)$ d'un objet aléatoire continu :

$$\chi_i = \chi_i(\lambda),\ \lambda \in [850, 1050], i = 1, \dots n.$$

Plus formellement, on définit un échantillon de données fonctionnelles comme étant une famille (χ_i) d'observations d'une variable aléatoire χ à valeurs dans un espace de dimension infini. Même si elle est explicitement satisfaite dans l'exemple spectrométrique précédent, la notion d'indépendance n'est pas requise dans une telle définition générale afin d'autoriser aussi la modélisation de problèmes où les χ_i sont des morceaux continus successifs de trajectoire d'un processus à temps continu. Ces problèmes ne sont pas à l'ordre du jour de ce travail et nous renvoyons à [7] (resp. à [21]) pour une modélisation linéaire (resp. non linéaire) de tels phénomènes fonctionnels dépendants.

Outre le fait qu'une telle modélisation va pouvoir donner naissance à de nombreuses méthodologies adaptées à ce type de données, on peut d'ores et déjà juger de son intérêt au travers de la constatation suivante. Si nous revenons à notre problème de spectrométrie, on sait que pour des raisons liées à la calibration de l'appareillage de spectrométrie le décalage vertical observé sur la Figure 1 est totalement non-informatif (voir [21] pour une discussion plus approfondie). Ainsi, la pratique courante amène à travailler sur les dérivées des courbes plutôt que sur les courbes elles-mêmes. Plus précisément, les données sur lesquelles la plupart des méthodes statistiques sont élaborées sont les dérivées secondes des courbes d'absorbance présentées dans la Figure 2.

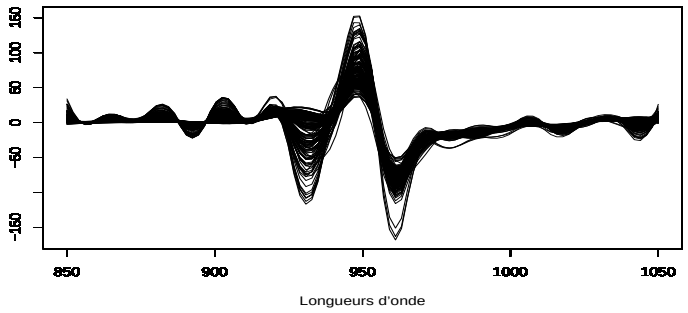

FIG. 2 – Les 215 dérivées seconde des courbes spectrométriques

Il est clair que ce pré-traitement des données, basée sur une opération de dérivation, n'est rendu possible que grâce à la vision fonctionnelle du problème qui est inhérente au modèle de régression posé dans ce paragraphe.

2.4 Les défis posés à la communauté statistique

Partant d'une telle modélisation fonctionnelle, les questions d'ordre méthodologique sont nombreuses. Peut-on (et comment) adapter les techniques statistiques classiques à ces échantillons d'un genre nouveau ? Doit-on au contraire redéfinir de nouvelles méthodes statistiques ? Peut-on se contenter de voir les jeux de données fonctionnelles comme des problèmes classiques de statistique en grande dimension et donc de les traiter au moyen des outils modernes basées sur les idées de sélection de modèle/variables ? Ou bien, au contraire, l'aspect fonctionnel qui se traduit par une très forte corrélation entre les variables doit-il être déterminant dans la construction de nouvelles méthodes statistiques ?

Bien entendu, ces questions méthodologiques entrainent de nombreuses questions touchant à la fois aux aspects mathématiques (validation asymptotique des méthodes proposées) et aux aspects appliqués (mise en œuvre et faisabilité des procédures). Nous es-

sayerons dans ce travail, en restant dans le cadre spécifique d'un problème de régression, de faire un tour d'horizon des principales méthodes qui existent déjà. Nous illustrerons leur comportement, et leur évidente complémentarité, au travers du problème de spectrométrie présenté ci-dessus.

2.5 Le problème de régression

Pour ce qui concerne ce travail, nous resterons dans le cadre simple d'un problème de prédiction d'une variable réelle Y, avec explicative fonctionnelle. Il s'agit d'une question classique en chimie quantitative, puisque la spectrométrie dans le proche infrarouge est beaucoup plus rapide et économique à effectuer qu'une analyse chimique directe d'une substance. Ainsi, pour le problème de contrôle de qualité alimentaire décrit ci-dessus, on dispose pour chacun des 215 morceaux de viande de la mesure Y_i du taux de graisse. L'objectif est de comprendre, à l'aide de cet échantillon, la relation existant entre la courbe spectrométrique χ et le taux de graisse Y, afin de pouvoir directement prédire le taux de graisse d'un nouveau steack haché en procédant simplement à son analyse spectrométrique. Concrètement, cela se traduit par un problème de régression avec variable explicative fonctionnelle (la courbe d'absorbance χ) et variable réponse scalaire (le taux de graisse Y).

La nature fonctionnelle de la variable χ amène diverses modélisations, et donc diverses méthodes statistiques, que nous passerons rapidement en revue dans ce travail et dont nous illustrerons les avantages et les inconvénients au travers de leurs comportements sur le problème concret de qualité alimentaire décrit ci-dessus. Pour tester ces comportements, notre échantillon de 215 couples (χ_i, Y_i) sera décomposé en un échantillon d'apprentissage (noté $\mathcal{A}$) de taille 160 sur lequel seront construites les diverses méthodes statistiques et un deuxième échantillon (noté $\mathcal{T}$) de taille 55 sur lequel les performances prédictives de ces méthodes seront mises à l'épreuve. On mesurera la performance d'une méthode $\mathcal{M}$, à partir d'un critère d'erreur quadratique :

$$ERR_{\mathcal{M}} = \frac{1}{55}\sum_{i\in\mathcal{T}}(\hat{Y}_i - Y_i)^2,$$

où $\hat{Y}_i$ est la prédiction pour Y_i obtenue pour chaque nouvelle courbe χ_i, $i \in \mathcal{T}$ à partir de la méthode $\mathcal{M}$.

Tout au long de ce travail nous illustrerons notre propos à partir de ce jeu de données et en calculant pour diverses méthodes cette erreur de prédiction. Comme il s'agit d'un jeu de données très couramment utilisé dans la littérature fonctionnelle, il est naturellement hors de question de décrire en détails toutes les procédures déjà testées sur ces données. Nous présenterons cependant dans l'appendice un récapitulatif des résultats que ces diverses méthodes de régression fonctionnelle ont fournis sur ces données.

3 Méthodes basées sur des modèles multivariés sélectifs

3.1 Introduction aux modèles sélectifs

La notion de modèle sélectif ("sparse model") s'est abondamment développée ces dernières années dans des contextes de régression multivariée pour lesquels le nombre

de variables est très grand (comparativement au nombre d'individus). Partant du fait que chaque donnée fonctionnelle (i.e. chaque courbe χ_i) n'est en réalité observée qu'en un nombre fini (mais grand) de points

$$X_i^j = \chi_i(t_j), j = 1, \dots p,$$

une première approche consiste à faire abstraction de l'aspect fonctionnel des variables χ_i et d'appréhender ce problème en dimension finie p. Dans ce contexte, il convient à la fois de choisir les points de la courbe qui sont le plus pertinents pour prédire la réponse Y et d'effectuer la prédiction de Y à partir de ces points là. Plus précisément, le modèle s'écrit à partir d'un sous-ensemble de variables $\{j_1, \dots, j_q\} \subset \{1, \dots, p\}$ sous la forme

$$Y_i = r(X_i^{j_1}, \dots, X_i^{j_q}) + \epsilon,$$

où

$$X_i^{j_k} = \chi_i(t_{j_k}),\ i = 1, \dots, n,\ k = 1, \dots, q.$$

Bien sûr, un des objectifs majeurs de ce genre d'approche est d'arriver à construire un ensemble de prédicteurs de dimension q très largement inférieure à la dimension initiale p du problème. Dans ce qui suit, nous allons présenter trois techniques construites sur ce schéma. La première sera basée sur une modélisation linéaire de la fonction de régression r (voir Paragraphe 3.2), tandis que la seconde (voir Paragraphe 3.3) sera basée sur une modélisation non-paramétrique. Finalement la complémentarité de ces deux approches sera mise en évidence au travers d'une troisième méthode décrite dans le Paragraphe 3.4. Dans le Paragraphe 3.5 toutes ces méthodes seront commentées à partir de leurs performances prédictives sur le problème spectrométrique précédent.

3.2 Construction d'un modèle sélectif linéaire

Les méthodes de choix de variables en régression linéaire multiple ont été intensément étudiées dans la littérature. Parmi ces méthodes certaines ont pour objectif de réduire au minimum le nombre de variables avec en ligne de mire des applications aux problèmes de très grande dimension. C'est naturellement ce dernier type de techniques de construction de modèle sélectif que nous allons chercher à utiliser dans le contexte fonctionnel. Partant d'un modèle linéaire de dimension p

$$Y_i = a_0 + \sum_{j=1}^{p} a_j \chi_i(t_j) + \epsilon,$$

la méthode LASSO introduite par [31] consiste à estimer les coefficients linéaires a_j en minimisant un critère de moindres carrés pénalisés

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} a_j \chi_i(t_j) \right)^2 + \lambda \sum_{j=1}^{p} |a_j|.$$

La pénalisation de type L_1 assure la nullité d'une grande majorité des coefficients estimés $\hat{a}_j$. D'un point de vue calculatoire, ce problème de minimisation peut-être résolu de manière rapide et efficace au moyen de l'algorithme LARS introduit par [14]. Pour ce qui concerne l'exemple spectrométrique présenté ci-dessus, les résultats sont donnés dans

la Table 1 : l'algorithme LARS amène un modèle linéaire sélectif de dimension réduite $q = 10$ signifiant que seuls 10 parmi les 100 points de discrétisation des courbes spectrométriques vont intervenir dans le modèle. Ce modèle de dimension 10 conduit à une erreur de prédiction sur l'échantillon test $\mathcal{T}$ de 8.4 (erreur dont nous pourrons juger par la suite qu'elle est relativement importante).

Dimension du problème	Nombre de points sélectionnés	Erreur de prédiction
$p = 100$	$q = 10$	$ERR_{LARS} = 8.4$

TAB. 1 – Résultats du modèle sélectif linéaire sur les données spectrométriques

3.3 Construction d'un modèle sélectif nonparamétrique

Pour pallier à la rigidité de la modélisation précédente, imposée par la condition de linéarité, on peut chercher à construire des modèles sélectifs non-linéaires

$$Y_i = r(\chi_i(t_{j_1}), \ldots, \chi_i(t_{j_q})) + \epsilon,$$

où la fonction r est simplement assujétie à des conditions de régularité. À q fixé, l'idée consiste à introduire un estimateur non-paramétrique $\hat{r}_h$ dépendant d'un paramètre de lissage $h = (h_1, \ldots, h_q)$ et de minimiser un critère standard de validation croisée

$$CV_q = \sum_{i \in \mathcal{A}} \left(Y_i - \hat{r}_h^{-i}(X_i) \right)^2 W(X_i),$$

où $W(.)$ est une fonction de poids et où $\hat{r}_h^{-i}$ est basé sur l'échantillon $\mathcal{A}$ privé de la $i$ème observation. La minimisation s'effectue à la fois sur le paramètre de lissage h et sur toutes les combinaisons possibles de q parmi les p points de discrétisation de la variable χ. Ainsi, pour chaque valeur de la dimension q on disposera d'un q-uplet $t_{opt_q} = (t_{j_1}, \ldots, t_{j_q})$. L'étude asymptotique de cette méthode a été conduite dans [19] et montre que ce q-uplet converge vers le q-uplet fournissant la meilleure prédiction théorique de Y dès que les estimateurs non-paramétriques $\hat{r}_h$ sont les estimateurs linéaires locaux (voir [15]), mais de manière générale n'importe quel estimateur non-paramétrique de type δ-suite pourrait être utilisable.

Il est clair qu'un des inconvénients majeurs de cette approche réside dans la lourdeur de sa mise en œuvre puisqu'il convient encore de choisir q. Pour ce faire, on utilisera un algorithme séquentiel (que nous appelerons $SASDA$) de type "forward-backward" qui consiste à :

Étape 1, Initialisation : Trouver t_{opt_1} en minimisant CV_1 ;

Étape 2, Forward : Trouver t_{opt_2} en minimisant CV_2 uniquement sur tous les couples de la forme (t_{opt_1}, t_j) ;

- Étape 3, Boucle : Continuer ainsi jusqu'à stabilisation de l'erreur de prédiction sur l'échantillon test $\mathcal{T}$;
- Étape 4, Backward : Regarder, toujours à l'aide du critère CV, si certains points sélectionnés peuvent être supprimés.

Pour ce qui concerne l'exemple spectrométrique présenté ci-dessus, les résultats sont donnés dans la Table 2 : l'algorithme SASDA amène un modèle non-paramétrique sélectif de dimension réduite $q = 4$ signifiant que seuls 4 parmi les 100 points de discrétisation des courbes spectrométriques vont intervenir pour la prédiction du taux de graisse Y. Ce modèle de dimension 4 conduit à une erreur de prédiction sur l'échantillon test $\mathcal{T}$ de 1.2 (erreur dont nous pouvons d'ores et déjà juger qu'elle est relativement faible, tout du moins en comparaison avec celles obtenues précédement par modélisation linéaire). Ces résultats sont obtenus en utilisant comme estimateurs non-paramétriques $\hat{r}_h$ ceux obtenus par la méthode linéaire locale (voir [15]).

Dimension du problème	Nombre de points sélectionnés	Erreur de prédiction
$p = 100$	$q = 4$	$ERR_{SASDA} = 1.2$

TAB. 2 – Résultats du modèle sélectif non-paramétrique sur les données spectrométriques

3.4 L'algorithme SALARS

Les bons résultats de l'algorithme non-paramétrique SASDA doivent être tempérés par la lourdeur des calculs que l'aspect séquentiel de la méthode ne saurait effacer à lui seul. En effet on peut aisément imaginer des jeux de données fonctionnelles observées sur une grille beaucoup plus dense que les 100 points de discrétisation de l'exemple étudié ici. Ainsi on pourrait penser que l'on en est réduit à choisir entre les bonnes performances de la méthode SASDA et la rapidité de la méthode LARS, ou vu de manière plus négative on devrait se résoudre à accepter les mauvaises performances statistiques de l'une ou bien les difficultés d'implémentation de l'autre ...

Devant ce genre de question, plutôt que d'opposer les deux méthodes précédentes, les idées de "boosting" vont s'attacher à les combiner afin de tirer partie des avantages de chacune. C'est ainsi que l'on peut proposer un nouvel algorithme de sélection de points informatifs qui va combiner la flexibilité de la méthode SASDA et la faisabilité de la méthode LARS. Cet algorithme, que nous appellerons SALARS, consiste à :

Étape 1 : Faire une pré-sélection au moyen de l'algorithme LARS ;

Étape 2 : Utiliser l'algorithme SASDA sur les points déjà pré-sélectionnés.

Pour ce qui concerne l'exemple spectrométrique présenté ci-dessus, les résultats sont donnés dans la Table 3 : l'algorithme SALARS amène un modèle non-paramétrique sélectif de dimension réduite $q = 4$ signifiant que seuls 4 parmi les 100 points de discrétisation des courbes spectrométriques vont intervenir pour la prédiction du taux de graisse Y. Ce modèle de dimension 4 conduit à une erreur de prédiction sur l'échantillon test $\mathcal{T}$ de 1.7.

Dimension du problème	Nombre de points sélectionnés	Erreur de prédiction
$p = 100$	$q = 4$	$ERR_{SALARS} = 1.7$

TAB. 3 – Résultats du modèle sélectif boosté sur les données spectrométriques

3.5 Bilan et commentaires

Sur le problème de spectrométrie qui nous intéresse, le bilan des performances prédictives des diverses méthodes de régression basées sur des modèles sélectifs est reporté dans la Table 4.

Caractéristiques du modèle	Méthode de prédiction	Erreur de prédiction
Multivarié/Linéaire	LARS	$ERR_{LARS} = 8.4$
Multivarié/Non-paramétrique	SASDA	$ERR_{SASDA} = 1.2$
Multivarié/Boosting	SALARS	$ERR_{SALARS} = 1.7$

TAB. 4 – Bilan des modèles sélectifs sur les données spectrométriques

La première chose que l'on peut constater est la très forte non-linéarité du phénomène, puisqu'un modèle de sélection linéaire de type LARS est bien moins performant que la méthode de sélection non-paramétrique SASDA. D'un autre côté, on constate que la lourdeur de mise en œuvre de la technique non-paramétrique de sélection de variables peut-être contournée par la technique de pré-sélection linéaire. Ainsi, l'algorithme SALARS n'amène qu'une faible perte en terme de puissance prédictive tout en étant d'une implémentation très rapide.

Dans l'état actuel de développement de ces méthodes, nous conseillons dans la pratique l'utilisation de l'algorithme SALARS afin de pallier au manque de souplesse du modèle linéaire sans mettre en cause sa rapidité d'implémentation. Le probème essentiel tient au manque de méthode directe d'estimation de la famille de points $(t_{j_1}, \ldots, t_{j_q})$ qui sont informatifs dans le modèle non-paramétrique. Par voie de conséquence, notre recommandation peut bien sur être amenée à changer en fonction de l'évolution des connaissances dans ce domaine.

Avant de conclure ce Paragraphe 3 il convient de rappeler que toutes les méthodes décrites juqu'ici sont basées sur une approche multidimensionnelle du probème qui n'intègre pas la continuité structurelle des données. Dans le Paragraphe 4 qui suit, nous allons présenter des approches alternatives basées sur la modélisation fonctionnelle du problème.

4 Méthodes basées sur des modèles purement fonctionnels

4.1 Introduction aux modèles fonctionnels

De manière générale, une approche fonctionnelle du problème de régression consiste à modéliser le lien entre la réponse Y (ici la variable scalaire "taux de graisse" à prédire) et une explicative fonctionnelle χ (ici la courbe spectrométrique) :

$$Y_i = r(\chi_i) + \epsilon_i, \ i = 1, \ldots, n.$$

L'objet à estimer est l'opérateur fonctionnel r défini sur l'espace E de dimension infinie dans lequel la variable χ prend ses valeurs (ici par exemple E peut-être un espace de fonctions suffisamment régulières et à support $[850, 1050]$). Dans le Paragraphe 4.2 nous

étudierons une méthode de prédiction basée sur une modélisation linéaire de l'opérateur r, tandis que le Paragraphe 4.3 sera basé sur une modélisation non-paramétrique de cet opérateur.

4.2 Construction d'un modèle fonctionnel linéaire

Le modèle de régression linéaire fonctionnelle consiste à supposer que l'espace E est de Hilbert et que l'opérateur de régression r est linéaire et continu. Ainsi, si l'on note $< . >$ le produit scalaire sur E, le théorème de représentation des opérateurs linéaires continus permet de ramener le problème de l'estimation de r à celui d'un élément $\beta(.) \in E$, puisque le modèle de régression peut alors s'écrire :

$$Y_i =< \chi_i, \beta > +\epsilon.$$

Dans le problème précis qui nous intéresse ici, on peut prendre pour E l'espace des fonctions de carré intégrable sur $[850, 1050]$ muni de son produit scalaire usuel, et on peut écrire le modèle de la manière suivante :

$$Y_i = \int_{850}^{1050} \beta(t)\chi_i(t)dt + \epsilon.$$

Comme en témoigne la revue bibliographique récente [9], de nombreuses techniques d'estimation du paramètre fonctionnel $\beta(.)$ ont été développées dans la littérature. Nous nous limiterons ici à l'estimateur basé sur les Splines de lissage (voir [10]), qui est un des plus couramment utilisé dans la littérature, et nous noterons $FLR - Spline$ la méthode de prédiction de Y basée sur cet estimateur de β. Cet estimateur $\hat{\beta}$ est obtenu en minimisant un critère de moindres carrés pénalisés

$$\sum_{i=1}^{n} (Y_i - < X_i, \beta >)^2 \ + \ \lambda||\beta||^2.$$

La minimisation de ce critère est effectuée parmi les β s'écrivant comme combinaisons linéaires de Splines. Ici λ est un paramètre de lissage qui contrôle la régularité de la fonction $\beta(.)$ en contraignant sa norme L_2 à ne pas être trop grande ; ce paramètre est choisi par validation croisée.

L'application de cette procédure sur les données spectrométriques étudiées ici amène une erreur de prédiction

$$ERR_{FLR-Spline} = 8.01.$$

À ce stade il faut souligner que ce résultat est lié directement à la modélisation linéaire sous jacente et non à la méthode particulière d'estimation que nous avons utilisée, puisqu'en effet la mise en œuvre de tout autre estimateur du paramètre β amène des erreurs de prédiction similaires. Nous avons aussi effectué les prédictions à partir de deux méthodes linéaires fonctionnelles alternatives aux splines de lissage : la régression sur composantes principales lisses introduite par [10] (notée désormais FLR-SPC) et la régression linéaire pré-lissée introduite par [27] (notée désormais FLR-PreSmooth). La Table 5 présente tous ces résultats.

Vu la similarité des erreurs de prédiction, nous ne garderons pour la suite en mémoire que le résultat de la méthode FLR-Spline que nous noterons désormais plus simplement FLR. Nous reviendrons plus longuement dans le Paragraphe 4.4 sur les conclusions à tirer de ces résultats.

Caractéristiques du modèle	Méthode de prédiction	Erreur de prédiction
Linéaire Fonctionnel	Splines de lissage	$ERR_{FLR-Spline} = 8.0$
Linéaire Fonctionnel	Comp. Princ. Lisses	$ERR_{FLR-SPC} = 8.3$
Linéaire Fonctionnel	Pré-lissage	$ERR_{FLR-PreSmooth} = 7.9$

TAB. 5 – Bilan des estimateurs linéaires fonctionnels sur les données spectrométriques

4.3 Construction d'un modèle fonctionnel non-paramétrique

Le modèle de régression fonctionnel non-paramétrique consiste à relacher l'hypothèse de linéarité sur l'opérateur de régression r. Ainsi le modèle s'écrit sous sa forme générale :

$$Y_i = r(\chi_i) + \epsilon_i,\ i = 1, \ldots, n,$$

où r est un opérateur soumis à des contraintes de régularité (continuité par exemple) qui sont relativement faibles (en comparaison avec le modèle linéaire précédemment introduit). Le premier avantage d'une telle généralité est de ne pas nécessiter, contrairement à la modélisation linéaire précédente, de structure Hilbertienne sur l'espace E de la variable fonctionelle puisque seules des notions d'ordre topologique seront nécessaires pour modéliser la régularité de r. De manière générale, on suppose que E est un espace muni d'une semi-métrique d et que la continuité de l'opérateur r s'entend au sens de la topologie associée à d. Nous verrons par la suite au travers de l'exemple spectrométrique traité ici qu'une telle généralité n'obéit pas simplement à un souci d'esthétique mathématique mais qu'elle permet d'obtenir des informations sur les données qu'une structure moins générale (Hilbert ou Banach par exemple) ne pourrait capturer. Ces avantages en terme de flexibilité du modèle statistique se traduisent naturellement par un accroissement des difficultés d'estimation. En effet, l'objet à estimer r est un opérateur (non-linéaire) défini sur E alors que la structure de Hilbert introduite dans le modèle linéaire permettait de simplifier ce problème en le ramenant à celui d'un élément $\beta(.)$ de E. Ce type de modèle a été récemment étudié et popularisé par [21], et les considérations précédentes sur le modèle vont justifier l'appellation non-paramétrique.

Concernant les techniques d'estimation, il est naturel de chercher à adapter les méthodes usuelles d'estimation non-paramétrique multidimensionnelle au nouveau cadre de variables fonctionnelles. Comme en témoigne la revue bibliographique récente [23], les méthodes qui se sont pour l'instant prêtées le mieux à une telle adaptation sont celles basées sur les idées de pondérations locales. Là encore, la structure topologique (directement héritée de la semi-métrique d) jouera un rôle prépondérant pour définir la notion de voisinage (et donc de localité). Plus précisément de tels estimateurs de r s'écrivent, pour une nouvelle courbe x, sous la forme

$$\hat{r}(x) = \sum_{i=1}^{n} Y_i W_{n,i}(x).$$

Les poids locaux $W_{n,i}$ font intervenir les variables χ_i se trouvant dans une boule centrée en x et de rayon h. La semi-métrique d contrôle la nature de ces boules et le paramètre h contrôle leur taille. Aussi bien les résultats théoriques que les diverses applications réalisées (voir [23] pour un échantillon exhaustif de références) attestent du fait que le bon comportement de tels estimateurs dépend de deux choses essentielles : la topologie dont l'espace est muni (autrement dit, la semi-métrique d) doit être adaptée aux données

fonctionnelles (χ_i) et le paramètre de lissage h doit être autorisé à dépendre de x afin de prendre en compte des structures locales (hétérogénéité de concentration par exemple) qui sont monnaie courante dans les jeux de données de très grande dimension et donc en particulier dans les jeux de données fonctionnelles. Ici nous en resterons à la méthode d'estimation de type k plus proches voisins, étudiée dans [8], et que nous appellerons désormais $FNPR-kNN$. Les pondérations sont définies par :

$$W_{n,i}(x) = K(\frac{d(\chi_i, x)}{h_k(x)}),$$

où $K(.)$ est un noyau (une fonction continue, positive et décroissante sur son support [0,1]) et $h_k(x)$ est le rayon de la plus petite boule contenant au moins un nombre fixé k de courbes χ_i :

$$h_k(x) = inf\{h, \#\{i, d(\chi_i, x) \leq h\} = k\}.$$

Cette méthode présente l'avantage de travailler sur des voisinages dont la taille s'adapte localement de manière automatique au travers d'un unique paramètre k (entier de surcroit), facilitant ainsi sa mise en œuvre. Pour ce qui concerne la semi-métrique d, il est naturel d'en essayer plusieurs avant d'en choisir une. Concrètement, l'aspect régulier des données (voir Figure 1) amène à considérer des semi-métriques de type L_2 sur les dérivées des courbes :

$$d(x, y) = \int_{850}^{1050} \left(x^{(q)}(\lambda) - y^{(q)}(\lambda)\right)^2 d\lambda.$$

Les paramètres k et q sont choisis par validation croisée.

Ainsi, pour le problème spectrométrique présenté ci-dessus, cette méthode a abouti au choix $q = 2$, confirmant une pratique courante en chimie quantitative qui est de travailler avec les dérivées des courbes (voir Figure 2) plutôt qu'avec les courbes elles-mêmes. Nous renvoyons à l'ouvrage [21] pour de plus amples discussions sur ces questions. Finalement, en choisisant k par validation croisée, cette méthode a produit une erreur de prédiction relativement faible de

$$ERR_{FNPR-kNN} = 1.9.$$

Signalons que la plupart des autres estimateurs non-paramétriques fonctionnels ont été testés sur ces mêmes données amenant des erreurs de prédiction similaires dès que les deux principes de base sont respectés (localité de la taille du voisinage et bon choix de semi-métrique). La Table 6 présente les résultats obtenus par la méthode précédente (méthode FNPR-Kernel), ainsi que par deux méthodes fonctionnelles de type linéaire local introduites par [3] et [4] (méthodes FNPR-LL-BG et FNPR-LL-BFV) qui sont des variantes de l'estimateur à noyau construites à partir de problème de minimisation. Tous ces résultats sont obtenus en utilisant un noyau $K(.)$ quadratique et en choisissant les divers paramètres de lissage par validation croisée.

Pour insister sur le rôle de la topologie (c'est-à-dire sur l'influence de q), il faut souligner que le choix $q = 0$ (qui équivaut à modéliser E comme une espace métrique) donne des résultats très mauvais puisque l'erreur de prédiction (même avec la méthode locale kNN) est très importante (de l'ordre de 71). Ainsi, l'utilisation d'espace semi-métrique s'avère ne pas être qu'un raffinement mathématique mais un outil indispensable pour l'analyse de telles données. De même, pour insister sur la nécessité de considérer des voisinages locaux, il faut souligner que (même avec $q = 2$) une méthode à noyau basée sur une

Caractéristiques du modèle	Méthode de prédiction	Erreur de prédiction
Non-Param. Fonctionnel	Plus proches voisins	$ERR_{FNPR-kNN} = 1.9$
Non-Param. Fonctionnel	Linéaire Local I	$ERR_{FNPR-LL-BG} = 1.6$
Non-Param. Fonctionnel	Linéaire Local II	$ERR_{FNPR-LL-BFV} = 2.9$

TAB. 6 – Bilan des estimateurs non-paramétriques fonctionnels sur les données spectrométriques

fenêtre globale $h(x) = h, \forall x$ donne elle aussi de très mauvais résultats puisque l'erreur est encore très importante (de l'ordre de 5.4).

On peut conclure des résultats de la Table 6 que la qualité des prédictions est d'avantage liée à la structure non-paramétrique du modèle qu'à la méthode d'estimation particulière qui a été mise en œuvre, à condition que cette méthode respecte les deux principes de base : un bon choix de topologie et une adaptation locale de la taille des voisinages. Pour la suite nous ne garderons en mémoire que le résultat obtenu par la méthode FNPR-kNN (car outre ses bonnes qualités prédictives elle présente l'avantage d'être très peu coûteuse en temps de calculs), et nous l'appellerons désormais plus simplement méthode FNPR.

4.4 Bilan et premiers commentaires

Le bilan des performances prédictives des diverses méthodes basées sur des modèles purement fonctionnels est résumé dans la Table 7.

Caractéristiques du modèle	Méthode de prédiction	Erreur de prédiction
Fonctionnel/Linéaire	Spline de lissage	$ERR_{FLR} = 8.0$
Fonctionnel/Non-paramétrique	k proches voisins	$ERR_{FNPR} = 1.9$

TAB. 7 – Bilan des modèles fonctionnels sur les données spectrométriques

La première conclusion à tirer est la très forte non-linéarité du phénomène liant le taux de graisse à la courbe spectrométrique. Ces aspects non-linéaires sont bien pris en compte par la méthode $FNPR$ qui est basée sur une modélisation non-paramétrique de l'opérateur de régression r, tandis qu'une modélisation linéaire standard (même fonctionnelle) de type FLR est incapable de les restituer.

5 Complémentarités des méthodes et boosting

5.1 Discussion générale

Les résultats présentés dans les Tables 4 et 7 attestent de la non-linéarité du phéomène puisqu'une approche non-paramétrique donne toujours de bien meilleurs résultats qu'une approche linéaire, et ce quelle que soit la manière d'appréhender les données (purement fonctionnelle ou multivariée). Bien entendu, cet aspect non-linéaire est propre à ce

problème de spectrométrie et ne saurait être pris comme conclusion générale. De manière plus précise, on peut constater que les deux types de modélisation amènent des erreurs similaires (avec un léger avantage toutefois pour la modélisation multivariée). Une conclusion hâtive pourrait amener à penser que le choix entre ces deux types d'approche est, finalement, de peu d'importance. Mais il est cependant légitime de s'interroger sur le fait de savoir si les deux approches capturent ou non les mêmes informations sur les données. Autrement dit, peut-on espérer qu'une combinaison de ces deux approches puisse amener des gains intéressants en terme d'erreur de prédiction ?

5.2 Une combinaison des avantages multivariés et fonctionnels

Afin de tirer profit des avantages respectifs de chaque modélisation, nous allons proposer une technique de boosting basée sur l'utilisation successive de la méthode multivariée sélective SALARS et de la méthode fonctionnelle FNPR. Concrètement, ce nouvel algorithme que nous appellerons MULT-FONC est construit selon les étapes suivantes :

Étape 1 : Appliquer l'algorithme sélectif SALARS pour le problème de régression initial $Y = r(\chi) + \epsilon$; Calculer les prédictions $\hat{Y}_i, i \in \mathcal{A} \cup \mathcal{T}$ obtenues par cette méthode ; Calculer les résidus $\hat{\epsilon}_i = \hat{Y}_i - Y_i, i \in \mathcal{A}$;

- Étape 2 : Appliquer la méthode fonctionnelle FNPR pour le nouveau problème de régression $\hat{\epsilon} = g(\chi) + \epsilon'$; Calculer les prédictions $\hat{\hat{\epsilon}}_i, i \in \mathcal{T}$ obtenues par cette méthode ;
- Étape 3 : Calculer les nouvelles prédictions boostées $\hat{Y}_i + \hat{\hat{\epsilon}}_i,\ i \in \mathcal{T}$.

Cet algorithme a été mis en œuvre sur les données précédentes. Pour mieux percevoir la portée du résultat, celui-ci est rapporté dans la Table 8 en même temps que sont rappelés ceux obtenus par la méthode purement fonctionnelle et par la modélisation multivariée sélective.

Caractéristiques du modèle	Méthode de prédiction	Erreur de prédiction
Multivarié sélectif	SALARS	$ERR_{SALARS} = 1.7$
Fonctionnel	k proches voisins	$ERR_{FNPR} = 1.9$
Boosting	SALARS + FNPR	$ERR_{MULT-FONC} = 0.7$

TAB. 8 – Résultat du boosting sur les données spectrométriques

5.3 Commentaires

Il est incontestable que l'approche de type boosting apporte un gain important en terme de qualité de prédiction. Bien que déjà relativement performantes, les deux modélisations (multivariée et fonctionnelle) ne capturent pas les mêmes effets et leur combinaison permet de diminuer considérablement les erreurs. Les modèles basés sur les techniques de sélection de variables permettent de choisir les points de discrétisation (en petit nombre) ayant un pouvoir prédictif optimal. En parallèle, les modèles purement fonctionnels prennent en compte l'aspect continu de chaque variable fonctionnelle et arrivent à récupérer d'autres informations (qui sont de nature différente de celles captées par les méthodes multivariées).

Les idées de boosting permettent de tirer profit des atouts respectifs des deux approches, donnant ainsi des résultats très largement supérieurs à ceux que chaque méthode peut donner séparément.

6 Conclusion générale

Les conclusions à tirer de cette rapide étude comparative de divers modèles/méthodes de régression pour variables fonctionnelles tiennent en deux idées essentelles.

La première idée est de dire que, de toute évidence, un jeu de données fonctionnelles est un échantillon statistique de très grande dimension. Par voie de conséquence, les méthodes récentes développées dans d'autres contextes (génomique par exemple) pour appréhender ces problèmes de grandes dimensions peuvent aussi s'avérer intéressantes en analyse de données fonctionnelles. Dans le même ordre d'idées, réciproquement, certaines avancées récentes en statistique fonctionnelle concernant la sélection de points de discrétisation informatifs peuvent aussi s'avérer fructueuses dans d'autres contextes non-fonctionnels mais de grande dimension. En particulier, la méthode SALARS décrite précédement pourrait aussi apporter des éléments nouveaux dans des problèmes (non-fonctionnels) de sélection de variables pour lesquels les hypothèses de linéarité s'avèreraient être inadéquates.

Le deuxième point important à retenir est le fait qu'un jeu de données fonctionnelles n'est pas qu'un simple échantillon statistique de grande dimension puisqu'il présente la particularité (due à la continuité des phénomènes observés) d'une très forte colinéarité entre les variables. Ainsi, les modèles relevant purement de la statistique multivariée sont incapables de prendre en compte cet aspect de continuité ; cependant, ils peuvent être avantageusement combinés avec d'autres modèles de type fonctionnel.

En fin de compte, l'extrême complexité structurelle d'un jeu de données fonctionnelles ouvre la porte à de très nombreux problèmes et méthodes statistiques. L'exemple étudié ici illustre bien la nécessité d'utiliser ces diverses méthodes dans un esprit de complémentarité plutôt que de compétitivité. Ainsi, bien qu'elles ne soient encore que très peu développées (à notre conaissance les travaux dans ce domaine se limitent à [26], [16], [24], [32] et [22]), les techniques de boosting sont promises à un avenir florissant en analyse de données fonctionnelles. Et la grande dimension apparaitra enfin, et plus que jamais, comme une chance (et non pas comme un fléau ...).

Appendice

Les données utilisées dans ce travail servent depuis quelques années de repère sur lequel toute nouvelle méthode statistique est testée. L'objectif de cet appendice est de recenser tous les résultats donnés par ces diverses méthodes. Certains travaux (comme par exemple ceux de [2]) sont volontairement exclus de cette liste parce qu'ils utilisent des variables explicatives additionnelles.

Tous ces résultats sont donnés en terme d'erreur de prédiction $ERR_{\mathcal{M}}$ (telle que définie dans le Paragraphe 2.5) mesurée sur un échantillon test $\mathcal{T}$; chaque méthode statistique $\mathcal{M}$ étant construite sur un échantillon d'apprentissage $\mathcal{A}$ et sans utiliser en aucune manière l'échantillon $\mathcal{T}$. Les résultats annoncés dans la Table 9 sont obtenus à partir des données fournies dans la rubrique *spectrometric dataset* du site *http ://www.math.univ-toulouse.fr/staph/npfda/npfda-datasets.html* et en prenant pour échantillon $\mathcal{A}$ (respecti-

vement pour échantillon $\mathcal{T}$) les 160 premières lignes de ce tableau (respectivement les 55 dernières). Pour tout ce qui concerne les modèles non-paramétriques fonctionnels la fonction noyau $K(.)$ est le noyau quadratique. Par ailleurs, les différents paramètres de lissage intervenant dans les méthodes ont toujours été choisis par validation croisée. Nous souhaiterions que cette Table 9 soit incrémentée, au fur et à mesure que de nouvelles techniques de prédiction verront le jour, en suivant le même schéma de construction pour les échantillons $\mathcal{A}$ et $\mathcal{T}$.

Outre les abbréviations déjà introduites précédemment dans cet article, nous utiliserons dans la Table 9 les suivantes :

(a) Codes R accessibles à partir du package *lars* sur le site *http ://cran.r-project.org/* ;
(b) Fonction R/S *pgsplus.spl* accessibles sur le site *http ://www.math.univ-toulouse.fr/staph/* dans la rubrique *logiciels en ligne* ;
(b) Codes R/S et manuel d'utilisation accessibles sur le site *http ://www.math.univ-toulouse.fr/staph/npfda/* ;
(d) Résultats obtenus à partir des courbes brutes ;
(e) Résutats obtenus à partir des dérivées secondes des courbes ;
(f) Ce modèle utilise toutes les déerivées successives des courbes ;
(g) Méthodes non basées sur l'estimation de l'espérance conditionnelle ;
(h) Étude menée par Adela Martinez (St J. de Compostelle, Espagne) ;
(i) Étude menée par Aldo Goia (Novara, Italie) ;
NP Modèle Non-paramétrique ;
SP Modèle semi-paramétrique ;
Lin Linéaire ;
Mult Modèle multivarié sélectif ;
NFG Estimateur à noyau construit (voir Paragraphe 4.3) avec une largeur de fenêtre globale fixe $h(x) = h, \forall x$.

Références

[1] A. Ait-Saidi, F. Ferraty, R. Kassa and P. Vieu — *Cross-validated estimation in the single functional index model.* Statistics, **42**, 475-494 (2008).

[2] G. Aneiros Perez and P. Vieu — *Semi-functional partial linear regression.* Statist. & Prob. Letters, **76**, 1102-10 (2006).

[3] A. Baillo and A. Grané — *Functional local linear regression with functional predictor and scalar response.* J. Multiv. Anal., **100**, 102-111 (2009).

[4] J. Barrientos-Marin, F. Ferraty and P. Vieu — *Locally modelled egression and functional data.* J. Nonparametric Statist., En cours d'impression (2009).

[5] P. Besse and J.O. Ramsay — *Principal components analysis of sampled functions.* Psychometrika, **51**, 285-311 (1986).

[6] C. Borggaard and H.H. Thodberg — *Optimal minimal neural interpretation of spectra.* Analytical chemistry, **64**(5), 545-551 (1992).

Méthodologie	Modèle	Méthode	Référence	Notes	Erreur
Multivariée	NP	SASDA	[19]		1.2
	Linéaire	LARS	[14]	(a)	8.4
	Boosting	SALARS	Paragraphe 3.4		1.7
Fonctionnelle	Linéaire	Splines	[10]	(b,h)	8.8
	...	ACP lisse	[10]	(b,h)	8.9
	...	Pré-lissage	[27]	(h)	8.5
Fonctionnelle	NP sur χ_i	NFG	[21]	(c,d)	148.2
	...	kNN	[8]	(c,d)	70.7
	NP sur χ_i''	NFG	[21]	(c,e)	5.4
	...	kNN	[8]	(c,e)	1.9
	...	Lin Loc I	[3]	(e)	1.6
	...	Lin Loc II	[4]	(e)	2.9
	...	Mediane Cond	[21]	(c,e,g)	4.8
		Mode Cond	[21]	(c,e,g)	2.9
Fonctionnelle	Additif	Backfitting	[22]	(f)	1.2
	SP	Single Index	[1]	(i)	3.3
	SP	Proj Pursuit	[18]	(i)	1.9
Boosting	Mult+NP	SALARS+kNN	Paragraphe 5		0.7

TAB. 9 – Récapitulatif des méthodes sur les données spectrométriques.

[7] D. BOSQ. — *Linear processes in functional spaces. Theory and Applications.* Lecture Notes in Statistics, **149**, Springer Verlag, new York (2000).

[8] F. BURBA, F. FERRATY AND P. VIEU. — *k-nearest neighbor method in functional nonparametric regression.* J. Nonparametric Statist., **21**, 453-469 (2009).

[9] H. CARDOT AND P. SARDA. — *Functional Linear Regression.* In *Handbook on Functional Data Analysis & related fields (Ed. F. Ferraty and Y. Romain).* Oxford University Press (2010).

[10] H. CARDOT, F. FERRATY AND P. SARDA. — *Spline estimators for the functional linear model.* Statistica Sinica, **13**, 571-591.

[11] J. DAUXOIS, A. POUSSE AND Y. ROMAIN. — *Asymptotic theory for the principal component analysis of a random vector function : some application to statistical inference.* Journal of Multiv. Anal., **12**, 136-154 (1982).

[12] M. DAVIDIAN, X. LIN, AND J.L. WANG — *Introduction to the Emerging Issues in Longitudinal and Functional Data Analysis.* Statist. Sinica, **14**(3), 613-629 (2004).

[13] J. DEVILLE. — *Méthodes statistiques et numériques de l'analyse harmonique.* Annales de l'INSEE, **15**, 3-97 (1974).

[14] B. EFRON, T. HASTIE, T. JOHNSTONE AND R. TIBSHIRANI. — *Least angle regression* Ann. Statist., **32**, 407-499 (2004).

[15] J. FAN AND I. GIJBELS. — *Local polynomial modelling and its applications.* Chapman and Hall, London (1996).

[16] B. Fernández de Castro and W. González Manteiga. — *Boosting for real and functional samples : an application to environmental problem.* Stoch. Environ. Res. Risk Assess., **22**(1), 27-37 (2008).

[17] F. Ferraty. — *Special issue on statistical methods and problems in infinite dimensional spaces.* J. Multivariate Analysis, **101**(2), 305-490 (2010).

[18] F. Ferraty, A. Goia, E. Salinelli and P. Vieu. — *Additive functional regression model based on projection pursuit directions.* En cours de rédaction (2010).

[19] F. Ferraty, P. Hall and P. Vieu. — *Most predictive design points for functional predictors.* En révision, (2010).

[20] F. Ferraty and Y. Romain. — *Handbook on Functional Data Analysis & related fields.* Oxford University Press (2010).

[21] F. Ferraty and P. Vieu. — *Nonparametric Functional Data Analysis.* Springer Series in Statistics, New York (2006).

[22] F. Ferraty and P. Vieu. — *Additive regression and boosting for functional data.* Computational Statist. and Data Analysis, **53**, 1400-13 (2009).

[23] F. Ferraty and P. Vieu. — *Kernel regression estimation for functional data.* In *Handbook on Functional Data Analysis & related fields (Ed. F. Ferraty and Y. Romain).* Oxford University Press (2010).

[24] J. Gertheiss and G. Tutz. — *Supervised feature selection in mass spectrometric based proteomic profiling by blockwise boosting.* Bioinformatics, **25**, 1076-77 (2009).

[25] W. González Manteiga and P. Vieu. — *Introduction to the Special Issue on Statistics for Functional Data.* Computational Statist. and Data Analysis, **51**(10), 4788-92 (2007).

[26] N. Krämer. — *Boosting functional data.* In *COMPSTAT2006. proceedings in Computational Statistics*, Heidelberg Physica, 1121-28 (2006).

[27] A. Martinez Calvo. — *Presmoothing in functional linear regression.* In : Functional and Operatorial Statistics. Physica-Verlag, Heidelberg, 223-229 (2008).

[28] J. Ramsay and B. Silverman. — *Functional Data Analysis.* Springer Series in Statistics, New York (1997).

[29] J. Ramsay and B. Silverman. — *Applied functional Data Analysis.* Springer Series in Statistics, New York (2002).

[30] J. Ramsay and B. Silverman. — *Functional Data Analysis (2nd Edition).* Springer Series in Statistics, New York (2005).

[31] R. Tibshirani — *Regression shrinkage and selection via the lasso.* J. Roy. Statist. Soc. Ser. B, **58**, 267-288 (1996).

[32] G. and J. Gertheiss. — *Feature extraction in signal regression : a boosting technique for functional data regression.* Preprint (2009).

[33] M. Valderrama, — *Introduction to the Special Issue on Modelling Functional Data in Practice.* Computational Statist., **22**(3), 331-334 (2007).