Regression on functional data: methodological approach with application to near-infrared spectrometry

Frédéric Ferraty

Regression on functional data: methodological approach with application to near-infrared spectrometry

Frédéric Ferraty (ferraty@math.univ-toulouse.fr)
University of Toulouse & Toulouse Mathematics Institute, France.

Abstract: We consider the situation when one observes a scalar response and a functional variable as predictor. For instance, in our petroleum industry problem, the response is the octane number of a gasoline sample and the functional predictor is a curve representing its near-infrared spectrum. The statistician community developed numerous models for handling such datasets and we focus here on four regression models: two standards as the functional linear model and the functional nonparametric regression, and two recently developed: the functional projection pursuit regression and a parsimonious model involving a nonparametric variable selection method. Each of these models are implemented with two datasets containing near-infrared spectrometric curves. A comparative study of these models is provided in order to emphasize their possible advantages and drawbacks. At last, a simple but useful methodological approach is then proposed in order to boost the two most recent regression models by combining the most relevant informations obtained by each of the studied models. We show on the spectrometric data how such an approach may lead to important improvements.

Keywords: boosting, functional data, functional linear regression, functional nonparametric regression, functional projection pursuit regression, nonparametric variable selection, near-infrared spectrometry

1 Introduction

Popularized by [36] and [35], functional data analysis (FDA) is a very active research area in the international statistical community. The development of this topic is essentially due to the joint progress of monitoring devices and computational tools allowing to collect and process highly dimensioned data. This kind of datasets come from the observation of some underlying continuous processus sampled at a grid of measurements. Near-infrared spectrometry provide benchmark examples coming from chemometrics. This is a non-destructive technology able to measure numerous chemical compounds in a wide variety of products (food industry, petroleum industry, wood industry, etc); see among others [31], [26]. For instance, let us consider a sample of 60 gasoline samples. Each sample is illuminated by a light beam at 401 equally spaced wavelengths $(\lambda_1, \ldots, \lambda_{401})$ in the near-infrared range 900-1700 nm. For each wavelength λ and each gasoline sample i, the absorption $X_i(\lambda)$ of radiation is measured. The ith discretized spectrometric curve is given by $X_i(\lambda_1), \ldots, X_i(\lambda_{401})$; Figure 1 displays the 60 spectrometric curves (the last 21 wavelengths were dropped to

make graphics readable). It is clear that all these curves involve some continuum

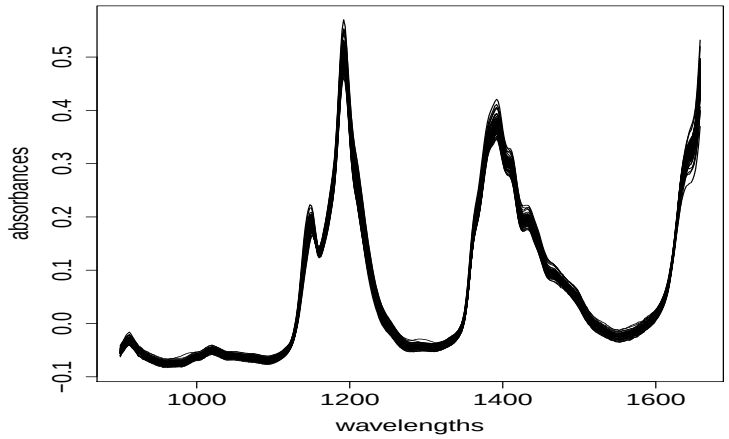


Figure 1: 60 near-infrared spectra sampled at 401 equally spaced wavelengths

in their structure even if they are observed at discrete points. The terminology *functional data* refer to this continuous feature. Figure 2 gives a benchmark example of such data introduced in [3] and dealing with 215 finely chopped pieces of pork meat. For the ith piece of meat one observes a spectrum of absorption $X_i(\cdot)$ sampled at 100 equally spaced wavelengths $\lambda_1, \ldots, \lambda_{100}$ from 850 to 1050 nm. The ith discretized spectrometric curve is given by $X_i(\lambda_1), \ldots, X_i(\lambda_{100})$. Throughout both these examples which will be our connecting thread, one can

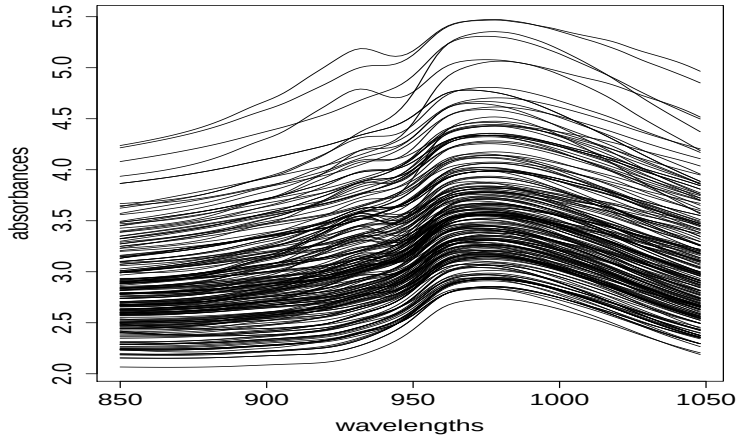


Figure 2: 215 spectrometric curves sampled at 100 equally spaced wavelengths

remark that the grid of measurements (i.e. wavelengths) for the spectrometric

curves is quite dense. It is worth noting here that there exist more pathological situations where one has at hand sparse measurements (i.e. each profile is observed at few points possibly randomly distributed) which requires particular methods. Although this is an important issue as demonstrated by the numerous publications around this topic (see for instance [30, 44, 43, 32, 42, 41]), it is out of our purpose.

The main aim of this paper is to present various ways of modelling nonlinear relationship in datasets containing functional data and to discuss methodological aspects. We focus on the special case when one regresses a scalar response on an explanatory functional variable. As in the standard multivariate setting, different families of regression models have been developed. We especially concentrate our attention on four regression models of various dimensionalities. Some of them are quite standard as the functional linear model (see [35] or [36] and references therein) or the functional nonparametric regression (see the monography [22] and [20] for general overview and related methods). Some other have been recently developed; this is the case of the functional projection pursuit (see [10], [16] and [19]) or the parsimonious nonparametric regression models involving nonparametric variable selection (see [18] and [17]). Contrarly to the functional linear model, the three other models are able to catch nonlinear relationship. Given these models and our two spectrometric examples, we propose a comparative study in order to emphasize their possible advantages or drawbacks. It is out of question to present in detail each of these statistical models as well as their theoretical properties. The reader will find useful references throughout this work which is voluntarily oriented toward practical and methodological aspects. We first describe the prediction problems in Section 2. The nonparametric functional regression, which is a model of high dimensionality, is presented in Section 3. In the opposite, Section 4 focuses on a model of low dimensionality: the functional linear regression. Section 5 is devoted to two recent regression models of intermediate dimensionality. The first is based on projection pursuit regression ideas whereas the second is a parsimonious model involving a nonparametric variable selection method. Section 6 proposes to boost the previous methods by taking into account the most relevant informations derived from each of them. We show on the spectrometric data how the combination of these models may lead to important improvements. Before concluding, Section 7 enumerates useful resources dealing with FDA, oriented towards practitioners and available online.

2 Functional data and prediction problems

Petroleum prediction problem. For each of 60 gasoline samples one collects a spectrometric curves (see Figure 1). Additionally, one knows the octane number for each gasoline sample. The goal is to predict the octane number from the observation of a new spectrometric curve. To this end, one observes n pairs $(X_i, Y_i)_{i=1,\ldots,n}$ where X_i (resp. Y_i) is the ith spectrometric curve (resp. response). The prediction problem is very simple and can be formulated via the

model $Y_i = r(X_i) + \epsilon_i$ for $i = 1, \ldots, n$. Figure 3 schematizes the problematic. The two first spectrometric curves are very similar and the responses also (the

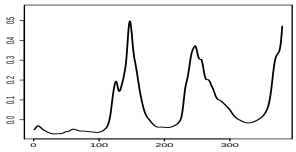
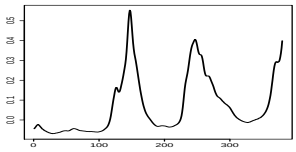
COVARIATE	RESPONSE
X_1 (near-infrared spectrum #1) 	Y_1 = octane number of gasoline sample #1 = 85.3
X_2 (near-infrared spectrum #2) 	Y_2 = octane number of gasoline sample #2 =85.25
$\vdots$	$\vdots$

Figure 3: Petroleum prediction problem

standard deviation is around 1.53).

Food prediction problem. For each of 215 pieces of meat one collects a spectrometric curves (see Figure 2). Separately, and for each piece of meat, one measures as response the fat content by means of analytic chemical process. The goal is to predict the fat content from the observation of a new spectrometric curve. Once again, one has at hand n pairs $(X_i, Y_i)_{i=1,\ldots,n}$ where X_i (resp. Y_i) is the ith spectrometric curve (resp. response). The prediction problem is the same as previously (i.e. $Y_i = r_2(X_i) + \epsilon_i$ for $i = 1, \ldots, n$) and Figure 4 schematizes the analogous problematic. The shape of both spectrometric curves are quite similar excepted the occurence of a small secondary bumb in the second spectrum; the responses are quite different (the standard deviation is around 12.74).

Finally, in these situations the statistical model admits the general writting $Y = r(X) + \epsilon$ where $r(\cdot)$ is an unknown operator modelling the relationship between X and Y; the statistical challenge consists in proposing a relevant estimator. Here, we focus our attention on regression models such that

$$\begin{array}{l} r(X) = E(Y|X) \text{ (i.e. } E(\epsilon|X) = 0) \text{ with the constraint that } r \text{ be-} \\ \text{longs to some set of } \mathcal{C}\text{; no additional assumption on the distribution} \\ \text{of } (X, Y) \text{ is required.} \end{array} \qquad (\mathcal{M})$$

This general modelling covers numerous situations and the set $\mathcal{C}$ concentrates all

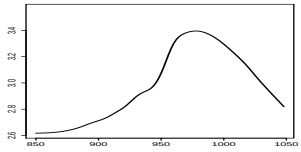
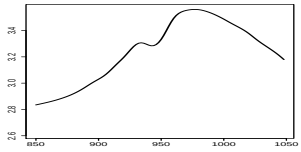
Covariate	Response
X_1 (near-infrared spectrum #1) 	Y_1 = fat content of piece #1 = 22.5
X_2 (near-infrared spectrum #2) 	Y_2 = fat content of piece #2 = 40.1
$\vdots$	$\vdots$

Figure 4: Chemometrics prediction problem

the model hypotheses. The nature and size of $\mathcal{C}$, which determines what people call "the dimensionality of the model", play a major statistical role. However, a general statistical principle says that the accuracy of the estimating procedure depends on the size of $\mathcal{C}$; larger is the size of $\mathcal{C}$, harder is the estimation of the unknown operator r. So, we propose in the remaining of this paper to discuss these aspects, mainly from a methodological and practical point of views.

3 Models of high dimensionality: pure nonparametric regression

Numerous references, theoretical developments and practical studies can be found in [22] which popularized nonparametric methodologies in the functional data field. From a mathematical point of view, functional data are defined as observations of some random variable X taking its values in some infinite-dimensioned space $\mathcal{F}$ endowed with the inner product $\langle \cdot, \cdot \rangle$. The datasets introduced previously are typical examples where the considered infinite-dimensioned spaces are just functions spaces. In this section, we especially focus on nonparametric regression when one considers a functional explanatory variable X and a scalar response Y; one observes n pairs (X_i, Y_i) identically distributed as (X, Y).

3.1 Nonparametric regression model

Considering a nonparametrically regression model amounts to refer to $(\mathcal{M})$ with the set of constraints $\mathcal{C}_{FNPR}$ containing only regularity constraints acting on r.

For instance, $\mathcal{C}_{FNPR}$ may be defined as the set of operators mapping $\mathcal{F}$ into $\mathbb{R}$ such that r is lipschitz:

$$\exists \nu > 0,\, \exists C > 0,\, \forall (x_1, x_2) \in \mathcal{F} \times \mathcal{F},\, |r(x_1) - r(x_2)| \leq C\, d(x_1, x_2)^{\nu},$$

where $d(\cdot, \cdot)$ is a proximity measure (i.e. a metric or more generally a semi-metric[1]) between two elements of $\mathcal{F}$. Of course, one can relax this set of constraints by considering only continuous operator. In the opposite, there are numerous ways of enriching $\mathcal{C}_{FNPR}$; for instance one can assume that for any $(x, u) \in \mathcal{F} \times \mathcal{F}$, it exists r_x such that:

$$r(x + \delta u) = r(x) + \delta \langle r_x, u \rangle + o(\delta), \tag{1}$$

as δ tends to zero; r_x is what we call the directional derivative of r at x along the direction u. Whatever the set $\mathcal{C}_{FNPR}$, it contains only regularity assumptions. Consequently, $\mathcal{C}_{FNPR}$ contains nonlinear operators and thus the nonparametric model is very flexible. This nonparametric feature is a key advantage, especially when there is no standard tool for displaying graphically the relationship between a scalar response and an explanatory functional variable. So, the difficulty of anticipating on the shape of the regression operator combined with no a priori information makes the nonparametric modelling a relevant method for exploring such a relationship.

3.2 Functional nonparametric regression in action

Before going on, let us remind how building an estimator of r. To this end, we focus on the kernel estimator which is a very popular way of estimating nonparametrically the regression operator. Let K be a positive asymmetric kernel function; then, the nonparametric kernel estimator r_{FNPR} of r is defined as:

$$r_{FNPR}(u) = \frac{\sum_{i=1}^{n} Y_i\, K\left\{h^{-1} d(X_i, u)\right\}}{\sum_{i=1}^{n} Y_i\, K\left\{h^{-1} d(X_i, u)\right\}},$$

where h is the so-called bandwidth which plays the role of the smoothing parameter. The simplicity of its writting as well as its ease of implemention makes the kernel estimator very useful.

In order to assess the predictive performance of the functional nonparametric regression, the original dataset $\{(X_i, Y_i);\, i = 1, \ldots, n\}$ are split into two subsamples: $\mathcal{L} = \{(X_i, Y_i);\, i \in L\}$ and $\mathcal{T} = \{(X_i, Y_i);\, i \in T\}$ with $L \cup T = \{1, \ldots, n\}$ and $L \cap T = \emptyset$. The learning sample $\mathcal{L}$ allows to build the estimator r_{FNPR} and to select automatically the bandwidth h via a cross-validation procedure. The testing sample provides the relative mean squared error of prediction:

$$RMSEP(r_{FNPR}) = \frac{\sum_{i \in \mathcal{T}} (Y_i - \widehat{Y}_i)^2}{\sum_{i \in \mathcal{T}} (Y_i - \overline{Y})^2},$$

[1] a semi-metric $d(\cdot, \cdot)$ is a metric such that $d(x_1, x_2) = 0$ does not imply that $x_1 = x_2$.

where, for all i in $\mathcal{T}$, $\widehat{Y}_i = r_{FNPR}(X_i)$ and $\overline{Y}$ is the average of Y based on the testing sample. The original dataset is randomly split 100 times which allows to compute 100 values for $RMSEP(r_{FNPR})$ and to display their distribution by means of a boxplot. At last, for all datasets, the size of the testing sample $\mathcal{T}$ represents 50% of the whole sample.

We focus here on the functional regression model $Y_i = r(X_i) + \epsilon_i$ where the X_i's are near-infrared spectrometric curve of the Y_i's are corresponding scalar responses (octane number or fat content). Based on the knowledge in the chemometrician community, near-infrared spectra suffers from a calibration problem due to the electronic devices. Considering successive derivatives of the spectrometric curves instead of the original curves themselves allows to overcome this problem. Unfortunately, we never know which derivative is the most informative. [21] studied the food dataset and pointed out that the twice differentiated curves are the most predicitve. This is why the kernel estimator r_{FNPR} is used with the $d(X_i, x) = \int \{X_i''(t) - x''(t)\}^2 dt$ where the notation f'' stands for the second derivative of any real-valued univariate function f. The top panel of Figure 5(a) displays the distribution of the relative mean squared error of prediction; the values are concentrated around 0.02 (the median). The bottom panel gives an idea on the accuracy of the predictions corresponding to one run (i.e. one pair of subsamples $(\mathcal{L}; \mathcal{T})$) where $RMSEP(r_{FNPR}) \simeq 0.02$. The predictions and the observations are very close and the functional nonparametric regression seems to be a relevant tool for predicting fat content. Concerning the octane dataset, Figure 5(b) proposes similar plots than those given in Figure 5(a) but now the kernel estimator involves the proximity measure $d(X_i, x) = \int \{X_i'(t) - x'(t)\}^2 dt$ where the notation f' stands for the first derivative of any real-valued univariate function f. The median of $RMSEP(r_{FNPR})$ is around 0.2. One can observe that the global predictive power is weaker than previously but the functional nonparametric regression still works well for predicting octane number.

3.3 Methodological aspects

It is worth noting that the functional nonparametric regression method involves some proximity measure $d(\cdot, \cdot)$ (i.e. the set of constraints depends on d: $\mathcal{C}_{FNPR,d}$). For instance, chemometricians know that spectrometer induces some calibration problem (see [29] for a general overview about this statistical problem). According to the nature (fat content, octane, or any other products like moisture, sucrose level, etc) of what we intend to predict from spectrometric curves, the most informative proximity measure $d(\cdot, \cdot)$ may change but we never know in advance which is the most relevant one. This is why from a methodological point of view, one can implement the functional nonparametric regression by pluging a family of proximity measures and select the most predictive one. This amounts to consider for the regression operator r a more general set of constraints $\mathcal{C}_{FNPR} = \cup_{d \in \mathcal{D}} \mathcal{C}_d$ where $\mathcal{D}$ stands for a family of proximity measures. This procedure is particularly interesting because in some cases it allows to improve significantly the predictive ability of the nonparametric regression

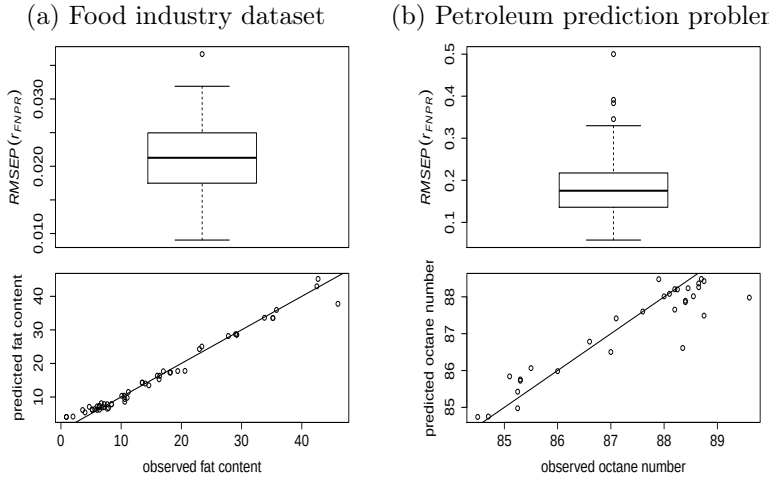


Figure 5: FNPR with original spectra: Out-of-sample predictive performance.

model. For instance, when one focuses on our spectrometric datasets, one can set $\mathcal{D} = \{d_k; k = 0, 1, \ldots, K\}$ with $d_k(u_1, u_2) := \int \{u_1^{(k)}(t) - u_2^{(k)}(t)\}^2 dt$ and where $f^{(k)}$ stands for the kth derivative of any function f (with the convention $f^{(0)} := f$). Here the family $\mathcal{D}$ is just indexed by the K first integers. Figure 6 gives an idea on how much one can expect to improve the predicitive power. Clearly, the twice (resp. once) differentiated spectrometric curves for predicting

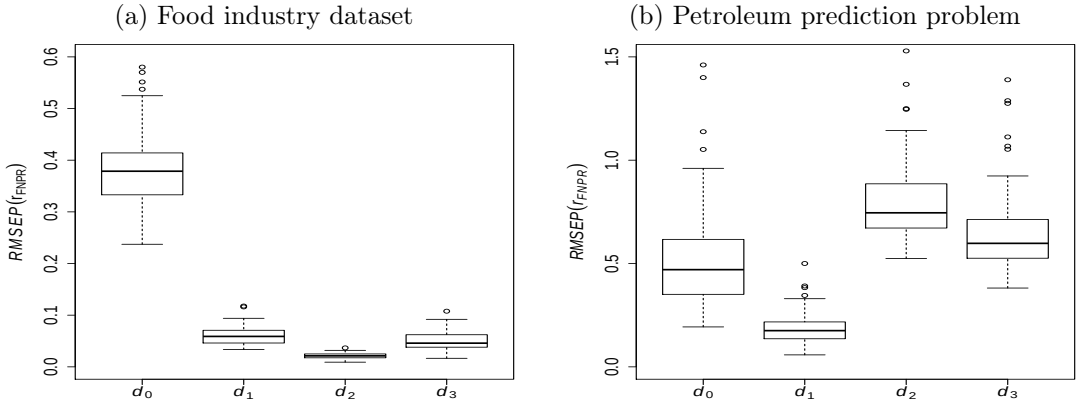


Figure 6: Spectrometry datasets: out-of-sample performance obtained with proximity measure based on successive derivatives

the fat content (resp. octane number) lead to the best out-of-sample performance and in any case the standard L_2-norm (i.e. d_0) degrades dramatically the results.

So, the proximity measure plays a major role from a practical point of view but not only. Indeed, it is important to emphasize that the proximity measure plays also a crucial role in the asymptotic behaviour of the kernel estimator (see for instance the discussion in Chapter 13 of [22]). To conclude this section, one can remark that flexibility of the functional nonparametric regression model comes from the huge size of the set of constraints $\mathcal{C}_{FNPR}$; larger is $\mathcal{C}_{FNPR}$, more flexible is the regression model. Equivalently to this notion of flexibility, statisticians introduced the terminology *dimensionality* which is a similar way to express the amount of flexibility of any model. When one says that a model is of high dimensionality, it expresses its high flexibility feature which is especially the case of the functional nonparametric regression model. However, considering models of high dimensionality may lead to several main drawbacks. Firstly, functional nonparametric regression model is not designed to produce graphical outputs allowing to interpret results. Secondly, considering model of high dimensionality produces lower rate of convergence than in the parametric framework (but it is normal because one considers model of higher dimensionality). Thirdly we have to face with the so-called *curse of dimensionality* meaning that higher is the dimension of the explanatory variable, larger should be the sample size to expect accurate predictions. Although the two first drawbacks are unescapable, the third one may be overcome. Indeed, amplified in the functional setting where explanatory variables live in some infinite-dimensioned space, the curse of dimensinality well known in the nonparametrician community is valid as soon as one considers a standard norm as proximity measure. But, if you do not reduce the proximity measures to standard norms (for instance semi-metric) and plug them as an additional "parameter" of the method, it is possible to weaken the curse of dimensionality impact. Figure 6 supports clearly this idea; with the same sample size, the use of semi-metrics significantly improves the predictive performance.

Before ending this section, let us remark that the functional nonparametric regression model is an interesting exploratory tool in that sense it points out the major predictive role played by the second (resp. first) derivative in the food (resp. petroleum) industry example.

4 Models of low dimensionality: the functional linear regression

4.1 Functinal linear regression model

Another way of modelling the relationship between a functional explanatory variable and a scalar response is to consider a set of constraint $\mathcal{C}$ much more rigid in the sense that one imposes the linearity of the regression operator r. This linearity assumption reduces considerably the dimensionalty of the regression model and the set of constraints becomes:

$$\mathcal{C}_{FLR} = \left\{ r : \mathcal{F} \to \mathbb{R}, \forall x \in \mathcal{F},\, r(x) = \langle\, x, \rho \rangle \right\},$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{F}$ and ρ is some unknown smooth functional parameter. This linear modelling is very popular in the functional data analysis community and numerous papers are available in the literature (see for instance the technical works [8], [6], [9] as well as [36] and [20] for general overviews and related methods).

4.2 Functinal linear regression model in action

Starting from the linear model $Y_i = \mu + \langle X_i, \rho \rangle + \epsilon_i$ for $i = 1, \ldots, n$, a standard estimating procedure consists in minimising some penalized sum of squares of the form $Q_\lambda(\rho) := \sum_{i=1}^{n} (Y_i - \overline{Y} - \langle X_i, \rho \rangle)^2 + P_\lambda(\rho)$ where $P_\lambda(\rho)$ is a penalty term depending of some (possibly multivariate) parameter λ and set $\widehat{\rho} := \inf_{\rho \in \mathcal{S}} Q_\lambda(\rho)$. Here, $\overline{Y}$ is the average of the Y_i's and stands for the estimation of the real parameter μ. The last minimization operates over a set $\mathcal{S}$ of smooth functions with good approximation properties and depending on the context (such as spline basis, wavelet basis, tensor product of splines, etc). In our situation, we only use spline estimator for deriving the unknonwn functional parameter ρ. Then, for all $x \in \mathcal{F}$, the estimator r_{FLR} of r is defined as

$$r_{FLR}(x) = \overline{Y} + \langle x, \widehat{\rho} \rangle.$$

To assess the predictive performance, we use the same criteria (i.e. $RMSEP(r_L)$) and follow the same procedure depicted previously (100 randomly splits for building 100 pairs of learning and testing subsamples). In addition, the choice of the penalty (smoothing) parameter λ is derived from a cross-validation procedure systematically based on the learning sample. Figure 7 displays the results for both datasets; top and middle panels are similar to that described in the previous section whereas those at the bottom plot the estimated functional parameter $\widehat{\rho}$. The estimated functional parameter is an interesting interpretable tool. It allows to identify ranges of wavelengths playing a minor (resp. major) role named the set of wavelengths λ such that $|\widehat{\rho}(\lambda)|$ is smaller (resp. greater) than some threshold value. However, the price to pay for getting interpretable tools implies a significant loss in terms of prediction. The median of the $RMSEP(r_{FLR})$ is around 0.07 (resp. 0.30) for the food (resp. petroleum) industry dataset.

5 Models of intermediate dimensionality

According to the previous developments, on one hand model with high dimensionality like the functional nonparametric regression model may lead to powerful predictive performance but no interpretable graphical tool is available. On the other hand, rigid model with low dimensionality like the linear one offers interpretable graphical output but with a possible loss in terms of predictive quality. In order to take advantages of each previous models, an interesting way consists in proposing regression models of intermediate dimensionality balancing predictive performance and interpretability need.

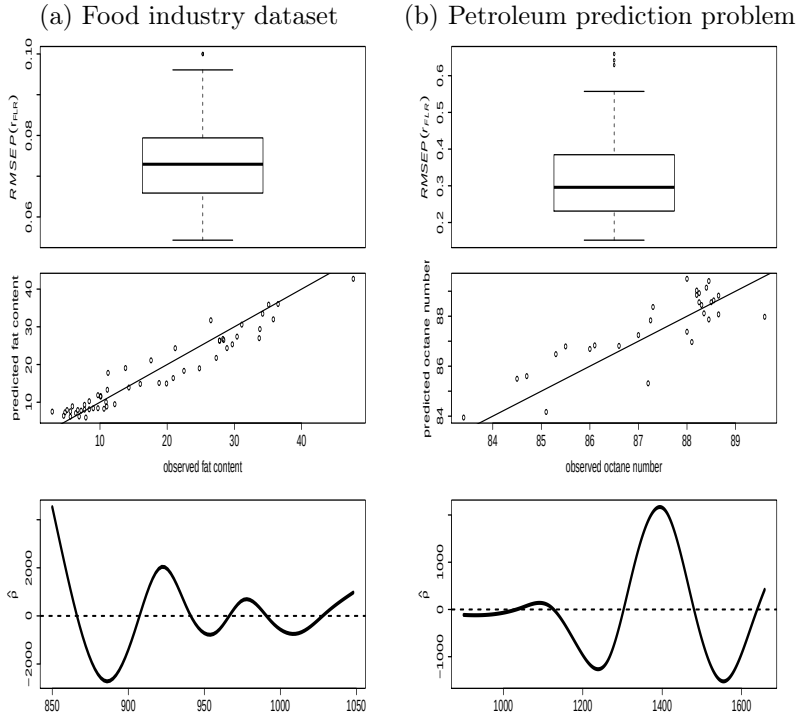


Figure 7: FLR with original spectra: out-of-sample performance and functional parameter.

5.1 Directional additive modelling

The terminology *directional additive modelling* encompasses models assuming that the functional covariate impacts on the scalar response only through a few relevant directions via nonlinear additive link functions. One of the simplest, the functional index model (see for instance [2] or [1]), assumes that there exists a real parameter μ, one functional direction ρ and one additive component g (i.e. real-valued function) such that $r(X) = \mu + g(\langle X, \rho \rangle)$; the real parameter μ, the informative direction ρ and the link function g are unknown and have to be estimated. Recent works ([10] and [16]) extended this idea to D directions $\rho_1, \ldots, \rho_D$ by developing the more sophisticated modelling

$$r(X) = \mu + g_1(\langle X, \rho_1 \rangle) + \cdots + g_D(\langle X, \rho_D \rangle)$$

where the D informative directions $\rho_1, \ldots, \rho_D$ and the D additive components $g_1, \ldots, g_D$ (also called link functions) are unknown smooth functions that have to be estimated; the unknown real parameter μ is also unknown. This model, called *functional projection pursuit regression* (FPPR) is an extension to the functional setting of the popular *projection pursuit regression* (see for instance [23] and [24]). Of course, the number of functional directions D has to be

reasonable in order to avoid overparametrization situation and consequently identifiability issue (see the recent works [27] and [45]). For this model, the set of constraints can be expressed as:

$$\mathcal{C}_{FPPR} = \{r : \mathcal{F} \to \mathbb{R}, \forall x \in \mathcal{F}, r(x) = \mu + g_1(\langle x, \rho_1 \rangle) + \cdots + g_D(\langle x, \rho_D \rangle)\}.$$

Clearly, the FPPR dimensionality is much higher than the functional linear regression one (i.e. $\mathcal{C}_{FLR} \subset \mathcal{C}_{FPPR}$); the FPPR is more flexible than FLR. Introducing several embedded functional parameters makes much more complex the estimating mechanism. Although various implementions of FPPR are available in the litterature (see again [10] and [16] for more details and references therein), in order to simplify this intensive computational algorithm, [19] proposed a new approach based on average derivative ideas. This last method is used here for deriving the D estimated functional directions $\widehat{\rho}_1, \ldots, \widehat{\rho}_D$ and the D estimated additive components $\widehat{g}_1, \ldots, \widehat{g}_D$. This new method, based on the nonparametric estimation of the functional directional derivative r_x defined in (1), allows to estimate simultaneously the D functional directions $\rho_1, \ldots, \rho_D$. Let r_{FPPR} be the FPPR estimator of r:

$$r_{FPPR}(x) = \overline{Y} + \widehat{g}_1(\langle x, \widehat{\rho}_1 \rangle) + \cdots + \widehat{g}_D(\langle x, \widehat{\rho}_D \rangle)$$

Out-of-sample performance. We compute $RMESP(r_{FPPR})$ according to the same scheme as for the previous models. Concerning the food industry dataset (Figure 8 (a)), the predictive ability of r_{FPPR} is higher than r_{FLR} but lower than

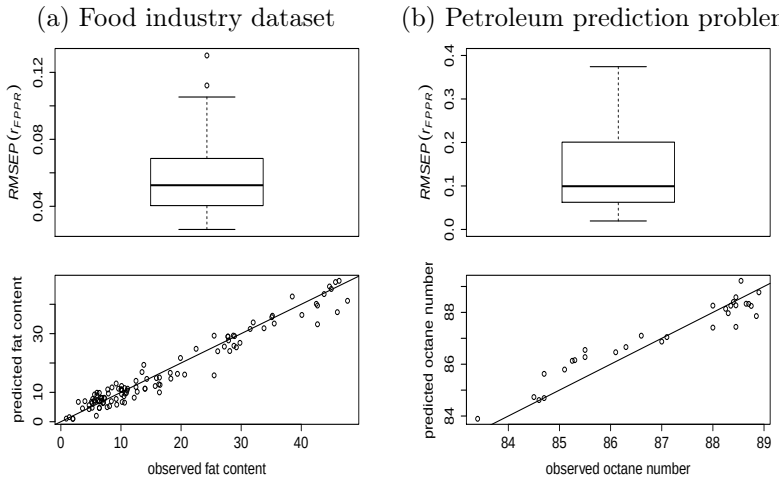


Figure 8: FPPR: out-of-sample performance.

r_{FNPR}; using more flexible model leads to better predictive accuracy. About the petroleum example, Figure 8 (b) shows the nice out-of-sample performance

of FPPR which is much better than FLR but also better than FNPR.

Interpretable outputs. In addition of its well predictive behavior, the interest of FPPR is to produce graphical tools through the estimated functional directions and additive components. To this end, FPPR is launched with the whole sample of both datasets. Figure 9 focuses on the food industry example for which a 2-dimensional FPPR has been implemented for studying the relationship between the fat content and the near-infrared spectra; one has at hand $D = 2$ estimated functional directions (i.e. $\widehat{\rho}_1$ and $\widehat{\rho}_2$) and also $D = 2$ estimated additive components (i.e. $\widehat{g}_1$ and $\widehat{g}_2$). To make easier the interpretation

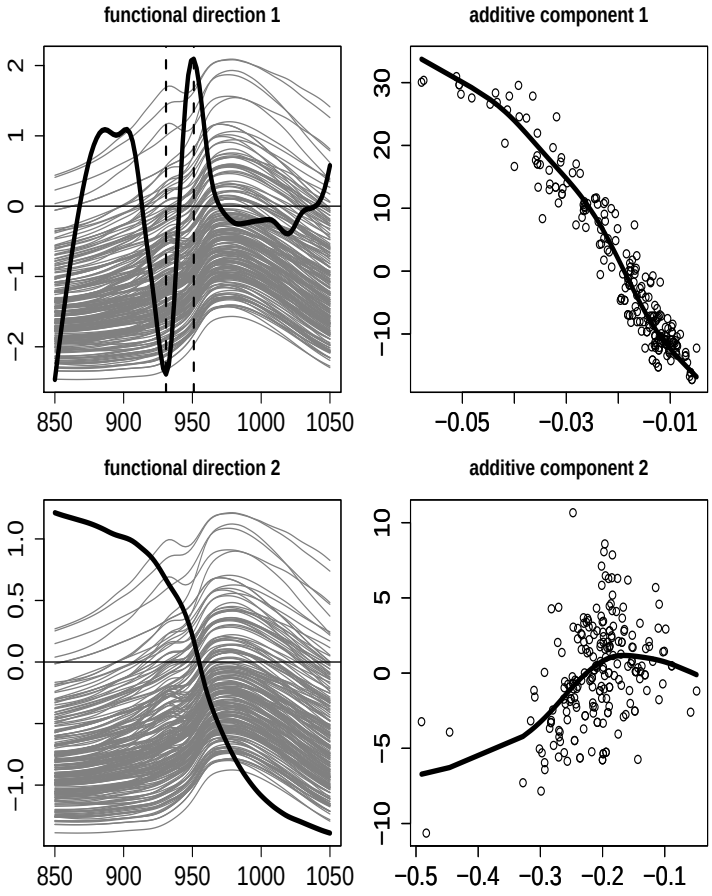


Figure 9: FPPR outputs for food industry dataset: estimated functional directions and additive components.

of the estimated functional directions (continuous black thick line), they have been superimposed on the 215 original spectra (in gray) in the left panels. One

remarks two heavy peaks (vertical dashed lines) on the first estimated functional direction $\widehat{\rho}_1$ (top-left panel). The first one (minimum) around 930 nm identifies clearly a secondary bumb which appears sometimes in the spectra; the second peak (maximum) corresponds to the hollow between the secondary bumb and the main one (when it occurs). The second estimated functional direction $\widehat{\rho}_2$ (bottom-left panel) reduces the role played by the middle of the spectra emphasized with $\widehat{\rho}_1$. Regarding associated additive components (right panels), they point out the need of considering some nonlinear shape. FPPR outputs for the petroleum dataset are displayed in Figure 10. Here, a simple one-dimensional FPPR is sufficient to describe the relationship between the octane number and the spectra. Then, only one functional direction $\widehat{\rho}_1$ and additive component $\widehat{g}_1$ are estimated; remember that the one-dimensional FPPR (i.e. $D = 1$) is also called functional index model. Similarly to left panels of Figure 9, $\widehat{\rho}_1$ (con-

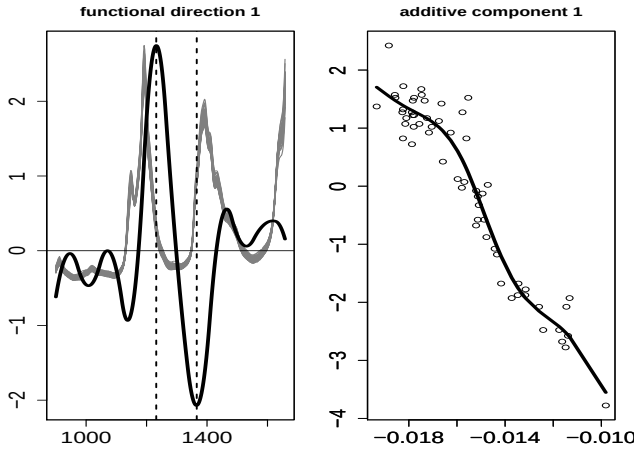


Figure 10: FPPR outputs for petroleum industry dataset: estimated functional direction and additive component.

tinuous black thick line) is superimposed on the 60 original spectra (in gray). This graphics indicates that the wavelengths playing a major role in terms of prediction are located just after 1200 nm and just before 1400 nm; they do not correspond to the peaks of the spectra. The right panel emphasizes again the nonlinear feature of the link function $\widehat{g}_1$.

5.2 Parsimonious nonlinear regression model

In our examples, if we forget the implicit order of the wavelengths, the ith discretized spectrometric curves $X_i(\lambda_1), \ldots, X_i(\lambda_p)$ can be viewed as a standard p-dimensional covariate $X = \{X_{i,1}, \ldots, X_{i,p}\}$ with, for $j = 1, \ldots, p$, $X_j = X_i(\lambda_j)$. So, our challenge is to regress nonlinearly a scalar response on a quite high-dimensional covariate. A different but useful way of handling such a situation consists to propose, for $i = 1, \ldots, n$, the following parsimonious nonparametric

regression:

$$Y_i \;=\; r_{\mathcal{J}}(X_i^{\mathcal{J}}) + \epsilon_i,$$

where $\mathcal{J}$ is an unknown small subset of $\{1, \ldots, p\}$ and $X_i^{\mathcal{J}}$ stands for the subvector $\{X_{i,j};\, j \in \mathcal{J}\}$ and $r_{\mathcal{J}}$ is an unknown multivariate smooth function. This model is parsimonious in that sense it assumes $E(Y_i|X_i) = E(Y_i|X_i^{\mathcal{J}})$. This means that only few covariates are nonparametrically active; the subset $\mathcal{J}$ is usually called active set of covariates. This model is a direct extension of the sparse linear regression methods intensively studied in the literature (see for instance least absolute shrinkage and selection operator [40], smoothly clipped absolute deviation [14], least angle regression [12], Dantzig selector [7] and [5] for a recent overview on this topic).

So, the main aim is to estimate the active subset $\mathcal{J}$ and the corresponding multivariate regression function $r_{\mathcal{J}}$. [18] developed a first approach by using a stepwise forward algorithm based on minimizing a cross-validation criterion. Recently, [17] improved significantly this last work by proposing a new algorithm enlarging the class of possible combinations of covariates retained at each step. Here, we implemented this new algorithm in order to estimate the active subset; a standard linear local regressor (see for instance [13]) is used to derive the estimator r_{NOVAS} of the corresponding multivariate regression function $r_{\mathcal{J}}$ where the abbreviation NOVAS stands for NOnparametric VAriable Selection. In this sparse model, the set of constraints $\mathcal{C}_{NOVAS}$ is just the set of real-valued multivariate functions satisfying regularity assumptions like continuity, differentiability, etc.

We again follow the same scheme for assessing the out-of-sample performance of NOVAS which are displayed in Figure 11. It is worth noting that

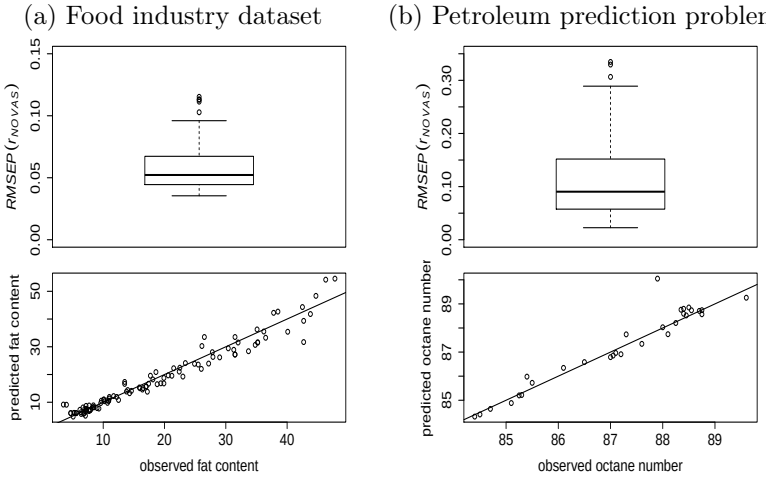


Figure 11: NOVAS with original spectra: out-of-sample performance.

the predictive power of NOVAS is similar to that obtained with the functional

projection pursuit regression (FPPR). What about the interpretability of the NOVAS method? Figure 12 gives an interesting answer to this question which is obtained when NOVAS is rerun on the whole sample for both datasets. In our

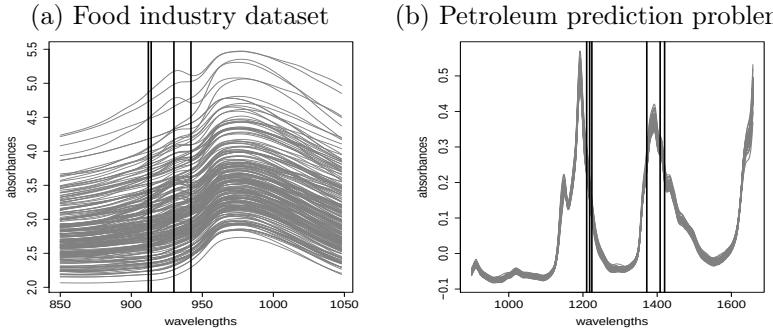


Figure 12: NOVAS with original spectra: selected wavelengths.

situation, selecting covariates amounts to retain wavelengths. So, the vertical black lines, superimposed on the original curves (in gray), identify the location of the selected wavelengths. About the food industry example (Figure 12 (a)), four wavelengths are retained: 912 nm, 914 nm, 930 nm and 942 nm. This result confirms what was observed with the FPPR: the secondary peak (around 930 nm) and the valley just after (around 942 nm) plays a major role. In addition, NOVAS emphasizes two wavelengths just before the secondary peaks (912 nm and 914 nm). Figure 12 (b) displays the six selected wavelengths for the petroleum industry dataset. It corroborates also the conclusions of FPPR: three wavelengths just after the first main peak (1210 nm, 1218 nm and 1224 nm) and one just before the second main peak (1372 nm). Two additional wavelengths are retained (1408 nm and 1420 nm) which highlight the importance of the range just after this second main peak.

6 Boosting approach

Boosting methodology is a generic statistical approach aiming to combine several methods. Generally, one has at hand several statistical technics for analyzing a given dataset. Most of the time, people implements them step by step and try to extracts all relevant informations. Another strategy consists in combining the obtained informations and to rerun the methods by integrating these new knowledges. This is what we propose to do here in a basic but efficience way.

1. *Starting point: FNPR.* The key point is the crucial information obtained thanks to FNPR (functional nonparametric regression): the twice (resp. once) differentiated curves are much more informative than the original ones for the food (resp. petroleum) industry dataset. So, the simple idea is to apply NOVAS (nonparametric variable selection) on the differentiated spectrometric curves.

2. *Nonparametric variable selection (NOVAS).* We propose to boost NOVAS by considering the once or twice differentiated spectra according to the targeted dataset instead of the original ones. Figure 13 details the results; the median of $RMSEP(r_{NOVAS})$ is around 0.009 (resp. 0.05) for the food (resp. petroleum) example. The predictive power is significantly improved when replacing original curves with their once or twice differentiated counterpart. Next plots (see Figure 14) locates (vertical black lines)

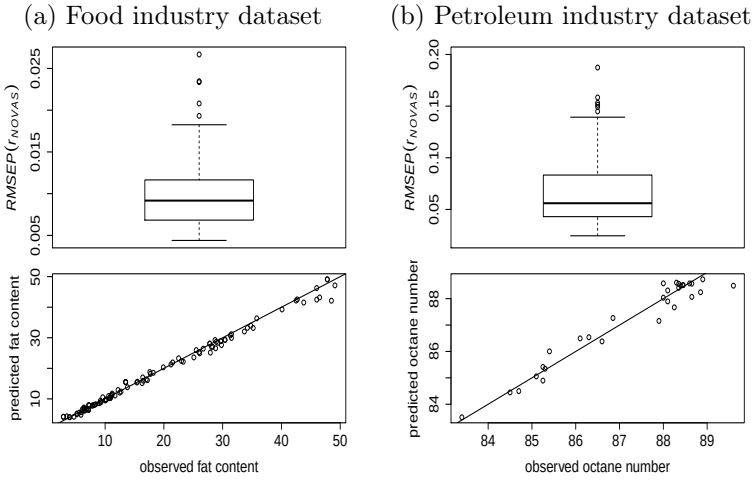


Figure 13: Out-of-sample performance of NOVAS: (a) (resp. (b)) uses twice (resp. once) differentiated curves.

the selected wavelengths for each dataset which are superimposed on their corresponding differentiated curves. About the food industry example (see

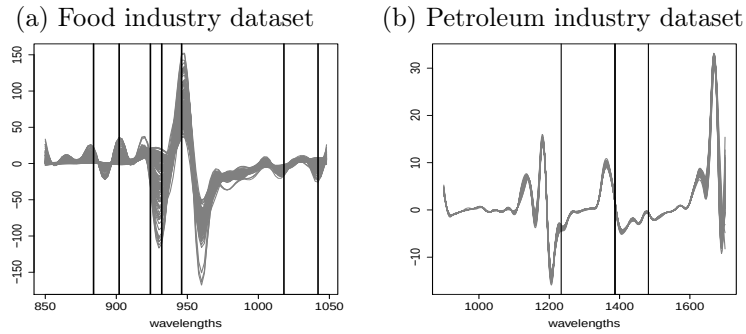


Figure 14: NOVAS: selected wavelengths.

Figure 14 (a)), the seven selected wavelengths (884 nm, 902 nm, 924 nm, 932 nm, 946 nm, 1018 nm and 1042 nm) are dispatched along the whole

range 850 nm - 1050 nm, identifying some particular shapes (valleys and peaks) of the twice differentiated curves. Figure 14 (b) focuses on the petroleum industry problem. For this dataset, only two wavelengths have been selected emphasizing the major role played by the wavelengths in the range 1200 nm - 1300 nm.

So, integrating in NOVAS the information derived from FNPR allows to observe a predictive gain in comparison with what we obtained when NOVAS was applied on the original curves.

3. *Back to the functional projection pursuit (FPPR).* Considering informations coming from FNPR and NOVAS, we propose to boost FPPR by taking benefit of our current knowledge. FNPR indicates that the twice (resp. once) differentiated spectra are more informative for the food (resp. petroleum) industry dataset. In addition, NOVAS tell us that the whole range of wavelengths (i.e. 850 nm - 1050 nm) is relevant for the food example whereas only wavelengths in the range 1200 nm - 1300 nm seems to be important for the petroleum predictive problem. Consequently: food industry dataset $\rightarrow$ FPPR is applied to the twice differentiated spectra and the whole range of wavelengths is considered, petroleum industry dataset $\rightarrow$ FPPR is applied to the once differentiated spectra and we take into account only wavelengths in the range 1200 nm - 1300 nm.

   (a) *Out-of-sample performance.* Figure 15 gives an idea on the predictive

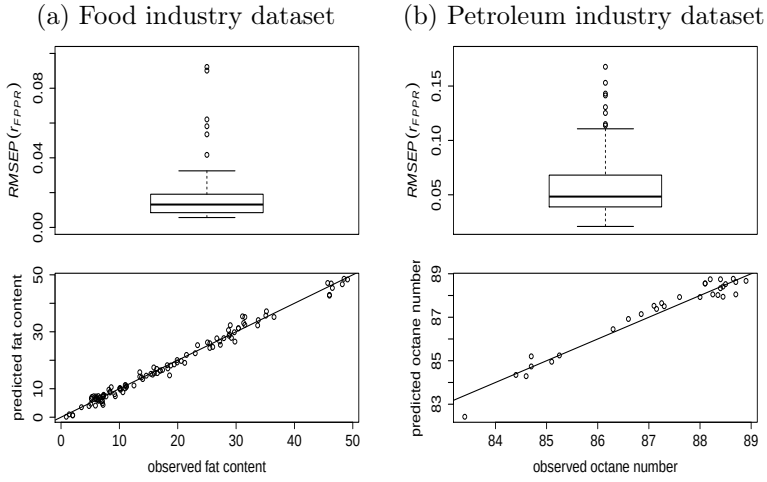


Figure 15: Out-of-sample performance of FPPR: (a) (resp. (b)) uses twice (resp. once) differentiated spectra.

quality of FPPR with conditions of use detailed just before. When comparing the results of FPPR with the original curves, it is clear that FPPR works much better in this setting.

(b) *Interpretable outputs.* For the food (resp. petroleum) industry example, a 2-dimensional (resp. 1-dimensional) FPPR is estimated. Figure 16 displays the outputs for the food industry example. The second derivatives of spectra are plotted in the background (in gray) with a suitable scale. The first estimated functional direction fits the

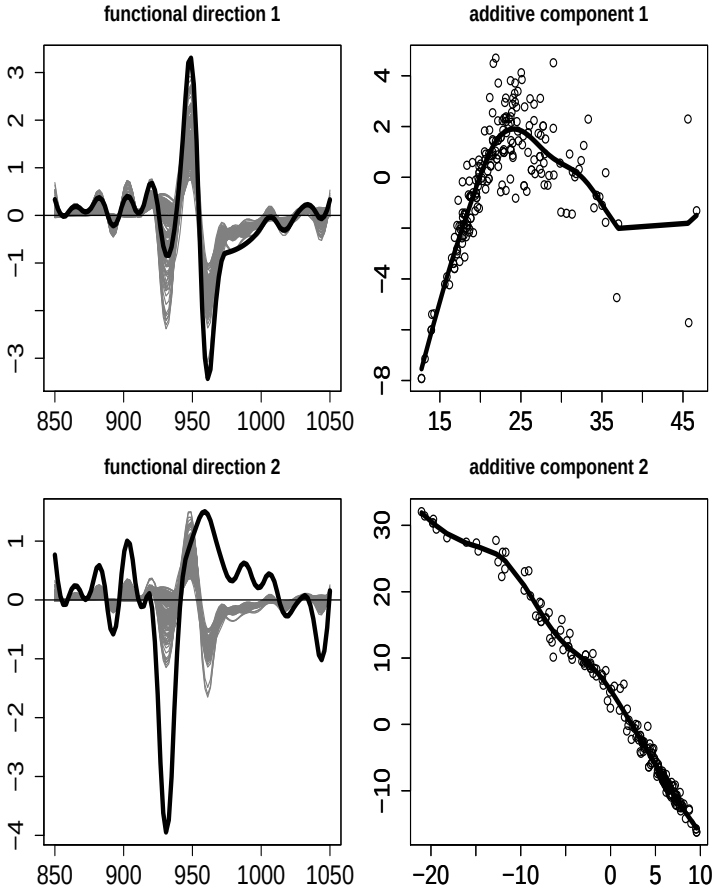


Figure 16: FPPR outputs for food industry dataset: estimated functional directions and additive components from twice differentiated spectra.

main shape of the twice differentiated spectra whereas the second estimated functional direction identifies clearly the first main minimum reached by the second derivative of the spectra (around 930 nm). These results corroborates what was obtained previously with NOVAS. About the petroleum industry dataset, FPPR outputs are displyed in Figure 17; the first derivative of the rescaled spectra are plotted in the background (in gray). The functional direction indi-

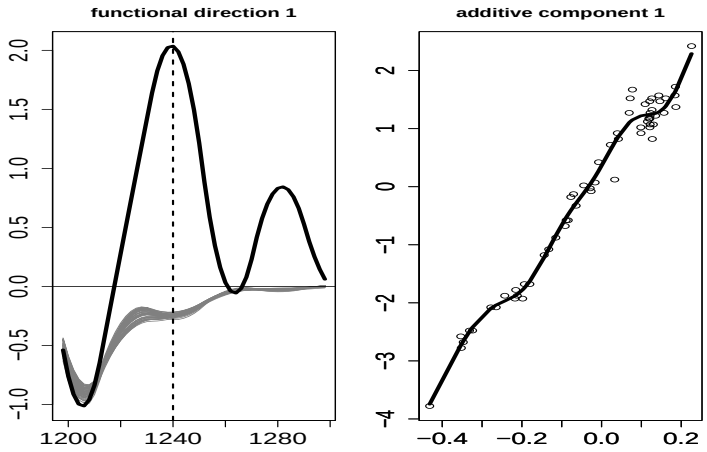


Figure 17: FPPR outputs for food industry dataset: estimated functional directions and additive components from twice differentiatrd spectra.

cates a slight valley (around 1240 nm) which seems to be important for predicting octane numbers.

As conclusion, one can say that combining informations derived from FNPR and NOVAS allows to use FPPR in a more efficience way.

7 Resources available online for functional data analysis

We focused only on four regression models to analyze our datasets. Of course, many other methods can be applied. So, before ending this work, one mentions for practitioners various resources available online dealing with functional data. The R [34] programming environment is a useful software integrating several packages developed for handling functional data:

- the *fda* package [37] for linear models dealing with functional data analysis,
- the *fdaMixed* package [28] for mixed model taking into account functional data,
- the *fda.usc* package [15] includes complementary exploratory and descriptive tools dealing with functional data analysis,
- the *fds* package [38] contains functional data sets; the petroleum industry dataset can be found in this package,
- the *fpca* package [33] deals with the restricted maximum likelihood estimation for functional principal components analysis,

- the *ftsa* package [25] for functional time series analysis,
- the *MFDF* package [11] is specially designed to model functional data in finance by using generalized linear model,
- the *rainbow* package [39] for visualizing functional data,

It is worth noting that all analyses and figures presented i this work have been provided by R. In addition of the previous R packages, other useful resources are available online:

- the companion website of [36] gives numerous complementary materials at www.functionaldata.org,
- the companion website of [22] proposes nonparametric approaches for handling functional data; datasets, R routines, examples of use and much more are available at http://www.math.univ-toulouse.fr/staph/npfda. It is worth noting that all methods presented in this work have been implemented with R; the R routines corresponding to FPPR and NOVAS will be available on this website,
- the PACE website which proposes useful materials and package for Functional Data Analysis and Empirical Dynamics written in Matlab; these methods are able to handle sparsely as well as densely sampled functional data.

8 Conclusion

The main contribution of this paper is to detail and compare the results of four regression models when explaining a scalar responses with near-infrared spectra. Although two of them are now very standard (the functional linear model and the functional nonparametric regression), the two others (functional projection pursuit and parsimonious model) are very recent. The intermediate dimensionality of these two new functional regression approaches is a key point; it provides useful interpretable outputs. Moreover, their flexibility is sufficiently high to catch nonlinear relationship leading to good predictive behaviour. But, if we boost them by integrating the most relevant informations coming from standard use of all these methods, it is possible to improve significantly their predictive performance as well as their interpretability.

Of course, our connecting thread in this work was two spectrometric datasets but the same methodology can be extended to other kind of functional data. For instance functional processes (time series may be viewed as a particular case of functional process; see for more details the monography [4]) provides numerous examples containing dependent functional data and the implementation of the presented methods remains valid for such datasets. In the near future, one can expect to develop useful interpretable tools for handling much more complex data like collection of surfaces, hyperspectral images, etc. It is worth noting

that the functional projection pursuit and the nonparametric variable selection can be implemented easily with high-dimensioned covariates (i.e. not necessary functional variable) with possible application to genomics and more generally to all domains dealing with high-dimensional data.

To conclude, models of intermediate dimensionality in the high-dimensional setting is certainly a highway for deriving new useful statistical methods.

References

[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] U. Amato, A. Antoniadis, and De Feis I. Dimension reduction in functional regression with application. *Comput. Statist. Data Anal.*, 50:2422–2446, 2006.

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

[4] D. Bosq. *Linear processes in function spaces: theory and applications*, volume 149. Springer Verlag, 2000.

[5] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[6] T. Cai and P. Hall. Prediction in functional linear regression. *Ann. Statist.*, pages 2159–2179, 2006.

[7] E. Candès and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35:2313–2351, 2007.

[8] H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statist. Probab. Lett.*, 45(1):11–22, 1999.

[9] H. Cardot, A. Mas, and P. Sarda. Clt in functional linear regression models. *Probab. Theory Related Fields*, 138(3-4):325–361, 2007.

[10] D. Chen, P. Hall, and H.-G. Müller. Single and multiple index functional regression models with nonparametric link. *Ann. Statist.*, 39:1720–1747, 2011.

[11] W. Dou. *MFDF: Modeling Functional Data in Finance*, 2009. R package version 0.0-2.

[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32:407–499, 2004.

[13] J. Fan and I. Gijbels. *Local Polynomial Modeling and its Applications*. Chapman and Hall, London, 1996.

[14] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360, 2001.

[15] M. Febrero-Bande and M. Oviedo de la Fuente. *fda.usc: Functional Data Analysis and Utilities for Statistical Computing (fda.usc)*, 2012. R package version 0.9.7.

[16] F. Ferraty, A. Goia, E. Salinelli, and P. Vieu. Functional projection pursuit regression. *TEST*, 22:293–320, 2013.

[17] F. Ferraty and P. Hall. An algorithm for nonlinear, nonparametric model choice and prediction. *Submitted work*.

[18] F. Ferraty, P Hall, and P. Vieu. Most-predictive design points for functional data predictors. *Biometrika*, 97(4):807–824, 2010.

[19] F. Ferraty, J. Park, and P. Vieu. Average derivative projection pursuit regression. *Submitted work*.

[20] F. Ferraty and Y. Romain, editors. *The oxford handbook of functional data analysis*. Oxford University Press New York, 2011.

[21] F. Ferraty and P. Vieu. The functional nonparametric model and application to spectrometric data. *Comput. Statist.*, 17(4):545–564, 2002.

[22] F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer, New York, 2006.

[23] J. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 1981:817–823, 1981.

[24] P.J. Huber. Projection pursuit. *Ann. Statist.*, 13:435–475, 1985.

[25] R. Hyndman and H. Shang. *ftsa: Functional time series analysis*, 2012. R package version 3.1.

[26] J. Kalivas. Two data sets of near infrared spectra. *Chemometr. Intell. Lab.*, 37(2):255–259, 1997.

[27] W. Lin and K. Kulasekera. dentifiability of single-index models and additive-index models. *Biometrika*, 94:496–501, 2007.

[28] B. Markussen. *fdaMixed: Functional data analysis in a mixed model framework*, 2011. R package version 0.1.

[29] H. Martens and T. Naes. *Multivariate calibration*. Wiley, 1992.

[30] H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *Ann. Statist.*, pages 774–805, 2005.

[31] B. Osborne and T. Fearn. *Near Infrared Spectroscopy in Food Analysis*. Wiley New York, 1986.

[32] J. Peng and H.-G. Müller. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Stat.*, pages 1056–1077, 2008.

[33] J. Peng and D. Paul. *fpca: Restricted MLE for Functional Principal Components Analysis*, 2011. R package version 0.2-1.

[34] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[35] J. Ramsay and B. Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer New York, 2002.

[36] J. Ramsay and B. Silverman. *Functional data analysis*. Springer New York, 2nd edition, 2005.

[37] J. Ramsay, H. Wickham, S. Graves, and G. Hooker. *fda: Functional Data Analysis*, 2012. R package version 2.2.8.

[38] H. Shang and R. Hyndman. *fds: Functional data sets*, 2011. R package version 1.6.

[39] H. Shang and R. Hyndman. *rainbow: Rainbow plots, bagplots and boxplots for functional data*, 2012. R package version 2.8.

[40] R. Tibshirani. Regression analysis and selection via the lasso. *J. R. Stat. Soc. B*, 58:267–288, 1996.

[41] W. Yang, H.-G. Müller, and U. Stadtmüller. Functional singular component analysis. *J. R. Stat. Soc. B*, 73(3):303–324, 2011.

[42] F. Yao and H.-G. Müller. Functional quadratic regression. *Biometrika*, 97(1):49–64, 2010.

[43] F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.*, 100(470):577–590, 2005.

[44] F. Yao, H.-G. Müller, and J.-L. Wang. Functional linear regression analysis for longitudinal data. *Ann. Statist.*, 33(6):2873–2903, 2005.

[45] M. Yuan. On the identifiability of additive index models. *Statist. Sinica*, 21:1901–1911, 2011.