Comput Stat (2015) 30:673–692
DOI 10.1007/s00180-014-0530-1

ORIGINAL PAPER

A partitioned Single Functional Index Model

Aldo Goia · Philippe Vieu

Received: 22 January 2014 / Accepted: 21 August 2014 / Published online: 2 September 2014

Abstract Given a functional regression model with scalar response, the aim is to present a methodology in order to approximate in a semi-parametric way the unknown regression operator through a single index approach, but taking possible structural changes into account. Our paper presents this methodology and illustrates its behaviour both on simulated and real curves datasets. It appears, from an example of interest in spectrometry, that the method provides a nice exploratory tool both for analyzing structural changes in the spectrum and for visualizing the most informative directions, still keeping good predictive power. Even if the main objective of this work is to discuss applied issues of the method, asymptotic behaviour is shortly described.

Keywords Functional predictor · Single Index Model · Additive models · Structural points · Spectrometric data

1 Introduction

The well-known functional regression model with scalar response (see Horváth and Kokoszka 2012; Ferraty and Vieu 2006 or Ramsay and Silverman 2005, for general discussions) postulates a relation between a real random variable Y and a random function X, which belongs to a functional space $\mathcal{F}$ of real functions defined on a compact interval I, via a real valued operator r as follows:

A. Goia (✉)
Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale,
Via Perrone, 18, 28100 Novara, Italy
e-mail: aldo.goia@eco.unipmn.it

P. Vieu
Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118, Route de Narbonne,
31062 Toulouse Cedex, France
e-mail: philippe.vieu@math.univ-toulouse.fr

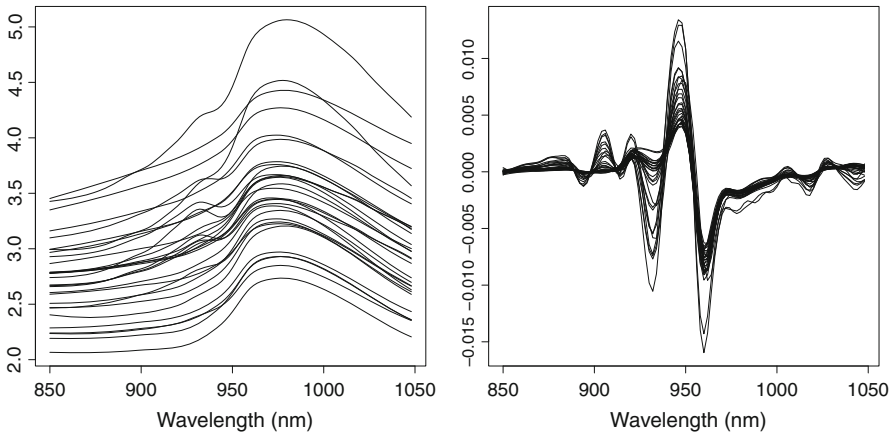

Fig. 1 Spectrometric curves from Tecator dataset (*left panel*) and their second derivatives (*right panel*)

$$Y = r[X] + \mathcal{E}$$

where $\mathcal{E}$ is a centered real random error uncorrelated with the regressor.

It is opportune to note that some parts of the curves, or even some of their particular points, may be more interesting than others in order to explain the relation between X and the response Y. Various approaches have recently been developed on this topic, including the partial no effect tests proposed in Cardot et al. (2004) in the context of linear models, the structural nonparametric tests introduced in Delsol et al. (2011) and Delsol (2013), or the methods based on variables selection as for instance in Ferraty et al. (2010) and Aneiros et al. (2011) for a nonparametric model or in McKeague and Sen (2010) in the functional linear framework.

Indeed one can suppose, in some situations, that specific parts of the whole curve X act in a different way for explaining the response Y. Hence, partitioning I in s contiguous sub-intervals I_j, and denoting X^j the restriction of X to I_j, one can write the following additive decomposition of the regression operator:

$$r[X] = \sum_{j=1}^{s} r_j\left[X^j\right]. \tag{1}$$

Consider for instance the problem of estimating the chemical composition of a given aliment by using spectrometric curves, namely the absorbances of light irradiated on the aliment varying the wavelength of emission. In the chemometric literature it is known that some features of the spectra (see Leardi 2003) or some specific parts of the spectrometric curves (see Delsol 2013) are more interesting than others to predict the proportion of a specific substance.

Figure 1 shows the near-infrared absorbance spectra corresponding to 215 pork samples, recorded on a Tecator Infratec Food and Feed Analyzer, and the second derivatives of such spectral curves. Such dataset has become a benchmark in functional regression studies: the aim is to predict the percentage of fat contained in each

sample of meat from its near-infrared spectrum. Some empirical evidences on such case study emerge from literature: as pointed out in Ferraty et al. (2013), the regression function exhibits a nonlinear nature; moreover the role of some specific points of the spectrometric curve in explaining the fat content has been emphasized in Ferraty et al. (2010). Combining previous observations, it is reasonable to expect that the decomposition (1) can lead to a regression model with better prediction ability, and which can be able to provide a key to better understand the relationship between predictor and response. These data will be presented with more details and analyzed further along this paper (see Sect. 4).

The decomposition (1) includes a broad class of modelizations: as, for instance, Linear models with functional coefficients having s points of jump discontinuity (see for instance Horváth and Reeder 2012), or Generalized Linear models (see for instance James 2002) that act on specific parts of the random curve. In these examples, the shape of r_js is entirely specified (since they are modeled in a parametric way): although that allows to give some interpretations of the estimated coefficients involved, it appears quite restrictive and the specification of the link is difficult to implement in the functional regression context.

On the other hand, a wide class of flexible and useful tools to modelize the regression operator r is represented by the Functional Single Index models (FSIM in the sequel). The main idea is to search the direction $\theta_0 \in \mathcal{F}$ along which the projection of the covariate X captures the most information on the response Y. This presents various interests: Firstly, it is avoiding problems due to the dimensionality which one can be meet in the full nonparametric approach (see Ferraty and Vieu 2002); Secondly, it is much more flexible than standard parametric/linear modelization (see James 2002); Finally, estimating the relevant functional direction θ_0 provides an easily interpretable tool.

The Single Index approach is well-known in the standard multivariate context: the interest both for its prediction abilities and for interpretability is attested by various works appeared in the last two decades: see Härdle et al. (1993), Härdle and Stoker (1989), or Xia and Härdle (2006) for a selected sample of references, and see Härdle et al. (2004) for a general presentation of semi-parametric approaches. The extensions to the functional framework of these ideas, such a functional semi-parametric methodology, have been intensively studied in the literature: conditions for identifiability for FSIM have been introduced in Ferraty et al. (2003) and several estimation techniques are proposed in Ait-Saïdi et al. (2008), Amato et al. (2006) and Ferraty et al. (2011). Moreover, this approach can be seen as the first step of the Functional Projection Pursuit regression developed in Ferraty et al. (2013).

The aim of this paper is to exploit the flexibility of FSIM in the additive decomposition (1) in order to treat situations when structural changes occur. More precisely we write:

$$r_j\left[X^j\right] = g_j\left(\int_{I_j} \theta_j(t)\, X(t)\, dt\right) \tag{2}$$

where g_j is an unknown real link function and θ_j are unknown directions such that $\int_{I_j} \theta_j^2(t)\, dt = 1$.

To the sake of simplicity, we will study specifically the introductory case $s = 2$. More in detail, we introduce an estimation procedure based on a backfitting algorithm where each term (2) is fitted by a procedure combining a spline approximation of the direction and the one-dimensional Nadaraya-Watson kernel regression estimate. Some considerations on the way to obtain asymptotic results are sketched: the crucial aspect emerging is the insensitiveness of the method to dimensionality effects. The selection of the breaking-point for cutting I into two parts is discussed and a fully data-driven method for that is presented. The study is completed with an extended empirical analysis based both on real and simulated data: as well as emphasizing on the good predictive performance of our method, the study highlights the interpretability of the functional directional outputs.

The paper is organized as follows. In Sect. 2 we deepen some technical aspects about the partitioned model and the estimation technique is described. Section 3 is devoted to some computational issues: in Sect. 3.1 the finite sample performances of the approach are illustrated through simulations, whereas the behaviour of the data-driven procedure for choosing the breaking-point is shown in Sect. 3.2. Finally, an application to the spectrometric dataset is presented in Sect. 4. A short discussion on asymptotics is provided in the final "Appendix".

2 Model and methodology

2.1 The partitioned FSIM

Our aim is to study the model (2), but in order to make things clearer we only detail the simplest case when $s = 2$; extensions to higher values of s are straightforward.

Let us fix some notations. Consider a functional r.v. $X = \{X(t), t \in I\}$ and the real r.v. Y defined over the same space. Without loss of generality, we take $I = [0, 1]$ and $\mathbb{E}[X(t)] = 0$ for all t. Define the regression model:

$$Y = r[X] + \mathcal{E} \tag{3}$$

where r is a real value operator and $\mathcal{E}$ is a real random error with finite variance and such that $\mathbb{E}[\mathcal{E}|X] = 0$ a.s.. As usually in the literature, we assume that X take values in the separable Hilbert space $L^2(I)$ of square integrable real functions.

Introduce a breaking-point $\lambda \in (0, 1)$ and split I into two subintervals in the following way:

$$I_1 = [0, \lambda] \qquad I_2 = (\lambda, 1].$$

We define the two-terms Partitioned Functional Single Index Model (PFSIM in the sequel) as

$$Y = \alpha + g_1\left(\int_{I_1} \theta_1(t)\, X(t)\, dt\right) + g_2\left(\int_{I_2} \theta_2(t)\, X(t)\, dt\right) + \mathcal{E} \tag{4}$$

where α is a real coefficient, g_1 and g_2 are some real smooth functions. For standard identifiability reasons one has to assume that the directions θ_j satisfy

$$\int_{I_1} \theta_1^2(t)\,dt = \int_{I_2} \theta_2^2(t)\,dt = 1$$

as well as

$$\int_{I_1} \theta_1(t)\,e_1(t)\,dt = \int_{I_2} \theta_2(t)\,f_1(t)\,dt = 1$$

where e_1 and f_1 are the first elements of some orthonormal bases of $L^2(I_1)$ and $L^2(I_2)$ respectively.

At this stage it is worth being noted the high degree of flexibility of the model. From one side it can be seen as a natural extension of the standard FSIM model as discussed for instance in Ait-Saïdi et al. (2008):

$$Y = \alpha + g\left(\int_I \theta(t)\,X(t)\,dt\right) + \mathcal{E},$$

as well, of course, as an extension of the basic linear model as discussed for instance in Cardot et al. (2003):

$$Y = \alpha + \int_I \theta(t)\,X(t)\,dt + \mathcal{E}.$$

More surprisingly, it can also be seen as a kind of extension of the fully nonparametric model (3) in the sense that it allows the use of an unsmooth operator r, while the nonparametric literature (see Ferraty and Vieu 2006) is based on continuity-type assumptions. Under this perspective, PFSIM provides a useful approximator for the regression operator:

$$r[X] \approx \alpha + g_1\left(\int_{I_1} \theta_1(t)\,X(t)\,dt\right) + g_2\left(\int_{I_2} \theta_2(t)\,X(t)\,dt\right) \tag{5}$$

with constraints $\int_{I_1} \theta_1^2(t)\,dt = \int_{I_2} \theta_2^2(t)\,dt = 1$. It should be noted that in such context, this decomposition is not unique: indeed, one can found two different couples of terms $\left\{\left(g_j, \theta_j\right)_{j=1,2}\right\}$ and $\left\{\left(\widetilde{g}_j, \widetilde{\theta}_j\right)_{j=1,2}\right\}$ such that $\sum_{j=1,2} g_j\left(\int_{I_j} \theta_j(t)\,X(t)\,dt\right)$ $= \sum_{j=1,2} \widetilde{g}_j\left(\int_{I_j} \widetilde{\theta}_j(t)\,X(t)\,dt\right)$. If that lack of unicity may cause problem for interpretating the outputs, it should be stressed that it has no effects on the two main features of the model, namely its interest for detecting existence of a possible breakpoint and its high degree of flexibility that will guarantee nice prediction performances.

2.2 Fitting the partitioned FSIM

Consider now the problem of estimating the link functions g_j and the directions θ_j in the model (4), from a sample $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ of r.v.s identically distributed as

(X, Y). In a first attempt we consider that the breaking point λ is known; the important question of estimating λ in practice will be addressed in Sect. 2.3. From the additive nature of the model, we propose a backfitting algorithm (see e.g. Hastie et al. 2009) in which each term is estimated by an alternating optimization strategy similar to the one used in Ferraty et al. (2013) and whose principle is illustrated in the following.

For $j = 1, 2$, consider the $(q_j + k_j)$-dimensional space of spline functions defined on I_j with order q_j and with $k_j - 1$ interior equispaced knots (with $q_j > 2$ and $k_j > 1$, integers) and let $\left\{B_s^j\right\}$ be normalized B-splines basis of such space. In such basis $\theta_j(t)$ is represented as $\delta_j^T \mathbf{B}_j(t)$, where $\mathbf{B}_j(t)$ is the vector of all the B-splines. To remove trivial ambiguity, each vector δ_j of coefficients is such that its first element is positive, and satisfies the normalization condition:

$$\delta_j^T \int_{I_j} \mathbf{B}_j(t)\, \mathbf{B}_j(t)^T\, dt\, \delta_j = 1. \tag{6}$$

The estimation procedure is based on the algorithm described below, which has been implemented in R code and exploits the Nelder-Mead optimization algorithm (see Nelder and Mead 1965). In the following we denote by $\{(x_i, y_i),\ i = 1, \ldots, n\}$ the observed values of the random pairs (X_i, Y_i).

- **Initialize -** Set $\widehat{\alpha} = n^{-1} \sum_{i=1}^n y_i$, initialize the current residuals $\widehat{\varepsilon}_i = y_i - \widehat{\alpha}$, and fix j ($j = 1$ or $j = 2$).
- **Cycle -** Find $\widehat{\delta}_j$ which minimizes over $\mathbf{d} \in \mathbb{R}^{q_j + k_j}$ the empirical quadratic cross-validation criterion:

$$CV_j(\mathbf{d}) = \frac{1}{n} \sum_{i=1}^n \left[\left(\widehat{\varepsilon}_i - \widehat{g}_j^{[-i]} \left(\mathbf{d}' \mathbf{b}_{j,i} \right) \right)^2 \right] \tag{7}$$

where $\mathbf{b}_{j,l} = \left\langle \mathbf{B}_j, x_l \right\rangle$ and

$$\widehat{g}_j^{[-i]}(z) = \sum_{l \neq i} \frac{K_j\left(\frac{z - \mathbf{d}' \mathbf{b}_{j,l}}{h_j}\right)}{\sum_{l \neq i} K_j\left(\frac{z - \mathbf{d}' \mathbf{b}_{j,l}}{h_j}\right)} \cdot \widehat{\varepsilon}_i$$

with K_j a kernel function and h_j a suitable smoothing parameter. As it is made conventionally in the additive models literature, we assume that $n^{-1} \sum_{l=1}^n \widehat{g}_j^{[-i]} \left(\widehat{\delta}_j' \mathbf{b}_{j,l} \right) = 0$.

Then, update the residuals

$$\widehat{\varepsilon}_i = y_i - \widehat{\alpha} - \widehat{g}_j^{[-i]} \left(\widehat{\delta}_j' \mathbf{b}_{j,l} \right)$$

and swap the value of the index j.

The process is continued until stabilization of the quadratic error measure $n^{-1} \sum_{i=1}^n (y_i - \widehat{y}_i)^2$, where $\widehat{y}_i$ are the estimated values.

The estimator is tuned by three couples of parameters: the order of splines q_1 and q_2, the number of knots k_1 and k_2, and the bandwidths h_1 and h_2. If the order of splines may be fixed to 3, the number of knots k_j has to be chosen conveniently in order to capture to complexity of the shape of the direction θ_j to estimate: classical Akaike Information Criterion, AIC, and the Schwartz Information Criterion, BIC (for a general presentation, see Burnham and Anderson 2002) can be useful in this view. Often, however the choice may be done heuristically. Finally, since the estimator of g_j is an usual nonparametric regression kernel estimator, the choice of the smoothing parameters h_j can be performed by data-driven selectors of the bandwidth such as cross-validation. Due to the nature of Nelder-Mead method, the proposed algorithm is inclined to get stuck in a local minimum: to alleviate this problem, one can use multiple random initialization of the parameters.

2.3 Data-driven breaking-point selection

While the previous procedure is defined for fixed value of the parameter λ, the question of how choosing it in practice is a natural one. The main idea for that is to use the value leading to the minimal prediction error. This work as follows:

- **Step 1** Choose a grid Λ of possible values for λ;
- **Step 2** Compute for each $\lambda \in \Lambda$ the estimates of the direction θ_j and the link functions g_j by running the algorithm defined in Sect. 2.2;
- **Step 3** Choose the value $\widehat{\lambda}$ which minimizes the cross-validation criterion

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i - \widehat{r}_{\lambda}^{[-i]}(x_i)\right)^2\right], \tag{8}$$

where $\widehat{r}_{\lambda}^{[-i]}$ is the leave-one-out version of the estimate computed along the step 2 for the value λ.

The selection method above is easy to implement but it could fail in detecting the "true" breaking-point due to non unicity problems of the approximation (5) stressed at the end of Sect. 2.1. Moreover, in case of misspecification of λ, the extreme flexibility of the approach leads however to obtain a good fitting which counterpoises the erroneous evaluation: the theoretical remarks in "Appendix" will make us clear how this procedure is accurate if one is only looking for predictive performance. Its behaviour on finite sample will be analyzed in Sect. 3.2.

3 Computational issues

3.1 Assessing the performances by simulations

We illustrate the finite sample performances of our procedure, comparing it to several linear and non-linear functional approaches in a series of simulation studies. To

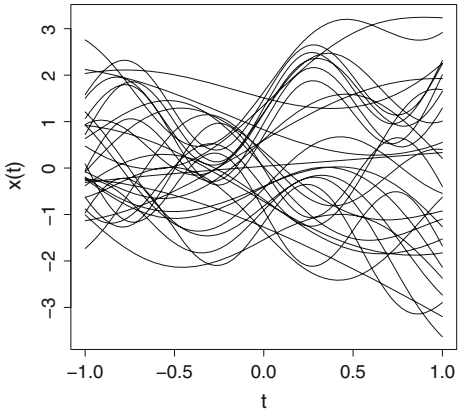

Fig. 2 A random selection of 30 functional predictors used in the simulation experiments

avoid introducing noises connected with a misspecification of the breaking-point, we suppose λ known.

Data were generated according to the following regression models:

$$Y_i = r\left[X_i\right] + \sigma \mathcal{E}_i \qquad i = 1, \ldots, n \tag{9}$$

where $n = 300, \mathcal{E}_i \sim \mathcal{N}(0, 1)$ and $\sigma^2 = \rho^2 Var\left(r\left[X\right]\right)$, ρ controls the signal-to-noise ratio (we used $\rho = 0.1, 0.3$). The functional covariates obey to

$$X_i(t) = a_i + b_i t^2 + c_i \exp(t) + \sin(d_i 2\pi t) \qquad t \in [-1, 1] \tag{10}$$

where a_i, b_i, c_i and d_i are real r.v.s independent and uniformly distributed over $(-1, 1)$, so that $\mathbb{E}\left[X_i(t)\right] = 0, t \in [-1, 1]$. Each functional predictor is discretized over a grid of 200 equispaced design points $\left\{t_j,\ j = 1, \ldots 200\right\}$ to obtain the 300×200 matrix $\left[x_i\left(t_j\right)\right]$. A random selection of these functional data is plotted in the Figure 2.

Regression operators $r\left[X_i\right]$ have been obtained as the sum of two terms acting on $I_1 = [-1, 0]$ and $I_2 = (0, 1]$. As illustrated in the following, they may be linear, generalized linear or full nonparametric terms, so as to cover a wide range of possible regression links and to show how PFSIM behaves in the different cases.

More in detail, we introduced the real functional coefficients:

$$\varphi_1(t) = \kappa_1 \cos(2\pi t^2) \qquad \varphi_2(t) = \kappa_2 \sin\left(\frac{3}{2}\pi t\right)^3$$

where κ_j are such that $\left(\int_{I_j}\left[\varphi_j(t)\right]^2 dt\right)^{1/2} = 1$, and the random functions, obtained by transformations of the original random data X_i:

$$m_1^{X_i}(t) = \sin\left(X_i(t)\right) \qquad m_2^{X_i}(t) = \sqrt{|X_i(t)|}.$$

Then we considered the following cases:

1. The regression operator is linear with a discontinuous functional coefficient:

$$r_1[X_i] = \int_{-1}^{1} \left[\varphi_1(t)\ \mathbf{1}_{t\in I_1} + \varphi_2(t)\ \mathbf{1}_{t\in I_2}\right] X_i(t)\, dt,$$

   where $\mathbf{1}_{t\in A}$ is the indicator function of subset A.
2. The regression operator is linear over I_1 and non linear over I_2. About the second addend, we analyzed both the case of generalized linear structure with cubic link function:

$$r_{2.a}[X_i] = \int_{-1}^{0} \varphi_1(t)\, X_i(t)\, dt + 4\left(\int_0^1 \varphi_2(t)\, X_i(t)\, dt\right)^3,$$

   and the one of a full nonparametric term:

$$r_{2.b}[X_i] = \int_{-1}^{0} \varphi_1(t)\, X_i(t)\, dt + \int_0^1 m_2^{X_i}(t)\, dt.$$

3. Both terms composing $r[X_i]$ are nonlinear:

$$r_{3.a}[X_i] = \sin\left(\pi \int_{-1}^{0} \varphi_1(t)\, X_i(t)\, dt\right) + 4\left(\int_0^1 \varphi_2(t)\, X_i(t)\, dt\right)^3,$$

   or full nonparametric:

$$r_{3.b}[X_i] = \int_{-1}^{0} m_1^{X_i}(t)\, dt + \int_0^1 m_2^{X_i}(t)\, dt.$$

We estimated the previous models with the algorithm illustrated in Sect. 2.2 over training-samples of size 200 with λ fixed to zero. We used cubic splines with the same number of internal knots: $k_1 = k_2 = 3$; the smoothing parameters h_j are selected by a leave-one-out cross-validation procedure. Prediction outcomes were quantified on test-sets of size $n.out = 100$, by the Relative Mean Square Error of prediction:

$$RMSE = \frac{\sum_{i=1}^{n.out}\left(y_i^{out} - \widehat{y}_i\right)^2}{\sum_{i=1}^{n.out}\left(y_i^{out} - \overline{y}\right)^2}$$

where y_i^{out} are the elements of the test-set, $\widehat{y}_i$ are the corresponding estimated values and $\overline{y} = n^{-1}\sum_{i=1}^{n} y_i^{out}$. Each simulation was repeated for 100 times to obtain a frequency distributions of the RMSE.

Prediction results with PFSIM were compared with those obtained from the following competitors:

1. Functional Single Index Model (FSIM) fitted with the first step of the alternating least square algorithm proposed in Ferraty et al. (2013): we used cubic splines

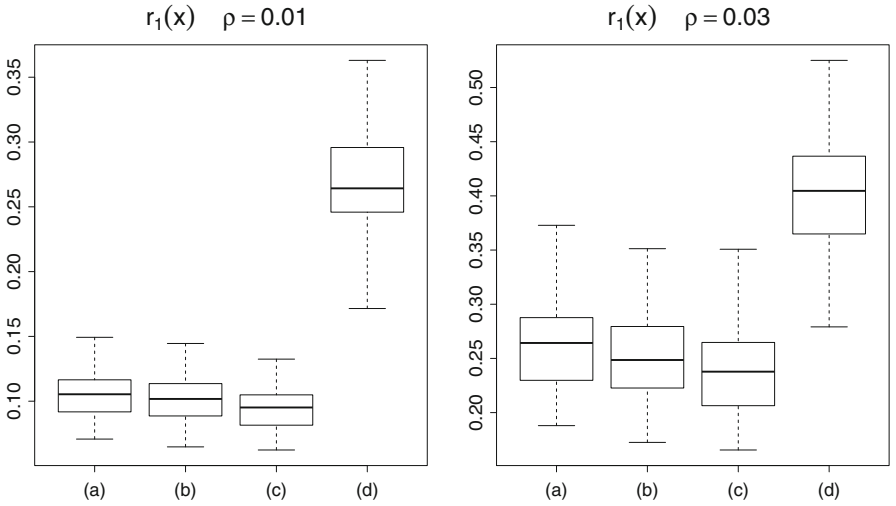

Fig. 3 RMSE for regression model involving operator r_1 with $\rho = 0.1$ and $\rho = 0.3$. **a** stands for PFSIM, **b** for FSIM, **c** for FLM and **d** for FNPM

with 5 knots and a leave-one-out cross-validation procedure for the selection of the bandwidth;

2. Functional Linear Model (FLM) where the functional coefficient was estimated with a penalized B-spline procedure (based on cubic splines with 20 internal knots) and the smoothing parameter in the penalization (controlling second derivatives) was selected with a cross-validation procedure (see e.g. Cardot et al. 2003);
3. Functional Nonparametric Model (FNPM) estimated using the κ -nearest neighbour functional kernel estimator (with κ chosen by local cross-validation) with proximity between curves measured with the classical L^2 norm (for details see Ferraty and Vieu 2006).

Comparison between the empirical distributions of $RMSE$s resulting from the above estimation strategies can be made analyzing Figs. 3, 4, 5, 6 and 7.

The simulations show that our method performs very well in all the examples, also in comparison with the other proposed methods. Indeed, when the model is linear (see Fig. 3) the PFSIM is practically equivalent to the FSIM and the FLM. Moreover, it produces the best prediction performances when the regression operator $r[X]$ is decomposable in two parts of the type $g_j\left(\int_{I_j} \theta(t) X(t) dt\right)$ (see Figs. 4 and 6). Finally, when a full nonparametric term appears, PFSIM widely outperforms FSIM and FLM estimators, and it is equivalent to the full nonparametric approach (see Figs. 5, 7).

From the study it emerges that our approach represents a valid alternative to the pure nonparametric one. Compared to this, one can bring out some advantages: the dimensionality problem is avoided, the task of choosing a "good" semi-norm is skipped, and, in same cases, it is possible to detect some latent structures in the regression operator when they exist.

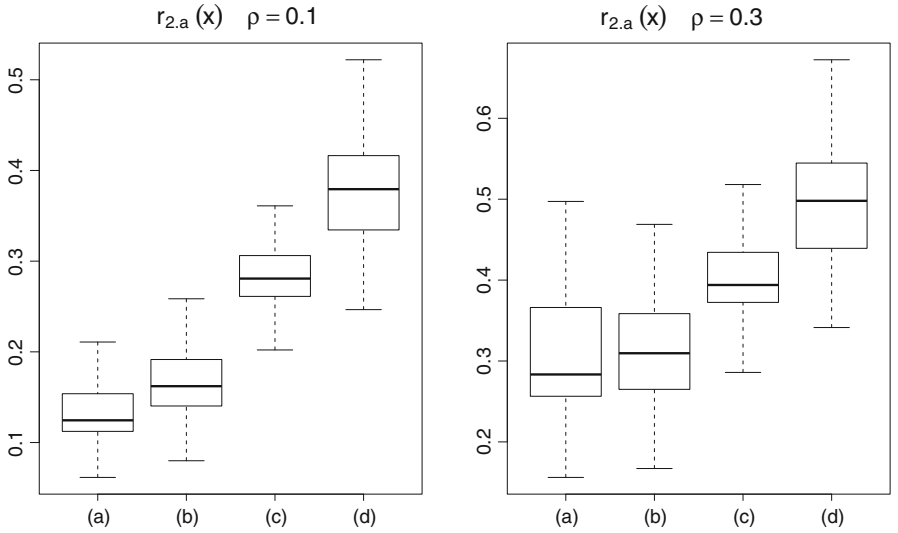

Fig. 4 RMSE for regression model involving operator $r_{2.a}$ with $\rho = 0.1$ and $\rho = 0.3$. **a** stands for PFSIM, **b** for FSIM, **c** for FLM and **d** for FNPM

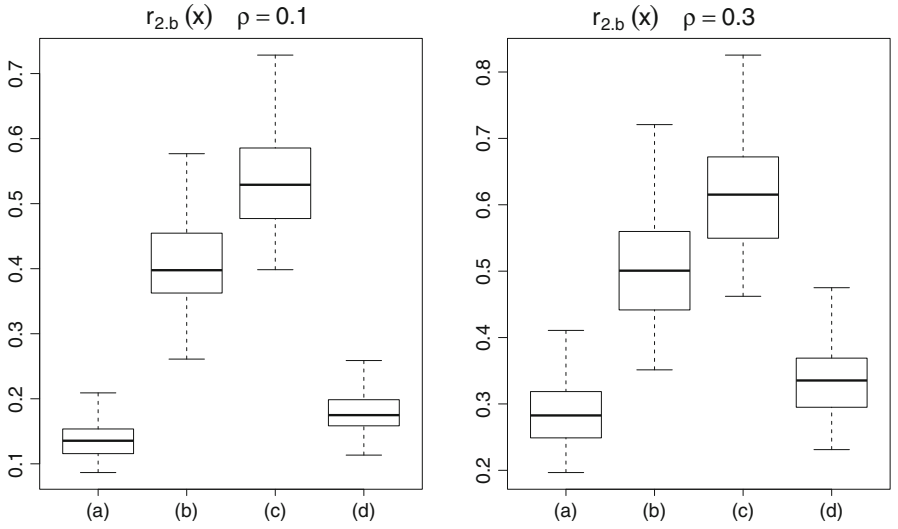

Fig. 5 RMSE for regression model involving operator $r_{2.b}$ with $\rho = 0.1$ and $\rho = 0.3$. **a** stands for PFSIM, **b** for FSIM, **c** for FLM and **d** for FNPM

To appreciate the latter aspect, we reproduce in Fig. 8 the estimates of link functions g_j and directions θ_j when the responses y_i are generated by a model with the regression operator $r_{2.a}$ and $\rho = 0.1$. In this case the graphs highlight the nature of the link between the predictor and the response: it is possible to detect the existence of a linear relation between the first part of the covariates and Y_i, and a nonlinearity in correspondence to the second part of the interval.

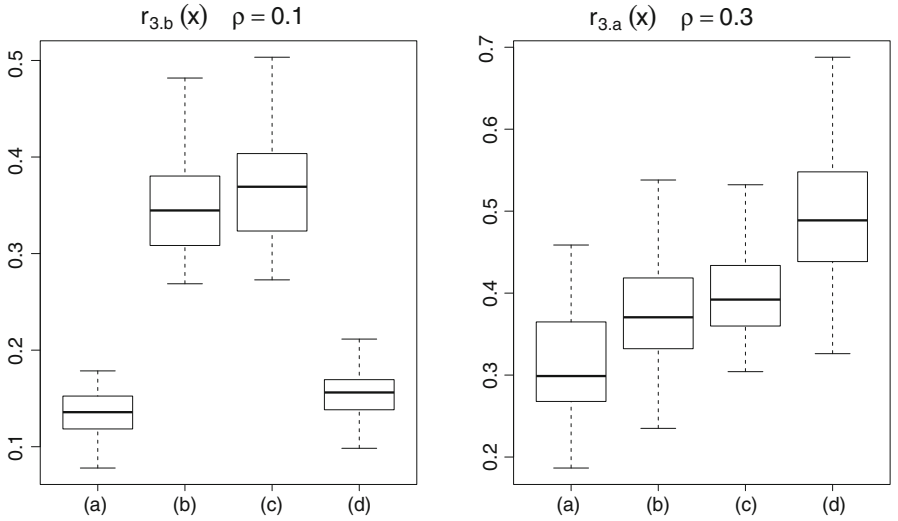

Fig. 6 RMSE for regression model involving operator $r_{3.a}$ with $\rho = 0.1$ and $\rho = 0.3$. **a** stands for PFSIM, **b** for FSIM, **c** for FLM and **d** for FNPM

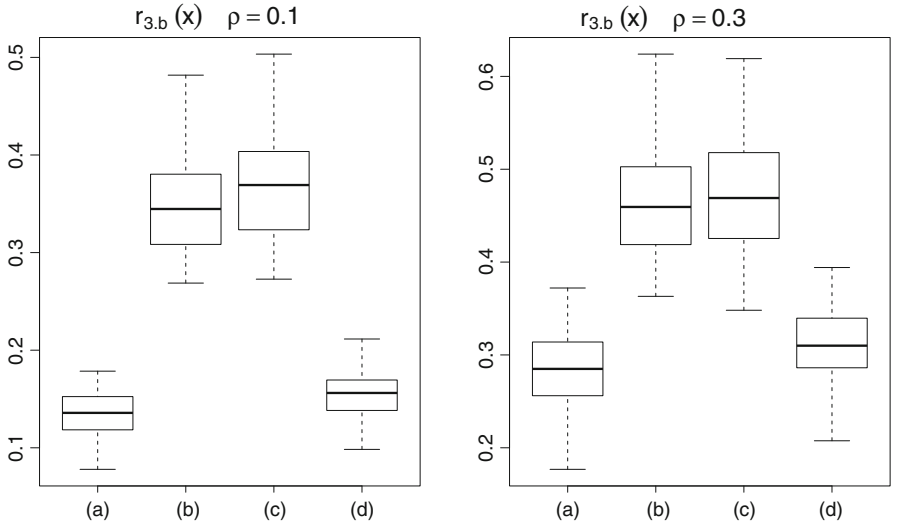

Fig. 7 RMSE for regression model involving operator $r_{3.b}$ with $\rho = 0.1$ and $\rho = 0.3$. **a** stands for PFSIM, **b** for FSIM, **c** for FLM and **d** for FNPM

3.2 Illustrating the selection of the breaking-point

To show how the selection algorithm described in Sect. 2.3 works in practice, in this section we illustrate the results of a simulation study conducted using the regression operators $r_{2.b}$ and $r_{3.a}$ defined in Sect. 3.1, where the "true" breaking-point is $\lambda_0 = 0$. Data have been generated according to (10) and the regression model (9) in the same

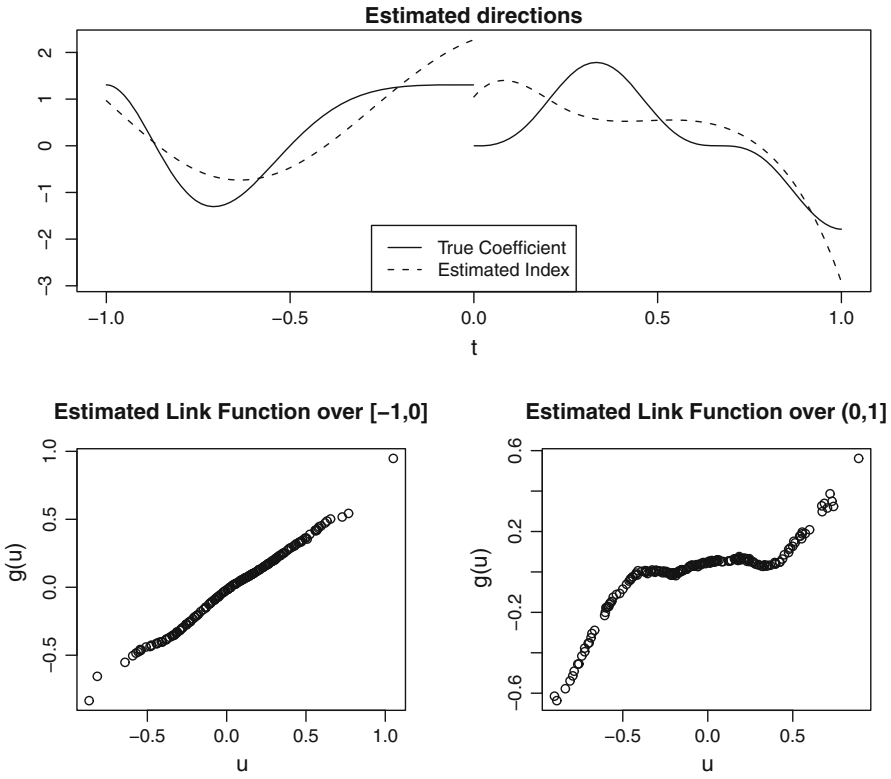

Fig. 8 Estimates of the directions θ_j (*top panel*) and link functions g_j (*bottom panels*) in the case of the regression operator $r_{2.a}$

way as in Sect. 3.1, with $\rho = 0.1$. Estimations were based on the same parameter conditions used in the previous section (cubic splines with $k_1 = k_2 = 3$, and h_j selected by a leave-one-out CV), and the search of the optimal λ is done between -0.9 and 0.9 with grid width 0.1 (i.e. $\Lambda = \{-0.9, -0.8, \ldots, 0.8, 0.9\}$). In order to evaluate the role of a misspecification of the breaking-point on the predictive abilities of the PFSIM, once $\widehat{\lambda}$ was identified, RMSEs have been computed over the test-set using both $\widehat{\lambda}$ as well as λ_0. The experiment has been replicated using 100 different random samples, of small, medium and moderately large sample sizes ($n = 50$, 100 and 200) in order to relate the identification of λ with the sample size.

Observing the smoothed distributions of the selected $\widehat{\lambda}$ varying the sample size n (see plots in Fig. 9) it emerges that, in the proposed examples, the selection method produced some estimates of the parameter λ slightly biased with a variability that decreases, as expected, with n. However, detection problems, ascribable to those raised in the end of Sect. 2.3, do not produce effects on the capacity of the PFSIM to provide good predictions, and this for the flexibility of the procedure. Indeed the box-plots in Figs. 10 and 11 show that the data-driven selected parameters $\widehat{\lambda}$ gives similar prediction errors as the true λ_0, which is unknown in practice. This fact justifies the employ of

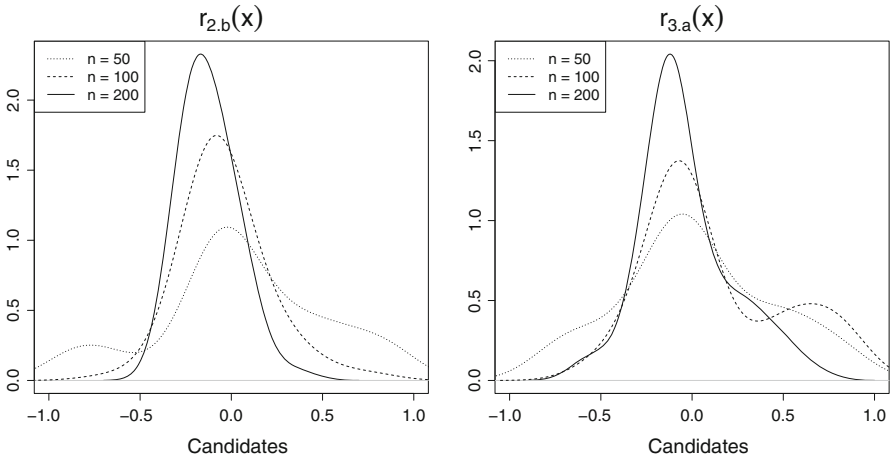

Fig. 9 Estimates of density of selected λ when one uses the regressor operators $r_{2.b}$ (*left panel*) and $r_{3.a}$ (*right panel*), varying the sample size n

the proposed cross-validation principle in applied frameworks, where the prediction aspect plays a central role.

4 Application to spectrometric datasets

To knowing the composition of an aliment, instead of relying on expensive chemical analysis, it is often preferable to obtain an estimate by spectroscopic analysis: a spectrometer measures the absorbtion of light emitted with different wavelengths by the studied substance. Absorbtion in function of the wavelength represents a functional data. During the last years, the use of various functional techniques has been widely explored for data of such nature (see for instance, Ferraty and Vieu 2002; Saeysa et al. 2008 or Ferraty et al. 2013).

In what follows, we illustrate an application of our PFSIM method in chemometric analysis: we use the well known Tecator dataset (available at lib.stat.cmu.edu/datasets/tecator): it consists in 215 spectra in the near infrared (NIR) wavelength range from 852 to 1,050 *nm*, discretized on a mesh of 100 equispaced measures, corresponding to as many finely chopped pork samples. The aim is to predict the fat content y_i, measured by chemical analysis, from the spectrometric curve. To avoid the well-known "calibration problem", due to the presence of shifts in the curves that cause noises, it is conventional to take as regressor x_i the second derivatives of spectrometric curves instead of the original ones (see Ferraty and Vieu 2002).

The regression methodology described in Sect. 2 has been applied over a learning-sample formed by the first 160 couples (x_i, y_i), and the goodness-of-fit evaluated over a test-set containing the remaining 55.

We proceeded with the selection of the breaking-point λ by using the Cross-validation procedure illustrated in Sect. 2.3 , with candidates between 860 and 1,040 with mesh width 10, so that $\Lambda = \{860, 870, \ldots, 1{,}030, 1{,}040\}$. The estimator for each

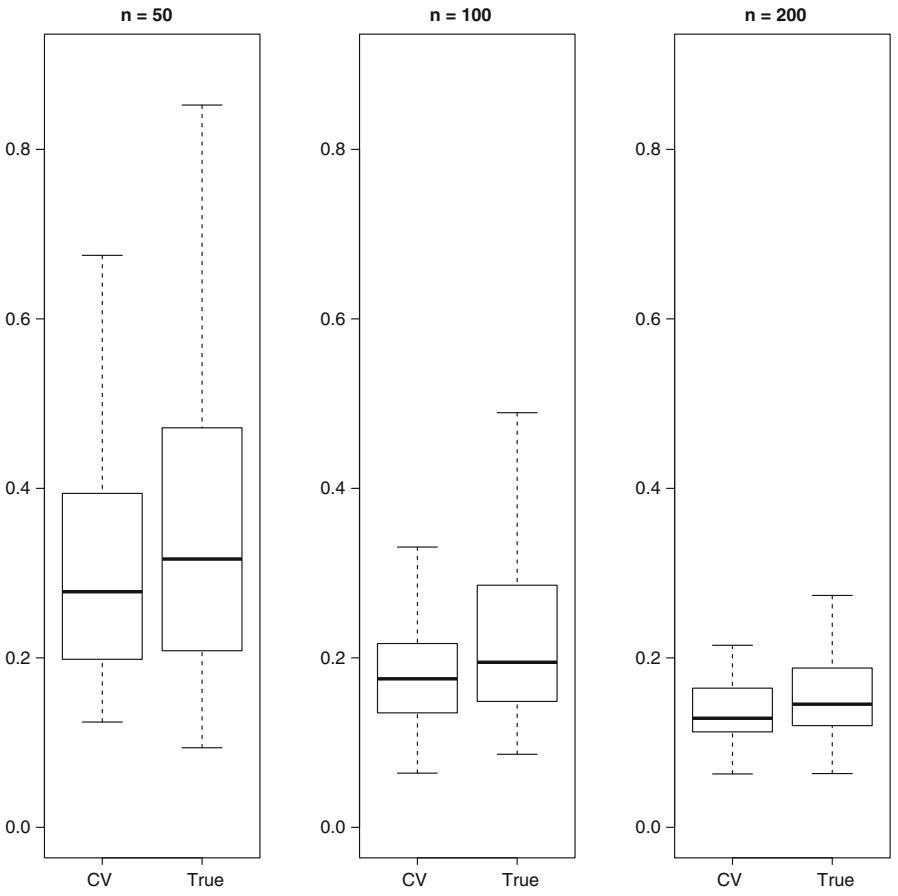

Fig. 10 Estimates of RMSEs when one uses λ and the true breaking-point for regression operator $r_{2.b}$ and varying n

$\lambda \in \Lambda$ was based on cubic splines with $k_1 = k_2 = 6$ internal knots. The minimum for $CV\ (\lambda)$ achieved at $\widehat{\lambda} = 960$ (see the left plot of Fig. 12).

Fixed the breaking-point at 960, the PSFIM applied to data by using cubic splines with $k_1 = 9$ and $k_2 = 5$ internal knots: this choice allows to improve sensibly the performances respect to take the same number of knots for either additive terms. Applying the estimated model to the testing sample, we obtained a square prediction error $MSE = \frac{1}{55}\sum_{i=1}^{55}\left(y_i^{out} - \widehat{y}_i\right)^2$ equal to 1.3916 and its relative version $RMSE$ equal to 0.00823. The out-of-sample prediction accuracy can be appreciated looking at the right panel of Fig. 12.

To provide an interpretation of estimated model, we analized the estimated additive terms. First we computed the explained variance by each component as the ratio between the empirical variance of $\widehat{g}_j\left(\int_{805}^{1050}\widehat{\theta}_j\,(t)\,x_j\,(t)\,dt\right)$ and the variance of response y_i. We obtained 0.96 for the first term and 0.02 for the second one: this

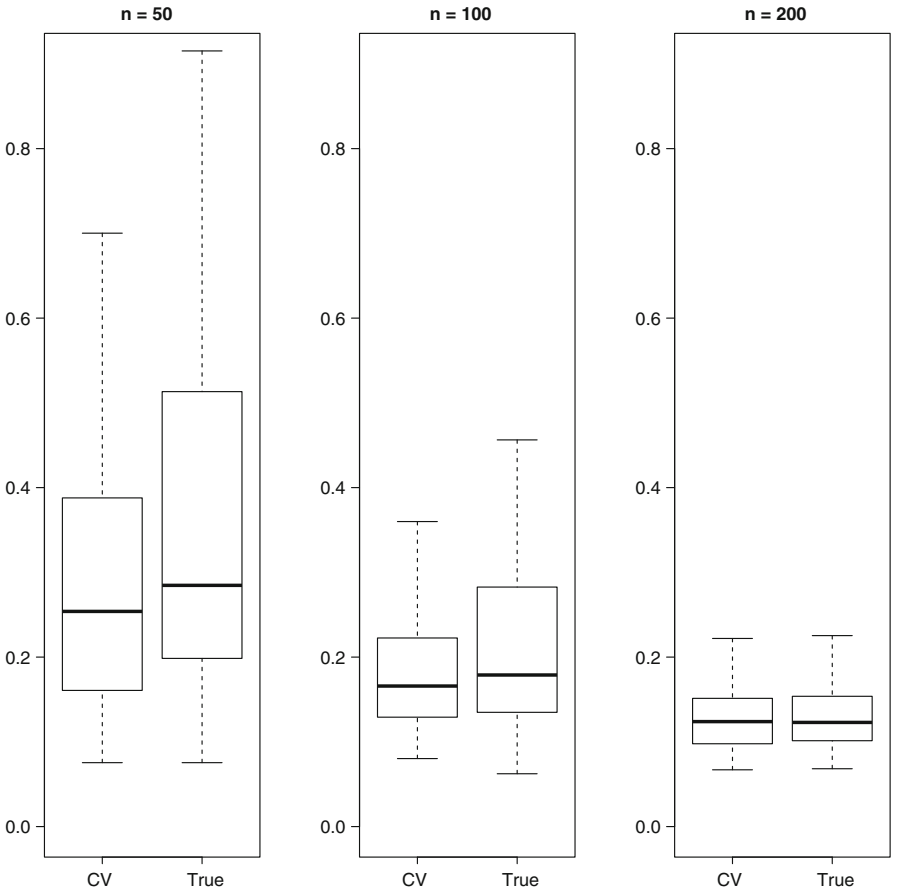

Fig. 11 Estimates of RMSEs when one uses λ and the true breaking-point for regression operator $r_{3.a}$ and varying n

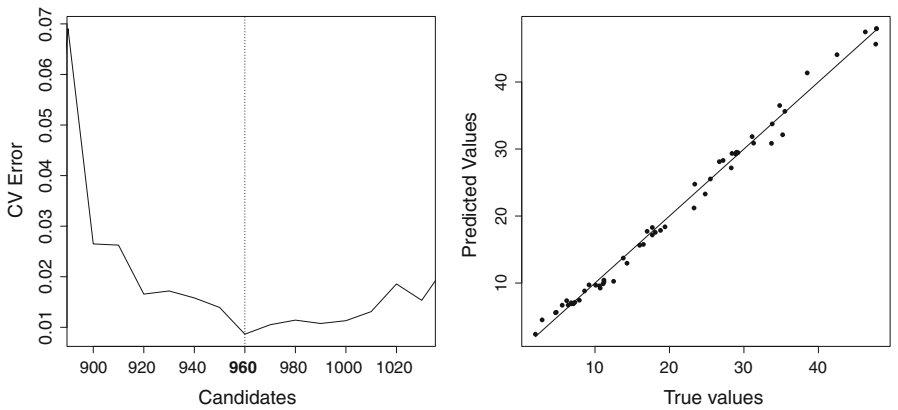

Fig. 12 Estimated λ and predicted values

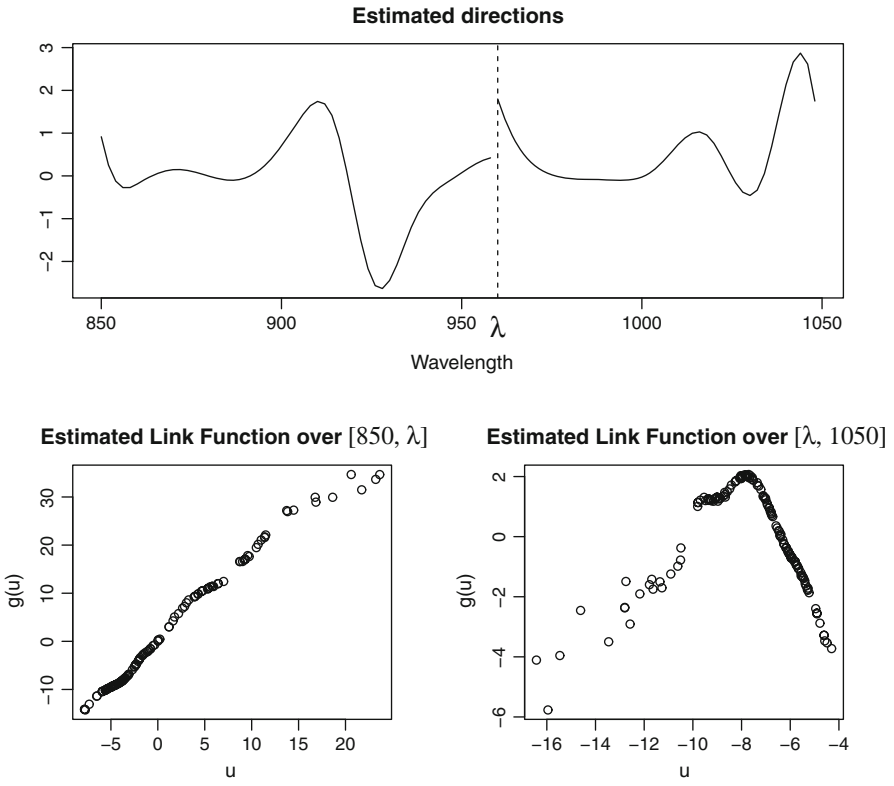

Fig. 13 Estimated directions θ_j and link functions g_j for spectrometric data

Table 1 Mean square errors MSE and relative MSE on the testing-set for PFSIM, FSIM, FLM and FNPM

Model	MSE	$RMSE \times 10^{-2}$
PFSIM	1.392	0.823
FSIM	3.704	2.191
FLM	8.439	4.992
FNPM	1.915	1.134

said us that the second part of the spectrum, corresponding to wavelengths longer than 960 *nm*, is in practice negligible in explaining the fat content. Therefore, to deepen on the nature of the link function over the relevant part of the spectrum, we look at the estimated directions plotted in Fig. 13: it appears that the wavelengths between 850 and 890 *nm* seem not relevant, whereas the ones in the range 890–950 are the most important. This is coherent with the results on selection of variables in Ferraty et al. (2010) where such interval appears the most interesting.

To conclude the analysis we compared the obtained results with ones gave by the same competitors used in Sect. 3.1. Reading the out-of-sample performances reported in Table 1 one can conclude that PFSIM is the best among the proposed techniques.

Because these data have been widely explored in literature, becoming a benchmark, it is possible to make a large comparison with a lot of methodologies. One can see, for instance, the summary Table 9 in Ferraty and Vieu (2011) and notice that our method appears to be one of the best in terms of prediction. In a nutshell, our method is of great interest on these data for exploratory purpose (see Fig. 13) but it is also one of the most powerful in terms of predictive performance (see Table 9 in Ferraty and Vieu 2011).

5 Conclusions

In this paper we have illustrated a methodology, in the framework of functional regression modeling with scalar response, which allows to approximate in a semi-parametric way the unknown regression operator through a single index approach, but taking possible structural changes into account. The novelty of the methodology consists in treating Single Index Model which can manage ruptures using unsmooth functional directions and additive link functions. In such perspective, our work can be included in the front of literature on selection variable, rather than only in the semi-parametric regression context. In that sense, our paper can be seen as taking part on the recent advances devoted to explore links between Functional Data Analysis and Variable Selection Procedures (see for instance Bongiorno et al. 2014).

An extensive simulation study, made to compare the predictive performance of the method with some classical functional regression competitor (parametric, semi-parametric and nonparametric), has pointed out the abilities of the proposed approach. Moreover it has been shown, through an application to a real benchmark data set, that the method reaches the usual goals of semi-parameric modelling in the sense that it combines good predictive power and interpretability of the outputs: indeed the results obtained are relevant and corroborated by former studies of the same data.

It should be noted that even though the implemented method rests on regressors which are curves, extensions to general functional objects, such as images and arrays, are always possible. Moreover, one can consider situations with a binary response variable.

Acknowledgments The first author thanks Enea G. Bongiorno for for valuable comments and suggestions. The second author wishes to thanks all the members of the STAPH group in Toulouse, for their long time support and comments. All the authors wish to thank the Associate Editor and two anonymous referees for their helpful remarks and suggestions which have led to substantial improvement of this paper.

Appendix: A few words on asymptotic behaviour

Cross-validation ideas have been firstly used in nonparametric framework for bandwidth selection in standard multivariate setting (see for instance Härdle and Marron 1985; Marron and Härdle 1986). Afterwards they have been extended for bandwidth selection in functional framework (see Rachdi and Vieu 2007), and more generally to other automatic parameter choices in functional data analysis, like choosing direction in functional single index modelling (see Ait-Saïdi et al. 2008), selecting the dimension

in functional projection pursuit regression (see Ferraty et al. 2013) or structural-points in complex regression models (see Ferraty et al. 2011). This is why cross-validation has been used in our work both for fitting the model in (7) and for choosing the break-point in (8).

Observing that the partitioned model (4) can be equivalently written as

$$Y = \alpha + g_1\left(\int_I \overline{\theta}_1(t)\, X(t)\, dt\right) + g_2\left(\int_I \overline{\theta}_2(t)\, X(t)\, dt\right) + \mathcal{E}$$

where

$$\overline{\theta}_j(t) = \begin{cases} \theta_j(t) & t \in I_j \\ 0 & \text{otherwise} \end{cases}$$

with $\overline{\theta}_1$ and $\overline{\theta}_2$ orthogonal by construction, PFSIM can be seen as some special case of the Functional Projection Pursuit Regression model developed recently in Ferraty et al. (2013). As a matter of conclusion, one could derive directly asymptotic results for the method proposed here just by straightforward adaptation of the proofs in the above mentioned paper. In particular one could get the following kinds of results:

i. Asymptotic optimality (in terms of minimal quadratic prediction error) of the estimates of the directions θ_j obtained in (7), just from the proof of Theorem 5 in Ferraty et al. (2013);
ii. Univariate rate of convergence of the estimates of the link functions g_j, just from the proof of Theorem 4 in Ferraty et al. (2013).

In the same spirit, by following the general methodology as presented in Ferraty et al. (2011) for structural parameter estimation, one could also get:

iii. Consistency of the selected parameter $\widehat{\lambda}$ towards the value $\lambda_0 \in \Lambda$ minimizing quadratic prediction errors;
iv. Asymptotic optimality (inside of Λ) of the data-driven value $\widehat{\lambda}$.

References

Ait-Saïdi A, Ferraty F, Kassa R, Vieu P (2008) Cross-validated estimations in the single-functional index model. Statistics 42:475–94

Amato U, Antoniadis A, De Feis I (2006) Dimension reduction in functional regression with applications. Comput Stat Data Anal 50:2422–2446

Aneiros G, Ferraty F, Vieu P (2011) Variable selection in semi-functional regression models. Recent advances in functional data analysis and related topics. Contrib Statist, Physica-Verlag, Heidelberg, pp 17–22

Bongiorno EG, Salinelli E, Goia A, Vieu P (eds) (2014) Contributions in infinite-dimensional statistics and related topics. Società editrice Esculapio, Bologna

Burnham KP, Anderson DR (2002) Model selection and multimodel inference, 2nd edn. Springer, New York

Cardot H, Ferraty F, Sarda P (2003) Spline estimators for the functional linear model. Stat Sinica 13:571–591

Cardot H, Goia A, Sarda P (2004) Testing for no effect in functional linear regression models, some computational approaches. Comm Stat Simul Comput 33:179–199

Delsol L, Ferraty F, Vieu P (2011) Structural test in regression on functional variables. J Multivar Anal 102:422–447

Delsol L (2013) No effect tests in regression on functional variable and some applications to spectrometric studies. Comput Stat 28:1775–1811
Ferraty F, Goia A, Salinelli E, Vieu P (2013) Functional projection pursuit regression. Test 22:293–320
Ferraty F, Hall P, Vieu P (2010) Most-predictive design points for functional data predictors. Biometrika 97:807–824
Ferraty F, Martinez Calvo A, Vieu P (2011) Thresholding in nonparametric functional regression with scalar response. Recent advances in functional data analysis and related topics. Contrib Statist, Physica-Verlag, Heidelberg, pp 103–109
Ferraty F, Park J, Vieu P (2011) Estimation of a functional single index model. Recent advances in functional data analysis and related topics. Contrib Statist, Physica-Verlag, Heidelberg, pp 111–116
Ferraty F, Peuch A, Vieu P (2003) Modèle à indice fonctionnel simple. Comptes Rendus Math Académie Sci Paris 336:1025–1028
Ferraty F, Vieu P (2002) The functional nonparametric model and applications to spectrometric data. Comput Stat 17:545–564
Ferraty F, Vieu P (2006) Nonparametric functional data analysis. Springer, New York
Ferraty F, Vieu P (2011) Richesse et complexité des données fonctionnelles. Revue Modulad 43:25–43
Härdle W, Hall P, Ichimura H (1993) Optimal smoothing in single-index models. Ann Stat 21:157–178
Härdle W, Marron JS (1985) Optimal bandwidth selection in nonparametric regression function estimation. Ann Stat 13:1465–1481
Härdle W, Müller N, Sperlich S, Werwatz A (2004) Nonparametric and semiparametric models. Springer Series in Statistics. Springer, New York
Härdle W, Stoker TM (1989) Investigating smooth multiple regression by the method of average derivatives. J Am Stat Assoc 84:986–995
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer Series in Statistics
Horváth L, Reeder R (2012) Detecting changes in functional linear models. J Multivar Anal 111:310–334
Horváth L, Kokoszka P (2012) Inference for functional data with applications. Springer, New York
James G (2002) Generalized linear models with functional predictors. J R Stat Soc B 64:411–432
Leardi R (ed) (2003) Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks. Elsevier, Amsterdam
Marron JS, Härdle W (1986) Random approximations to some measures of accuracy in nonparametric curve estimation. J Multivar Anal 20:91–113
McKeague IW, Sen B (2010) Fractals with point impact in functional linear regression. Ann Stat 38:2559–258
Nelder JA, Mead R (1965) A simplex algorithm for function minimization. Comput J 7:308–313
Rachdi M, Vieu P (2007) Nonparametric regression for functional data: automatic smoothing parameter selection. J Stat Plan Inference 137:2784–2801
Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, New York
Saeysa W, De Ketelaerea B, Darius P (2008) Potential applications of functional data analysis in chemometrics. J Chemometr 22:335–344
Xia Y, Härdle W (2006) Semi-parametric estimation of partially linear single-index models. J Multivar Anal 97:1162–1184