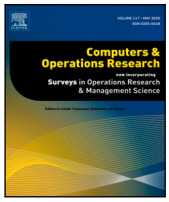
Computers & Operations Research 152 (2023) 106152

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor



On optimal regression trees to detect critical intervals for multivariate functional data

Rafael Blanquero [a], Emilio Carrizosa [a], Cristina Molero-Río [a,*], Dolores Romero Morales [b]

[a] Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Seville, Spain
[b] Copenhagen Business School (CBS), Frederiskberg, Denmark

ARTICLE INFO

Keywords:
Optimal randomized regression trees
Multivariate functional data
Critical intervals detection
Nonlinear programming

ABSTRACT

In this paper, we tailor optimal randomized regression trees to handle multivariate functional data. A compromise between prediction accuracy and sparsity is sought. Whilst fitting the tree model, the detection of a reduced number of intervals that are critical for prediction, as well as the control of their length, is performed. Local and global sparsities can be modeled through the inclusion of LASSO-type regularization terms over the coefficients associated to functional predictor variables. The resulting optimization problem is formulated as a nonlinear continuous and smooth model with linear constraints. The numerical experience reported shows that our approach is competitive against benchmark procedures, being also able to trade off prediction accuracy and sparsity.

1. Introduction

Decision trees (Loh, 2014) are state-of-the-art classification and regression methods consisting of recursively partitioning the sample with splits based on conditions imposed over the predictor variables of the model. They show excellent learning performance, are conceptually simple, appealing in terms of interpretability (Freitas, 2014; Goodman and Flaxman, 2017; Hu et al., 2019; Lin et al., 2020; Meinshausen, 2010) because of their rule-based nature, computationally cheap, and routines and packages to train them are available in popular languages such as Python and R. These characteristics make them suitable in many applications ranging from policy making to health care. For example, they were recently useful to evaluate housing systems for homeless youth (Chan et al., 2018), to forecast demand at a big online retailer (Januschowski et al., 2022), to predict the evolution of COVID-19 (Benítez-Peña et al., 2021), and to diagnose chest pathologies and predict length-of-stay and 48 h mortality (Soenksen et al., 2022). See, for instance, Aghaei et al. (2022) for more applications.

In recent times, and thanks to the availability of more powerful hardware and the dramatic advances in optimization solvers (Bixby, 2012), there has been an increased interest by the Mathematical Optimization community to develop novel approaches to build optimal decision trees, where the traditional greedy splitting procedure is replaced by a global strategy in which all the splits along the tree are decided in one go. These new paradigms include Mixed-Integer Linear Optimization (Aghaei et al., 2022; Ahuja et al., 1993; Bertsimas and Dunn, 2017; Dunn, 2018; Firat et al., 2019; Günlük et al., 2021; Verwer and Zhang, 2017; Verwer et al., 2017, 2019; Zantedeschi et al., 2020; Zhu et al., 2020), Continuous Optimization (Blanquero et al., 2020b, 2021a, 2022), Constraint Programming (Verhaeghe et al., 2019), SAT (Narodytska et al., 2018; Yu et al., 2020) and Dynamic Programming (Demirović et al., 2022) approaches, as reviewed in Carrizosa et al. (2021a).

Optimal decision trees have focused on the analysis of multivariate data. Nevertheless, data may present other kinds of complexities coming from the analysis of images (Wang et al., 2021) or texts (Ramon et al., 2020), the time- (Saha et al., 2021) and spatial-dependence (Georganos et al., 2021) of the data, as well as the hierarchical, relational and network (Óskarsdóttir et al., 2022) structure of the data, to name a few. The success of optimal decision trees when dealing with multivariate data makes it promising to tailor them to such complexities. In this work, we focus on the analysis of multivariate functional data. Functional Data Analysis (Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012; Ramsay and Silverman, 2002, 2005) is the field of Statistics that extends the classic multivariate analysis to handle observations of functional nature, and has applications in areas such as biomedicine (Leng and Müller, 2005), chemistry (Blanquero et al., 2016), health sciences (Strzalkowska-Kominiak and Romo, 2021), meteorology (James et al., 2009) and econometrics (Laukaitis and Račkauskas, 2005). The reader is referred to Cuevas (2014), Goia and Vieu (2016), Wang et al. (2016) for reviews and a special issue on the topic.

* Corresponding author.
E-mail addresses: rblanquero@us.es (R. Blanquero), ecarrizosa@us.es (E. Carrizosa), mmolero@us.es (C. Molero-Río), drm.eco@cbs.dk (D. Romero Morales).

https://doi.org/10.1016/j.cor.2023.106152
Received 22 May 2022; Received in revised form 3 November 2022; Accepted 10 January 2023
Available online 13 January 2023

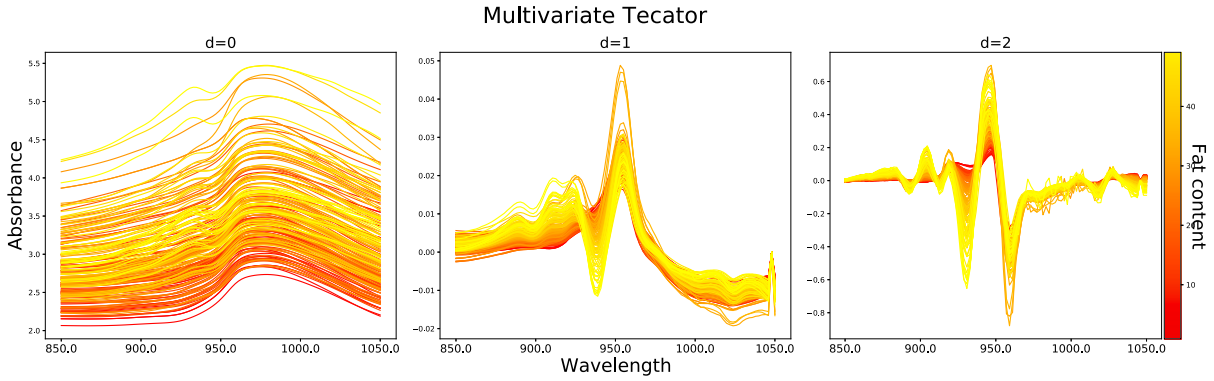


Fig. 1. Graphical representation of the multivariate Tecator dataset with $p = 3$ functional predictor variables: the raw data series ($d = 0$), the series of the first derivative ($d = 1$) and the series of the second derivative ($d = 2$). Each observation in the dataset is colored according to the response variable. The higher the value of the response variable, the colder the color.

Although functional data is usually associated with time dependence, one can think of features of very different nature such as the wavelength in a spectrometry, as in the well-known Tecator data set (Febrero-Bande and de la Fuente, 2012) which will be used below to illustrate our contribution.

Functional data may appear as predictor variables, response variables or both simultaneously. Modeling the relationship between a scalar response and functional predictor variables (Berrendero et al., 2018; Cai and Hall, 2006; Fan et al., 2015) through an optimal decision tree is the setting of this work. Such is the case of Tecator data set, which consists of the near-infrared absorbance spectra of samples of finely chopped pork, where the aim is to predict the percentage of its fat content. In Fig. 1, the multivariate version of the Tecator data set is illustrated.

In principle, optimal decision trees, as any other standard multivariate approach, could address the analysis of functional data after discretizing the functions and converting them to vectors. Yet, in general, the direct use of such techniques for functional data may have dramatic consequences (Horváth and Kokoszka, 2012; Jiménez Cordero, 2019), for instance, the curse of dimensionality since lots of points are needed to summarize the information in a curve (Vieu, 2018). Furthermore, the intrinsic characteristics of functional data may not be fully exploited (Borggaard and Thodberg, 1992): multivariate approaches ignore the ordering and the spacing of a set of data values (Griswold et al., 2008), and the strong correlation between measurements in two consecutive instants is not taken into account, potentially yielding multicollinearity problems (Hastie et al., 1995). These reasons motivate the development of models that take advantage of the functional nature directly (Balakrishnan and Madigan, 2006; Belli and Vantini, 2021).

In this paper, we propose a Continuous Optimization approach, the Sparse Optimal Randomized Regression Tree for multivariate Functional Data (S-ORRT-FD), based on the methodology in Blanquero et al. (2022). This methodology was shown to be competitive in terms of prediction accuracy against benchmarks, as well as scalable with respect to the size of the training sample, and flexible enough to incorporate desirable properties in addition to sparsity (Blanquero et al., 2020b, 2022), such as fairness (Carrizosa et al., 2021b; Zafar et al., 2017), by aiming to avoid predictions that discriminate against sensitive features; the cost-sensitivity (Blanquero et al., 2021a; Benítez-Peña et al., 2019; Blanquero et al., 2021b) for groups of individuals in which prediction errors are more critical, by ensuring an acceptable accuracy performance for them; or local explainability (Blanquero et al., 2022; Carrizosa et al., 2021a; Ribeiro et al., 2016), where the goal is to identify the predictor variables that have the largest impact on the individual predictions. The reader is referred to Carrizosa et al. (2021a) for more details on these properties.

In this framework, it may happen that simply using a finite number of instants (Aneiros and Vieu, 2014, 2016; Berrendero et al., 2019; Kong et al., 2016) and/or intervals (Blanquero et al., 2020a; Grollemund et al., 2019) in the domain of the functional predictor variables is sufficient to produce accurate predictions. S-ORRT-FD controls explicitly the number of intervals that are critical for prediction, as well as their length, and assigns to each of them a single coefficient per functional variable that is constant for the whole interval. This has, in turn, the advantage of being more interpretable and saves both monitoring and storage costs. In the case of degenerate intervals with length equal to zero, critical instants would be detected instead. Fig. 2 illustrates the detection of ten critical intervals addressed by S-ORRT-FD for the multivariate Tecator data set. A sparse solution in terms of length sparsity can be observed in the sense that one would not need to know the values of the curves in the whole domain in order to make accurate predictions.

Furthermore, we can control local and global sparsities (Febrero-Bande et al., 2019), that is, the number of coefficients and the number of functional predictor variables being used along the tree, respectively.

The paper is organized as follows. In Section 2, we introduce the S-ORRT-FD and its nonlinear mathematical optimization formulation. Our computational experience is reported in Section 3. We illustrate that S-ORRT-FD is competitive with state-of-the-art regression benchmarks. Moreover, we show our ability to trade off prediction accuracy and sparsity, in the form of controlling the number of critical intervals and the proportion of the curves to be used for prediction. Finally, conclusions and possible lines of future research are provided in Section 4.

2. Sparse optimal randomized regression trees for multivariate functional data

2.1. Introduction

Let $\mathcal{I}$ be a given set of N individuals. Each individual $i \in \mathcal{I}$ is represented by a pair $(\boldsymbol{x}_i, y_i)$. The first element $\boldsymbol{x}_i \in \mathcal{F}^p$ represents the multivariate functional data, composed by p functional predictor variables, i.e., $\boldsymbol{x}_i = \left(x_{i1}(\cdot), \ldots, x_{ip}(\cdot)\right)$, where $x_{ij}(\cdot) : [\underline{s}, \overline{s}] \to \mathbb{R}$, $j = 1, \ldots, p$, belongs to the set $\mathcal{F}$ of Riemann integrable functions in the interval $[\underline{s}, \overline{s}]$. Note that numerical predictor variables can also be included in our framework, by simply defining them as constant functions. The second element of the pair, $y_i \in \mathbb{R}$, indicates the value of the response variable.

With the Sparse Optimal Randomized Regression Trees for multivariate Functional Data (S-ORRT-FD), addressed in this paper, we extend the approach introduced in Blanquero et al. (2020b, 2021a, 2022), Carrizosa et al. (2021a) to additionally handle multivariate functional data. An optimal binary regression tree of a given depth D is to be built, obtained by controlling simultaneously prediction accuracy

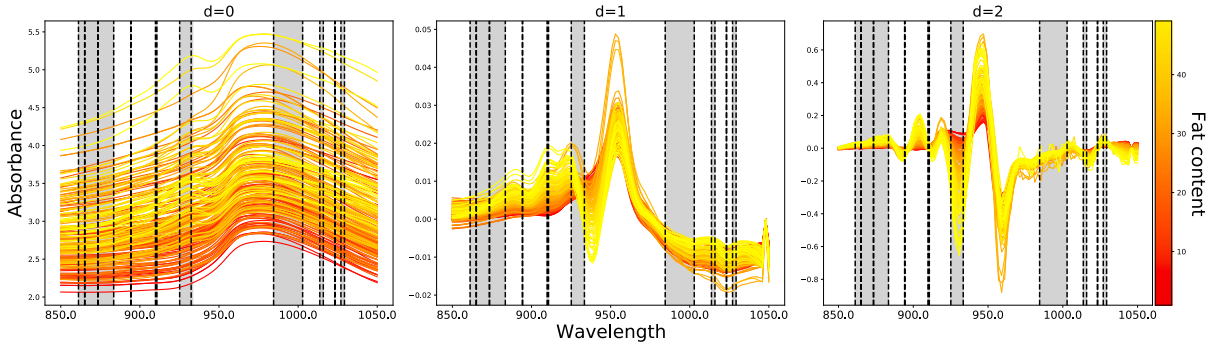


Fig. 2. Detection of ten critical intervals addressed by S-ORRT-FD for the multivariate Tecator data set.

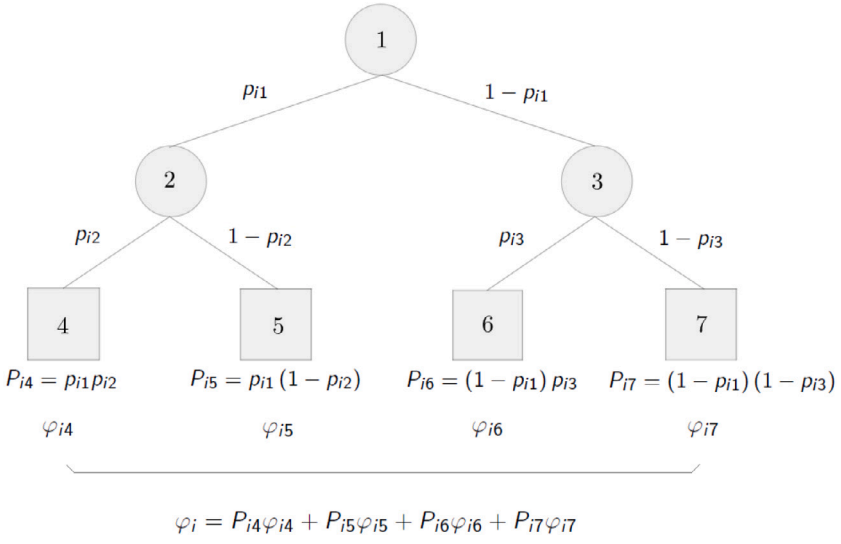


Fig. 3. Sparse optimal randomized regression tree of depth $D = 2$ for multivariate functional data with $\{p_{it}\}_{t=1,2,3}$, $\{P_{it}\}_{t=4,5,6,7}$ and $\{\varphi_{it}\}_{t=4,5,6,7}$ defined in Eqs. (2), (3) and (4), respectively.

and some kind of sparsity. Fig. 3 shows the structure of an S-ORRT-FD of depth $D = 2$.

S-ORRT-FDs are modeled by means of a Non-Linear Optimization (NLO) formulation. Oblique cuts are implemented at the set of branch nodes τ_B. Linear predictions are associated to the set of leaf nodes τ_L. The usual deterministic yes/no rule at each branch node is replaced by a probabilistic decision rule, induced by a univariate continuously differentiable cumulative density function (CDF) F, evaluated over the vector of functional predictor variables. With this, we have the probability of each individual in the sample falling into every leaf node, that will represent the weights that the linear predictions will have in the estimated outcome value.

In order to model oblique splits at branch nodes $t \in \tau_B$, we need to define, for each functional predictor variable $j = 1, \dots, p$ and each $t \in \tau_B$, the functional decision variables $a_{jt}(\cdot) : [\underline{s}, \overline{s}] \to \mathbb{R}$ that denote the coefficient functions, as well as the decision variables μ_t, $t \in \tau_B$ as intercepts. Then, the probability of individual $i = 1, \dots, N$ going down the left branch at branch node $t \in \tau_B$ would be defined by:

$$F\left(\frac{1}{p}\sum_{j=1}^{p}\int_{\underline{s}}^{\overline{s}} a_{jt}(s)x_{ij}(s)ds \; - \; \mu_t\right). \tag{1}$$

With the aim of detecting critical intervals for prediction, we will assume that $a_{jt}(\cdot), j = 1, \dots, p,\ t \in \tau_B$, are piecewise constant functions. Let H denote the number of pieces where constants are different from zero, or critical intervals, and $a_{jth} \in \mathbb{R}, j = 1, \dots, p, t \in \tau_B, h = 1, \dots, H$, the decision variables representing the coefficient of functional variable j at branch node t in the critical interval h. New decision variables are to be defined to represent the lower and upper ends of such intervals, namely $[s_{2h-1}, s_{2h}], h = 1, \dots, H$, where s_h, $h = 1, \dots, 2H$, should belong to the interval $[\underline{s}, \overline{s}]$. In this way, each functional decision variable turns into

$$a_{jt}(s) = \begin{cases} \dfrac{a_{jth}}{s_{2h} - s_{2h-1}}, & \text{if } s \in [s_{2h-1}, s_{2h}], h = 1, \dots, H, \\ 0, & \text{otherwise} \end{cases},$$

where coefficients a_{jth} have been scaled according to the length of their corresponding critical interval. See Fig. 4 for an example of a piecewise constant function with $H = 3$ critical intervals.

We have to add the constraints $s_h \le s_{h+1}$, $j = 1, \dots, p$, $h = 1, \dots, 2H - 1$, to ensure that intervals are ordered and do not overlap.

Let $\boldsymbol{a}$, $\boldsymbol{\mu}$ and $\boldsymbol{s}$ denote the $p \times |\tau_B| \times H$-matrix, the $|\tau_B|$-vector and the $2H$-vector of the coefficients $\boldsymbol{a} = (a_{jth})_{j=1,\dots,p,\ t\in\tau_B,\ h=1,\dots,H}$, $\boldsymbol{\mu} = (\mu_t)_{t\in\tau_B}$, and $\boldsymbol{s} = (s_h)_{h=1,\dots,2H}$, respectively. Then, the probability of individual $i = 1, \dots, N$ going down the left branch at branch node $t \in \tau_B$, defined in Eq. (1), turns into:

$$p_{it}\left(\boldsymbol{a}_{\cdot t\cdot}, \mu_t, \boldsymbol{s}\right) = F\left(\frac{1}{p}\sum_{j=1}^{p}\sum_{h=1}^{H}\frac{a_{jth}}{s_{2h} - s_{2h-1}}\int_{s_{2h-1}}^{s_{2h}} x_{ij}(s)ds \; - \; \mu_t\right). \tag{2}$$

The expression $\boldsymbol{a}_{\cdot t\cdot}$ denotes the $p \times H$-matrix of the coefficients in $\boldsymbol{a}$ related to branch node t.

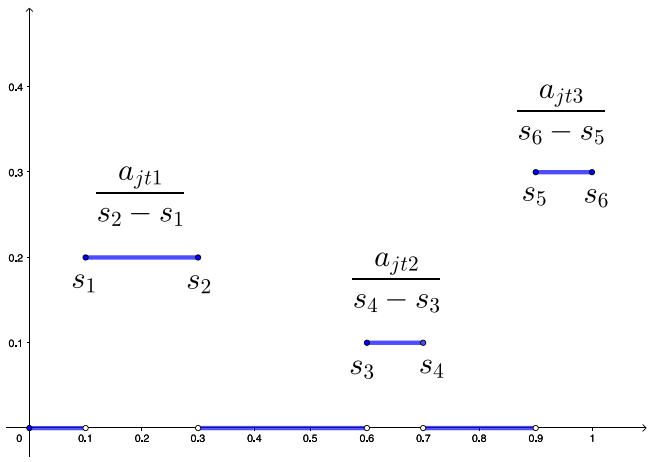


Fig. 4. Example of a piecewise constant function $a_{jt}(\cdot)$ with $H = 3$ critical intervals and domain $[0, 1]$.

The probability of individual $i = 1, \dots, N$ falling into leaf node $t \in \tau_L$ is:

$$P_{it}(\boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{s}) = \prod_{t_l \in \mathcal{N}_L(t)} p_{it_l}\left(\boldsymbol{a}_{\cdot t_l \cdot}, \mu_{t_l}, \boldsymbol{s}\right) \prod_{t_r \in \mathcal{N}_R(t)} \left(1 - p_{it_r}\left(\boldsymbol{a}_{\cdot t_r \cdot}, \mu_{t_r}, \boldsymbol{s}\right)\right), \quad (3)$$

where $N_L(t)$ and $N_R(t)$ are the sets of ancestor nodes of leaf node t whose left and right branch, respectively, takes part in the path from the root node to leaf node t, $t \in \tau_L$.

Similarly to oblique cuts, real decision variables $\tilde{\boldsymbol{a}} = \left(\tilde{a}_{jth}\right)_{j=1,\dots,p,\ t\in\tau_L,\ h=1,\dots,H}$ and $\tilde{\boldsymbol{\mu}} = \left(\tilde{\mu}_t\right)_{t\in\tau_L}$ are to be defined in order to provide linear predictions at leaf nodes $t \in \tau_L$ for each individual $i = 1, \dots, N$:

$$\varphi_{it}\left(\tilde{\boldsymbol{a}}_{\cdot t \cdot}, \tilde{\mu}_t, \boldsymbol{s}\right) = \sum_{j=1}^{p} \sum_{h=1}^{H} \frac{\tilde{a}_{jth}}{s_{2h} - s_{2h-1}} \int_{s_{2h-1}}^{s_{2h}} x_{ij}(s) ds \; - \; \tilde{\mu}_t, \quad (4)$$

where coefficients $\tilde{a}_{jth}$ have been scaled according to the length of their corresponding critical interval. The expression $\tilde{\boldsymbol{a}}_{\cdot t \cdot}$ denotes the $p \times H$-matrix of the coefficients in $\tilde{\boldsymbol{a}}$ related to leaf node t.

2.2. The formulation

With these parameters and decision variables, the S-ORRT-FD reads as the following NLO problem with linear constraints:

$$\min_{\substack{(\boldsymbol{a},\boldsymbol{\mu})\in\mathbb{R}^{(pH+1)|\tau_B|} \\ (\tilde{\boldsymbol{a}},\tilde{\boldsymbol{\mu}})\in\mathbb{R}^{(pH+1)|\tau_L|} \\ \boldsymbol{s}\in[\underline{s},\bar{s}]^{2H}}} \quad \frac{1}{N}\sum_{i=1}^{N}\Big(\sum_{t\in\tau_L} P_{it}(\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{s})\,\varphi_{it}\left(\tilde{\boldsymbol{a}}_{\cdot t\cdot}, \tilde{\mu}_t, \boldsymbol{s}\right) - y_i\Big)^2 \quad (5)$$

$$+ \; \lambda^G \sum_{j=1}^{p} \left\|\left(\boldsymbol{a}_{j\cdot\cdot}, \tilde{\boldsymbol{a}}_{j\cdot\cdot}\right)\right\|_\infty \quad (6)$$

$$+ \; \lambda^{\text{length}} \sum_{h=1}^{H} \left(s_{2h} - s_{2h-1}\right) \quad (7)$$

$$\text{s.t.} \quad s_h \le s_{h+1}, \; h = 1, \dots, 2H-1, \quad (8)$$

where the expressions $\boldsymbol{a}_{j\cdot\cdot}$ and $\tilde{\boldsymbol{a}}_{j\cdot\cdot}$ denote the $|\tau_B| \times H$- and the $|\tau_L| \times H$-matrices of the coefficients in $\boldsymbol{a}$ and $\tilde{\boldsymbol{a}}$ relating to functional predictor variable j, respectively.

The first term, prediction accuracy, is equal to the mean squared error (MSE) over the training sample between the actual response values and the predictions returned by S-ORRT-FD, which are weighted averages of the linear predictions along the different leaf nodes, where the weights in such average depend on the probability of individual i falling into such leaf node.

The second term, parametrized by λ^G, addresses global sparsity, which is modeled by the inclusion of a penalization term that controls whether a given functional predictor variable is ever used across the whole tree. Recall that each functional predictor variable may appear at either branch (in the oblique cuts) and leaf (in the linear predictions) nodes. Then, the ℓ_∞-norm is used as a group penalty function, by forcing all the coefficients linked to the same functional predictor variable to be shrunk simultaneously along all branch and leaf nodes. Local sparsity can also be controlled by penalizing the ℓ_1-norm of such coefficients instead, as done in previous works of the authors (Blanquero et al., 2020b, 2022).

It may occur that several functional predictor variables are related; for instance, one functional characteristic could appear together with its derivatives. In such cases, the corresponding ℓ_∞-norm term will comprise the coefficients of all of these series in order to force that all of them shrink to zero at the same time. The incorporation of the derivatives may be decisive to enhance prediction accuracy, as illustrated in Section 3.3.

The third term, parametrized by λ^{length}, controls the proportion of the curves to be used in the prediction returned by S-ORRT-FD. This is done by penalizing the length of the critical intervals. Note that if the length of a critical interval $[s_{2h-1}, s_{2h}]$ tends to zero, a critical instant would be detected instead, thanks to the integral form of the mean value theorem. This situation occurs for high enough values of λ^{length}.

2.3. A smooth reformulation

Problem (6)–(8) is non-smooth due to the ℓ_∞-norm appearing in the objective function. Recall that F is assumed to be continuously differentiable, therefore the MSE inherits smoothness. By rewriting this regularization term using new decision variables, we can formulate S-ORRT-FD as a smooth problem, thus solvable with standard continuous optimization solvers, as done in our computational section. Let $\boldsymbol{\beta} = \left(\beta_j\right)_{j=1,\dots,p}$, the regularization term of Problem (6)–(8) can be rewritten as follows:

$$\left\|\left(\boldsymbol{a}_{j\cdot\cdot}, \tilde{\boldsymbol{a}}_{j\cdot\cdot}\right)\right\|_\infty = \max\left(\left\{\left|a_{jth}\right|\right\}_{\substack{t\in\tau_B \\ h=1,\dots,H}} \cup \left\{\left|\tilde{a}_{jth}\right|\right\}_{\substack{t\in\tau_L \\ h=1,\dots,H}}\right) = \beta_j, \; j = 1, \dots, p,$$

where $\beta_j \ge 0$. We also need to impose $\beta_j \ge \pm\, a_{jt}$, $j = 1, \dots, p$, $t \in \tau_B$, and $\beta_j \ge \pm\, \tilde{a}_{jt}$, $j = 1, \dots, p$, $t \in \tau_L$. Hence, we have that Problem (6)–(8) is equivalent to the following smooth reformulation:

$$\min_{\substack{(\boldsymbol{a},\boldsymbol{\mu})\in\mathbb{R}^{(pH+1)|\tau_B|} \\ (\tilde{\boldsymbol{a}},\tilde{\boldsymbol{\mu}})\in\mathbb{R}^{(pH+1)|\tau_L|} \\ \boldsymbol{s}\in[\underline{s},\bar{s}]^{2H},\ \boldsymbol{\beta}\in\mathbb{R}^p}} \quad \frac{1}{N}\sum_{i=1}^{N}\Big(\sum_{t\in\tau_L} P_{it}(\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{s})\,\varphi_{it}\left(\tilde{\boldsymbol{a}}_t, \tilde{\mu}_t, \boldsymbol{s}\right) - y_i\Big)^2 \quad (9)$$

$$+ \; \lambda^G \sum_{j=1}^{p} \beta_j \quad (10)$$

$$+ \; \lambda^{\text{length}} \sum_{h=1}^{H} \left(s_{2h} - s_{2h-1}\right) \quad (11)$$

$$\text{s.t.} \quad s_h \le s_{h+1}, \; h = 1, \dots, 2H-1, \quad (12)$$

$$\beta_j \ge \pm a_{jth}, j = 1, \dots, p, t \in \tau_B, h = 1, \dots, H, \quad (13)$$

$$\beta_j \ge \pm \tilde{a}_{jth}, j = 1, \dots, p, t \in \tau_L, h = 1, \dots, H. \quad (14)$$

Several benefits derive from the above formulation. First, the size of the feasible region, that is, the number of decision variables as well as the number of constraints, is independent of the training sample size N and, therefore, our approach scales up when N grows. This is in line with the scalability experiments performed by the authors on tabular data in Blanquero et al. (2022). Second, once chosen the cumulative density function and provided the penalization parameters by the user, our model is parameter free. One would need to solve one single optimization problem, thus avoiding the highly costly parameter tuning present in other strategies such as SVR-FD (Blanquero et al., 2020a), against which we will compare in our computational experience. Third,

Table 1
Information about the functional data sets considered.

Data set	N	$\#points$	$[\underline{s}, \overline{s}]$	Source
Tecator	215	100	[850, 1050]	Febrero-Bande and de la Fuente (2012)
Sunflower	111	309	[0, 1]	Picheny et al. (2019)
Sugar	268	571	[275, 560]	Aneiros and Vieu (2014)
FHV	1500	100	$[0, 2\pi]$	Ferraty et al. (2010)

it is straightforward to include additional constraints that are able to model desirable properties as those mentioned in Section 1, in the same way that it was done in previous works from the authors (Blanquero et al., 2021a, 2022).

3. Computational experiments

The aim of this section is to illustrate the performance of our sparse optimal randomized regression trees for multivariate functional data. Section 3.1 gives details on the procedure followed to test our approach. In Section 3.2 we discuss the prediction accuracy of our approach, against several benchmark regression methods. Finally, in Section 3.3 we illustrate our ability to trade in some of our prediction accuracy for a gain in sparsity in the proportion of the curves to be used.

3.1. Setup

Well-known publicly available functional data sets have been chosen for the computational experiments. Table 1 lists their names together with their number of observations, the number of points where the evaluations of the original functions are known and the domain, as well as the source where they can be downloaded. All the data sets are univariate, that is, $p = 1$, and coming from real-world applications, except for FHV, which is simulated. More specifically, Tecator consists of the near-infrared absorbance spectra of 215 samples of finely chopped pork, measured at 100 wavelength points discretized from 850 to 1050 mm. The response variable is the percentage of fat content. Sunflower consists of a set of 111 climate evapotranspiration daily recordings of sunflower cultivations at 309 time points. The response variable is the annual grain yield in tons per hectare. Sugar consists of the fluorescence absorbance spectra of 268 samples of sugar, measured at 571 wavelength points discretized from 275 to 560 mm. The response variable is the percentage of ash content. FHV consists of 1500 curves with shape $x_i(s) = \sum_{l=1}^{3} U_{il} \cos[(3+l)\,s] + \sum_{l=1}^{3} V_{il} \sin[(4+l)\,s] + W_i\,(s-\pi)^2, i = 1, \ldots, 1500$, discretized in 100 points in the interval $[0, 2\pi]$, where $U_{il}, V_{il} \sim U([0,1]), l = 1, 2$, and $U_{i3}, V_{i3} \sim \mathcal{N}(0, 0.25), \forall i$. The response variable responds to the model $r(x_i) + \gamma_i$, where $r(x_i) = x_i\left(\frac{48\pi}{99}\right) + 2x_i\left(\frac{58\pi}{99}\right)x_i\left(\frac{128\pi}{99}\right)$ and $\gamma_i \sim_{iid} \mathcal{N}\left(0, \sigma_\gamma^2\right)$ with $\sigma_\gamma^2 = 5\%var\{r(x_i)\}$. See Fig. 5 for a graphical representation of them.

Two performance criteria, namely prediction accuracy and sparsity in the proportion of the curves to be used, are assessed. The prediction accuracy is evaluated by the sum of the squared errors (SSE) for the sake of comparison with benchmarks. The lower the SSE, the better the model in terms of prediction accuracy. The sparsity is evaluated by δ^{length}, as follows:

$$\delta^{\text{length}} = \left(1 - \frac{1}{\overline{s} - \underline{s}} \sum_{h=1}^{H} \left(s_{2h} - s_{2h-1}\right)\right) \times 100.$$

The higher the δ^{length}, the better the model in terms of sparsity.

A ten-fold cross-validation procedure has been applied, and Table 2 and Fig. 6 represent the average results of such ten test subsets.

The logistic CDF has been chosen for our experiments:

$$F(\cdot) = \frac{1}{1 + \exp(-(\cdot)\,\gamma)},$$

with a large value of γ, namely, $\gamma = 512$. We will illustrate that this small level of randomization is enough for obtaining good results.

The formulation has been implemented using the `scipy.optimize` package (Virtanen et al., 2020) in Python 3.7 (Python Core Team, 2015). As solver, we have used the SLSQP method (Kraft, 1988), since it is considered to be one of the most efficient computational methods to solve general NLO problems. SLSQP solves the optimization problem iteratively with a gradient descent strategy. The optimal descent direction is addressed by reducing the NLO problem to a standard quadratic programming subproblem. The resulting sequence of iterations converges to a local minimum of the NLO problem. For this reason, we will consider a multistart approach in our computational experience. The response variable has been normalized to the $[-1, 1]$ interval. As seen in Fig. 5, in real-world applications, the functional predictor variables are known for a finite set of points. Hence, smoothing techniques were applied as a preprocessing step so that an approximation to the original function can be obtained. In our study, each individual has been previously turned into a smooth cubic spline approximation using the `scipy.interpolate` package, where the domain $[\underline{s}, \overline{s}]$ has been shifted to $[0, 1]$ w.l.g. For this reason, the decision variables $\boldsymbol{s}$ have been restricted to the $[0, 1]$ interval. The decision variables $\boldsymbol{a}$ and $\boldsymbol{\mu}$ have been restricted to the $[-1, 1]$ interval for the sake of numerical stability with exponentials. Our experiments have been conducted on a PC, with an Intel® Core™ i7-7700 CPU 3.60 GHz processor and 32 GB RAM.

3.2. Results for S-ORRT-FD

In this section, we focus on testing the prediction accuracy of our approach. S-ORRT-FD with depths $D = 1, 2$ and H in the grid $\{1, \ldots, 10\}$ is compared against two benchmark regression methods. The former is SVR-FD (Blanquero et al., 2020a), a state-of-the-art machine learning model for multivariate functional data based on Support Vector Regression. The same grid for H is used in SVR-FD, as well as three different variants $d = 0, 1, 2$, which include, respectively, the situations where just the information of the raw functional data, or their monotonicity, or both their monotonicity and convexity are considered. The latter is Random Forest (RF) (Breiman, 2001), a sophisticated tree-based regression method competitive in terms of prediction accuracy. In contrast to S-ORRT-FD and SVR-FD, RF has no direct control on the domain of the functional predictor variables being used.

For S-ORRT-FD, since local searches may get stuck at bad local optima, we have followed a multistart approach, where the process is repeated 500 times starting from different initial solutions. The solutions found for H are given as initial solutions to the optimization problem for $H+1$, starting with random initial solutions for $H = 1$. The computing time taken by the S-ORRT-FD for a batch of ten values of H and 500 initial solutions typically ranges from 40 s (in Sunflower for $D = 1$) to 500 s (in FHV for $D = 2$). For SVR-FD, the results are taken from Blanquero et al. (2020a). They are comparable since ten-fold cross-validation was also performed, and the response variable was normalized to the $[-1, 1]$ interval. The default parameter setting in `randomForest` (Liaw and Wiener, 2002) R package has been used for running RF.

With respect to SVR-FD, S-ORRT-FD is competitive according to the results obtained in Table 2 for the data sets considered. The best values of SSE for each value of H between SVR-FD and S-ORRT-FD have been highlighted in bold. For Tecator, S-ORRT-FD outperforms SVR-FD for values of $H \geq 3$. For Sunflower, S-ORRT-FD beats SVR-FD for values of $H \leq 5$. For Sugar, S-ORRT-FD tends to dominate in terms of prediction accuracy for each value of H, while for FHV it is the other way around.

With respect to RF, S-ORRT-FD always manages to find a small number of critical intervals H for which a better prediction accuracy is achieved for the data sets considered. This is the case for Tecator with $H = 1$ and $D = 1$, Sunflower with $H = 1$ and $D = 1$, Sugar with $H = 2$ and $D = 1$, and FHV with $H = 5$ and $D = 2$.

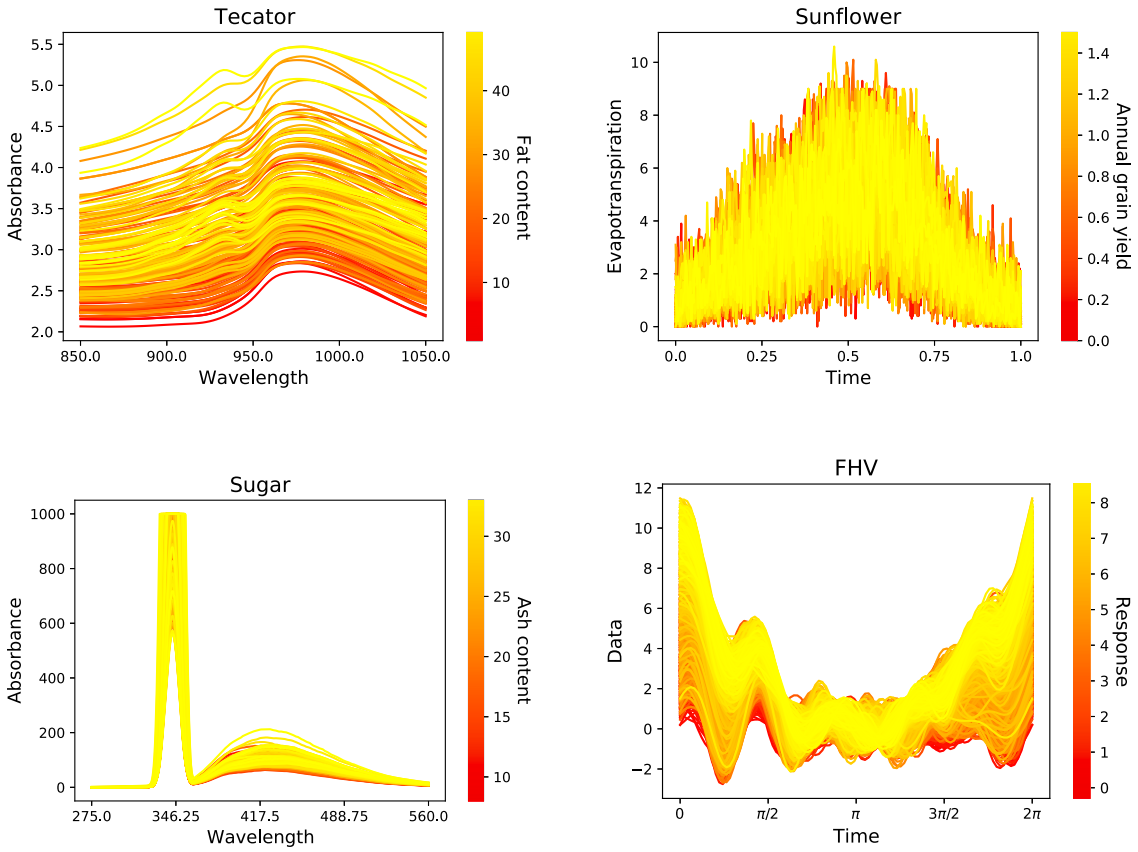


Fig. 5. Graphical representation, for each data set, of the evaluations of the original predictor variable function (Y-axis) known for a set of points in their domain (X-axis). Each observation in the data set is colored according to the response variable. The higher the value of the response variable, the colder the color.

Table 2
Comparison in terms of out-of-sample average SSE between our method S-ORRT-FD with depth D and benchmark methods RF and SVR-FD with information of derivatives up to order d.

Data set	Method	Out-of-sample average SSE									
		H									
		1	2	3	4	5	6	7	8	9	10
Tecator	S-ORRT-FD $D=1$	4.74	0.52	**0.21**	0.22	0.20	0.19	0.18	0.17	0.16	0.16
	S-ORRT-FD $D=2$	4.89	0.55	**0.21**	**0.14**	**0.17**	**0.13**	**0.13**	**0.11**	**0.10**	**0.11**
	SVR-FD $d=0$	0.25	**0.23**	0.24	0.29	0.41	0.40	0.45	0.45	0.49	0.50
	SVR-FD $d=1$	0.47	0.24	0.37	0.38	0.42	0.47	0.49	0.50	0.52	0.53
	SVR-FD $d=2$	**0.23**	0.36	0.39	0.43	0.47	0.55	0.58	0.60	0.61	0.61
	RF	2.13									
Sunflower	S-ORRT-FD $D=1$	**2.36**	2.69	**3.19**	**2.83**	**3.41**	6.00	6.50	4.62	4.34	4.65
	S-ORRT-FD $D=2$	2.57	**2.67**	3.68	4.24	4.21	4.73	6.40	5.91	5.85	5.40
	SVR-FD $d=0$	4.69	3.90	4.12	4.19	4.09	4.01	3.99	3.99	**3.92**	3.98
	SVR-FD $d=1$	4.29	3.86	4.11	4.32	4.24	4.08	4.07	4.01	9.95	3.98
	SVR-FD $d=2$	3.83	3.86	3.80	3.92	4.03	**3.91**	**3.93**	**3.96**	4.11	**3.94**
	RF	2.43									
Sugar	S-ORRT-FD $D=1$	**1.66**	**0.87**	2.12	2.06	2.44	**1.32**	**1.34**	**1.30**	**1.09**	**1.12**
	S-ORRT-FD $D=2$	1.76	1.26	**1.36**	2.28	**1.50**	1.58	1.85	1.44	3.74	2.75
	SVR-FD $d=0$	1.80	1.83	1.84	**1.84**	1.82	1.83	1.86	1.84	1.85	1.87
	SVR-FD $d=1$	2.08	1.81	1.82	1.88	1.84	1.88	1.85	1.83	1.82	1.84
	SVR-FD $d=2$	1.76	1.88	1.92	1.90	1.90	1.90	1.88	1.86	1.86	1.88
	RF	1.21									
FHV	S-ORRT-FD $D=1$	3.95	2.59	1.12	0.86	0.71	0.61	0.47	0.35	0.37	0.34
	S-ORRT-FD $D=2$	3.62	2.50	0.89	0.60	0.46	0.44	0.44	0.42	0.41	0.38
	SVR-FD $d=0$	**0.08**	**0.08**	**0.08**	**0.10**	**0.14**	**0.15**	**0.16**	**0.18**	0.19	0.21
	SVR-FD $d=1$	0.23	0.12	0.11	0.13	0.15	0.17	0.17	**0.18**	**0.18**	**0.18**
	SVR-FD $d=2$	0.12	0.13	0.12	0.14	0.15	0.16	0.17	0.19	0.21	0.25
	RF	0.51									

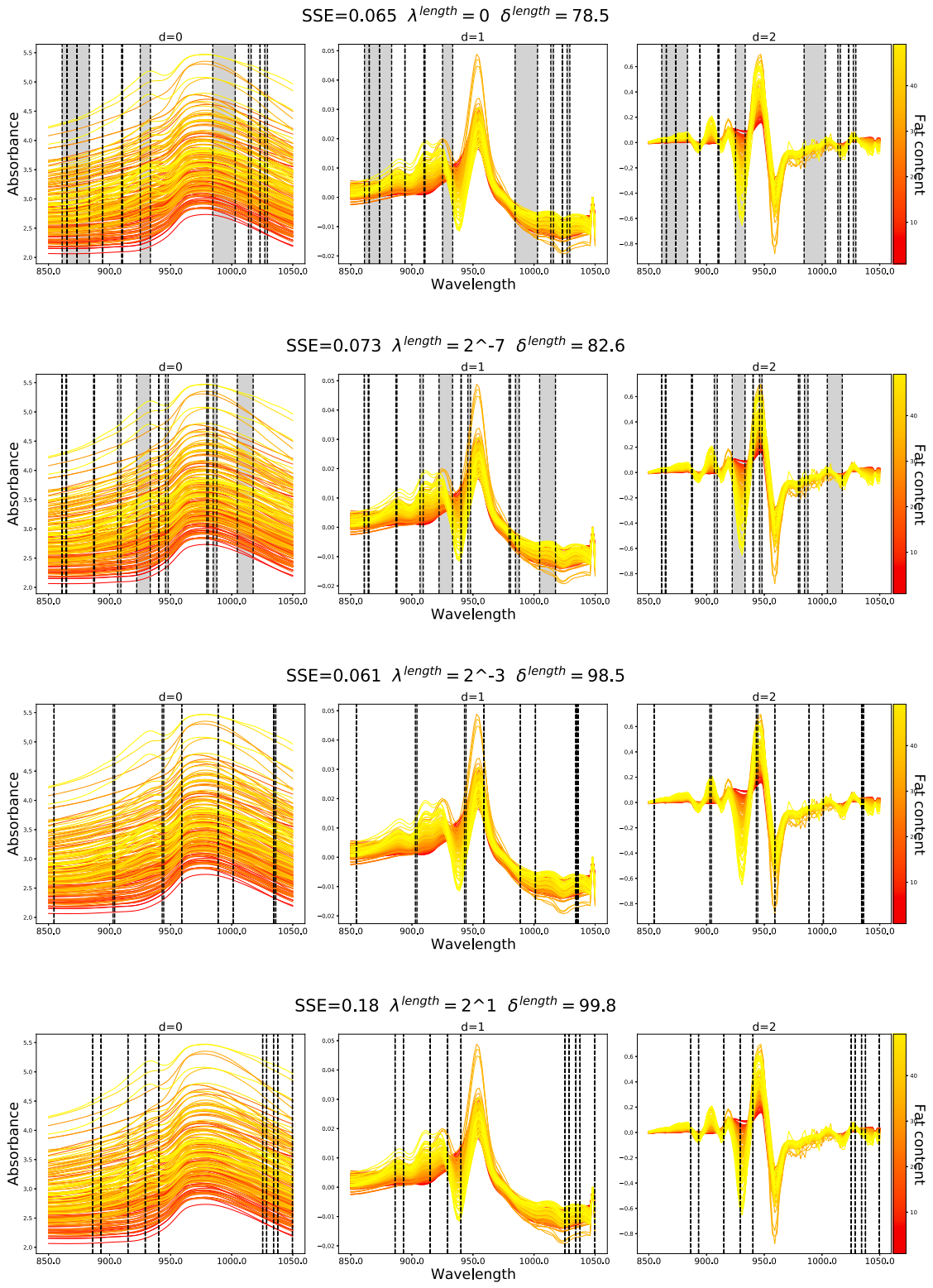


Fig. 6. Critical intervals detection for the Tecator data set as a function of λ^{length}. Two performance criteria, prediction accuracy and sparsity, are evaluated by the out-of-sample SSE and δ^{length}, respectively, where δ^{length} represents the proportion of the curve not being used.

In summary, these numerical results illustrate that S-ORRT-FD is competitive with the state-of-the-art machine learning model tailored to functional data, SVR-FD, and outperforms the standard benchmark regression method RF. Thanks to the design of our methodology, S-ORRT-FD shows to be versatile enough in order to capture nonlinear relationships between the functional predictor variables and the response variable. Furthermore, our approach can easily control global desirable properties such as sparsity, as seen in the next section.

A general observation derived from Table 2 is that it could happen that the larger the number of critical intervals H or the depth D, the larger the SSE. This might be due to different reasons. First, we note

that dealing with a higher H and/or D involves a tree structure with a higher number of parameters to be estimated, which may lead to overfitting. Second, although the model increases in complexity, we have fixed the number of starting points in the multistart approach to 500 independently of the configuration tested. Finally, a higher H can also yield autocorrelation problems.

3.3. Prediction accuracy and sparsity tradeoff

In the previous section, we have shown that considering a small number of critical intervals H, good results for S-ORRT-FD are achieved in terms of prediction accuracy. Nevertheless, the solution obtained could be not sparse, in the sense that the whole domain of the functional predictor variables could have been used for prediction, hence not giving information of the periods of highest relevance for predictions. In this section, we illustrate that both performance criteria, prediction accuracy and sparsity, can be improved. S-ORRT-FD can easily incorporate higher-order information to enhance prediction accuracy, at the same time of being able to trade some of it for a gain in sparsity, measured as the proportion of the curve not being used.

For the sake of conciseness, we illustrate this in the multivariate Tecator data set. We have solved Problem (9)–(14) with depth $D = 1$ and $H = 10$ for the sparsity parameters $\lambda^G = 0$ and λ^{length} in the grid $\{0\} \cup \{2^r,\ -10 \leq r \leq 1,\ r \in \mathbb{Z}\}$. We start solving the optimization problem with $\lambda^{\text{length}} = 0$ for 500 random initial solutions, and we continue for larger values of λ^{length}. The solutions found for fixed λ^{length}, are given as initial solutions to the next problem to be solved in the grid. This is an example of a multivariate functional setting, where $p = 3$ functional predictor variables, the raw data series ($d = 0$), the series of the first derivative ($d = 1$) and the series of the second derivative ($d = 2$), are used. See Fig. 1 for a graphical representation of them. The first and second derivatives are approximated by first using the finite difference method as explained below, and then smoothing the resulting sequences of increments as discussed in Section 3.1.

Let $x(s)$ denote the absorbance at wavelength s, where $s \in [850, 1050]$, being discretized into 100 equally spaced points, with step Δs. The first derivative was achieved with the forward difference method for points $\alpha = 1, \ldots, 99$ and the backward for $\alpha = 100$, as seen in the following equations:

$$x'(s_\alpha) = \begin{cases} \dfrac{x(s_{\alpha+1}) - x(s_\alpha)}{\Delta s}, & \alpha = 1, \ldots, 99, \\ \dfrac{x(s_{100}) - x(s_{99})}{\Delta s}, & \alpha = 100. \end{cases}$$

The second derivative was achieved with the forward differencing method for $\alpha = 1$, the central for $\alpha = 2, \ldots, 99$ and the backward for $\alpha = 100$, as seen in the following equations:

$$x''(s_\alpha) = \begin{cases} \dfrac{x(s_3) - 2x(s_2) + x(s_1)}{(\Delta s)^2}, & \alpha = 1, \\ \dfrac{x(s_{\alpha-1}) - 2x(s_\alpha) + x(s_{\alpha+1})}{(\Delta s)^2}, & \alpha = 2, \ldots, 99, \\ \dfrac{x(s_{100}) - 2x(s_{99}) + x(s_{98})}{(\Delta s)^2}, & \alpha = 100. \end{cases}$$

Fig. 6 illustrates the different solutions obtained for the out-of-sample average SSE and δ^{length} in the grid of λ^{length} considered. Three plots are depicted as a function of λ^{length}: the original functional predictor variable for Tecator data set, the first derivative and the second derivative. Each individual is colored depending on their value of the response variable. In gray, critical intervals detected for one of the folds are represented, and coincide for the three functional predictor variables. For $\lambda^{\text{length}} = 0$, an out-of-sample SSE of 0.065 is achieved, better than the value 0.160 observed in Table 2, which means that higher-order information was valuable for S-ORRT-FD prediction in this data set. As expected, the larger the value of λ^{length}, the smaller the proportion of the curve to be used. Indeed, we can enhance sparsity δ^{length} from 78.5% to 99.8%, at the cost of slightly damaging prediction accuracy SSE from 0.065 to 0.180, still begin superior to SVR-FD and RF results in Table 2.

4. Conclusions and future research

Many approaches on building optimal decision trees for multivariate data have been recently proposed in the literature. In this paper, we tailor the continuous optimization approach proposed previously by the authors to construct a regression tree, in order to consider multivariate functional data in addition. Critical intervals for prediction are detected simultaneously. While being competitive in terms of prediction accuracy to benchmark methods, including RF, our approach can directly control the desired number of critical intervals, as well as the number of functional predictor variables and the proportion of the curves to be used along the tree.

Several extensions to our approach are of interest. First, the coefficient functions $a_{jt}(\cdot)$ and $\tilde{a}_{jt}(\cdot)$ defined at branch and leaf nodes, respectively, could be extended to more sophisticated functions rather than piecewise constant ones. Second, other losses such as the mean absolute error or quantile regression can be considered, yielding optimization problems which deserve further analysis. Third, tailoring desirable properties such as cost-sensitivity, fairness or local explainability, to the context of Functional Data Analysis deserves further investigation. Fourth, it is known that bagging trees tends to enhance accuracy. An appropriate bagging scheme of our approach, where a collection of trees is built in order to have a global control on certain desirable properties, is an interesting open question. A parallelization framework would be suitable to make the training of the collection of trees tractable. Finally, tailoring optimal trees to other kinds of complex data that are not captured appropriately by standard implementations of these models such as text, image or network data is also an attractive research avenue.

CRediT authorship contribution statement

Rafael Blanquero: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Emilio Carrizosa:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Cristina Molero-Río:** Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Dolores Romero Morales:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

Data availability

Data will be made available on request.

Acknowledgments

This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant agreement ID: 822214), FQM-329, P18-FR-2369, US-1381178 (Junta de Andalucía, Spain) and PID2019-110886RB-I00 (funded by MCIN/AEI/10.13039/501100011033, Spain). This support is gratefully acknowledged.

References

Aghaei, S., Gomez, A., Vayanos, P., 2022. Strong optimal classification trees. arXiv preprint arXiv:2103.15965.

Ahuja, R., Magnanti, T., Orlin, J., 1993. Network Flows: Theory, Algorithms, and Applications. Prentice Hall, New Jersey.

Aneiros, G., Vieu, P., 2014. Variable selection in infinite-dimensional problems. Statist. Probab. Lett. 94, 12–20.

Aneiros, G., Vieu, P., 2016. Sparse nonparametric model for regression with functional covariate. J. Nonparametr. Stat. 28 (4), 839–859.

Balakrishnan, S., Madigan, D., 2006. Decision trees for functional variables. In: Sixth International Conference on Data Mining. ICDM'06, pp. 798–802.

Belli, E., Vantini, S., 2021. Measure inducing classification and regression trees for functional data. Stat. Anal. Data Min. 15 (5), 553–569.

Benítez-Peña, S., Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., 2019. Cost-sensitive feature selection for Support Vector Machines. Comput. Oper. Res. 106, 169–178.

Benítez-Peña, S., Carrizosa, E., Guerrero, V., Jiménez-Gamero, M.D., Martín-Barragán, B., Molero-Río, C., Ramírez-Cobo, P., Romero Morales, D., Sillero-Denamiel, M.R., 2021. On sparse ensemble methods: An application to short-term predictions of the evolution of COVID-19. European J. Oper. Res. 295 (2), 648–663.

Berrendero, J.R., Bueno-Larraz, B., Cuevas, A., 2019. An RKHS model for variable selection in functional linear regression. J. Multivariate Anal. 170, 25–45.

Berrendero, J.R., Cuevas, A., Torrecilla, J.L., 2018. On the use of reproducing kernel Hilbert spaces in functional classification. J. Amer. Statist. Assoc. 113 (523), 1210–1218.

Bertsimas, D., Dunn, J., 2017. Optimal classification trees. Mach. Learn. 106 (7), 1039–1082.

Bixby, R.E., 2012. A brief history of linear and mixed-integer programming computation. Doc. Math. 2012, 107–121.

Blanquero, R., Carrizosa, E., Chis, O., Esteban, N., Jiménez-Cordero, A., Rodríguez, J.F., Sillero-Denamiel, M.R., 2016. On extreme concentrations in chemical reaction networks with incomplete measurements. Ind. Eng. Chem. Res. 55 (44), 11417–11430.

Blanquero, R., Carrizosa, E., Jiménez-Cordero, A., Martín-Barragán, B., 2020a. Selection of time instants and intervals with support vector regression for multivariate functional data. Comput. Oper. Res. 123, 105050.

Blanquero, R., Carrizosa, E., Molero-Río, C., Romero Morales, D., 2020b. Sparsity in optimal randomized classification trees. European J. Oper. Res. 284 (1), 255–272.

Blanquero, R., Carrizosa, E., Molero-Río, C., Romero Morales, D., 2021a. Optimal randomized classification trees. Comput. Oper. Res. 132, 105281.

Blanquero, R., Carrizosa, E., Molero-Río, C., Romero Morales, D., 2022. On sparse optimal regression trees. European J. Oper. Res. 299 (3), 1045–1054.

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., Sillero-Denamiel, M.R., 2021b. A cost-sensitive constrained lasso. Adv. Data Anal. Classif. 15, 121–158.

Borggaard, C., Thodberg, H.H., 1992. Optimal minimal neural interpretation of spectra. Anal. Chem. 64 (5), 545–551.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Cai, T.T., Hall, P., 2006. Prediction in functional linear regression. Ann. Statist. 34 (5), 2159–2179.

Carrizosa, E., Molero-Río, C., Romero Morales, D., 2021a. Mathematical optimization in classification and regression trees. TOP 29 (1), 5–33.

Carrizosa, E., Restrepo, M.G., Romero Morales, D., 2021b. On clustering categories of categorical predictors in generalized linear models. Expert Syst. Appl. 182, 115245.

Chan, H., Rice, E., Vayanos, P., Tambe, M., Morton, M., 2018. From empirical analysis to public policy: Evaluating housing systems for homeless youth. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 69–85.

Cuevas, A., 2014. A partial overview of the theory of statistics with functional data. J. Statist. Plann. Inference 147, 1–23.

Demirović, E., Lukina, A., Hebrard, E., Chan, J., Bailey, J., Leckie, C., Ramamohanarao, K., Stuckey, P.J., 2022. MurTree: Optimal classification trees via dynamic programming and search. J. Mach. Learn. Res. 23 (26), 1–47.

Dunn, J., 2018. Optimal Trees for Prediction and Prescription (Ph.D. thesis). Massachusetts Institute of Technology.

Fan, Y., James, G.M., Radchenko, P., 2015. Functional additive regression. Ann. Statist. 43 (5), 2296–2325.

Febrero-Bande, M., de la Fuente, M.O., 2012. Statistical computing in functional data analysis: The R package fda.usc. J. Stat. Softw. 51 (4), 1–28.

Febrero-Bande, M., González-Manteiga, W., Oviedo De La Fuente, M., 2019. Variable selection in functional additive regression models. Comput. Statist. 34 (2), 469–487.

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

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice, Vol. 76. Springer.

Firat, M., Crognier, G., Gabor, A., Hurkens, C., Zhang, Y., 2019. Column generation based math-heuristic for classification trees. Comput. Oper. Res. 116, 104866.

Freitas, A., 2014. Comprehensible classification models: A position paper. ACM SIGKDD Explor. Newslett. 15 (1), 1–10.

Georganos, S., Grippa, T., Gadiaga, A.N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S., 2021. Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int. 36 (2), 121–136.

Goia, A., Vieu, P., 2016. An introduction to recent advances in high/infinite dimensional statistics. J. Multivariate Anal. 146, 1–6.

Goodman, B., Flaxman, S., 2017. European union regulations on algorithmic decision-making and a "right to explanation". AI Mag. 38 (3), 50–57.

Griswold, C.K., Gomulkiewicz, R., Heckman, N., 2008. Hypothesis testing in comparative and experimental studies of function-valued traits. Evolution 62 (5), 1229–1242.

Grollemund, P.-M., Abraham, C., Baragatti, M., Pudlo, P., 2019. Bayesian functional linear regression with sparse step functions. Bayesian Anal. 14 (1), 111–135.

Günlük, O., Kalagnanam, J., Li, M., Menickelly, M., Scheinberg, K., 2021. Optimal decision trees for categorical data via integer programming. J. Global Optim. 81, 233–260.

Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. Ann. Statist. 23 (1), 73–102.

Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications, Vol. 200. Springer Science & Business Media.

Hu, X., Rudin, C., Seltzer, M., 2019. Optimal sparse decision trees. In: Advances in Neural Information Processing Systems 32. pp. 7265–7273.

James, G.M., Wang, J., Zhu, J., 2009. Functional linear regression that's interpretable. Ann. Statist. 37 (5A), 2083–2108.

Januschowski, T., Wang, Y., Torkkola, K., Erkkilä, T., Hasson, H., Gasthaus, J., 2022. Forecasting with trees. Int. J. Forecast. 38 (4), 1473–1481.

Jiménez Cordero, M.A., 2019. Classification and Regression with Functional Data: A Mathematical Optimization Approach (Ph.D. thesis). University of Seville.

Kong, D., Xue, K., Yao, F., Zhang, H.H., 2016. Partially functional linear regression in high dimensions. Biometrika 103 (1), 147–159.

Kraft, D., 1988. A Software Package for Sequential Quadratic Programming. Tech. Rep., DFVLR-FB 88-28, DLR German Aerospace Center — Institute for Flight Mechanics, Köln, Germany.

Laukaitis, A., Račkauskas, A., 2005. Functional data analysis for clients segmentation tasks. European J. Oper. Res. 163 (1), 210–216.

Leng, X., Müller, H.-G., 2005. Classification using functional data analysis for temporal gene expression data. Bioinformatics 22 (1), 68–76.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.

Lin, J., Zhong, C., Hu, D., Rudin, C., Seltzer, M., 2020. Generalized and scalable optimal sparse decision trees. In: International Conference on Machine Learning. pp. 6150–6160.

Loh, W.-Y., 2014. Fifty years of classification and regression trees. Internat. Statist. Rev. 82 (3), 329–348.

Meinshausen, N., 2010. Node harvest. Ann. Appl. Stat. 4 (4), 2049–2072.

Narodytska, N., Ignatiev, A., Pereira, F., Marques-Silva, J., 2018. Learning optimal decision trees with SAT. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. IJCAI-18, pp. 1362–1368.

Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., Reynkens, T., 2022. Social network analytics for supervised fraud detection in insurance. Risk Anal. 42 (8), 1872–1890.

Picheny, V., Servien, R., Villa-Vialaneix, N., 2019. Interpretable sparse SIR for functional data. Stat. Comput. 29 (2), 255–267.

Python Core Team, 2015. Python: A dynamic, open source programming language. Python Software Foundation, URL https://www.python.org.

Ramon, Y., Martens, D., Provost, F., Evgeniou, T., 2020. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. Adv. Data Anal. Classif. 14 (4), 801–819.

Ramsay, J., Silverman, B., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer, New York.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis. Springer, New York.

Ribeiro, M., Singh, S., Guestrin, C., 2016. "Why should I trust You?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

Saha, A., Basu, S., Datta, A., 2021. Random forests for spatially dependent data. J. Amer. Statist. Assoc. 1–19.

Soenksen, L.R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H.M., Li, M.L., Fuentes, I., Bertsimas, D., 2022. Integrated multimodal artificial intelligence framework for healthcare applications. NPJ Digit. Med. 5 (1), 1–10.

Strzalkowska-Kominiak, E., Romo, J., 2021. Censored functional data for incomplete follow-up studies. Stat. Med. 40 (12), 2821–2838.

Verhaeghe, H., Nijssen, S., Pesant, G., Quimper, C.-G., Schaus, P., 2019. Learning optimal decision trees using constraint programming. In: The 25th International Conference on Principles and Practice of Constraint Programming. CP2019.

Verwer, S., Zhang, Y., 2017. Learning decision trees with flexible constraints and objectives using integer optimization. In: Salvagnin, D., Lombardi, M. (Eds.), Integration of AI and OR Techniques in Constraint Programming: 14th International Conference, CPAIOR 2017, Padua, Italy, June 5-8, 2017, Proceedings. pp. 94–103.

Verwer, S., Zhang, Y., Ye, Q., 2017. Auction optimization using regression trees and linear models as integer programs. Artificial Intelligence 244, 368–395.

Verwer, S., Zhang, Y., Ye, Q., 2019. Learning optimal classification trees using a binary linear program formulation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 1625–1632.

Vieu, P., 2018. On dimension reduction models for functional data. Statist. Probab. Lett. 136, 134–138.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 17, 261–272.

Wang, J.-L., Chiou, J.-M., Müller, H.-G., 2016. Functional data analysis. Annu. Rev. Stat. Appl. 3, 257–295.

Wang, P., Fan, E., Wang, P., 2021. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognit. Lett. 141, 61–67.

Yu, J., Ignatiev, A., Stuckey, P., Le Bodic, P., 2020. Computing optimal decision sets with SAT. In: International Conference on Principles and Practice of Constraint Programming. pp. 952–970.

Zafar, M., Valera, I., Gomez Rodriguez, M., Gummadi, K., 2017. Fairness constraints: Mechanisms for fair classification. In: Artificial Intelligence and Statistics. PMLR, pp. 962–970.

Zantedeschi, V., Kusner, M., Niculae, V., 2020. Learning binary trees via sparse relaxation. arXiv preprint arXiv:2010.04627.

Zhu, H., Murali, P., Phan, D., Nguyen, L., Kalagnanam, J., 2020. A scalable MIP-based method for learning optimal multivariate decision trees. Adv. Neural Inf. Process. Syst. 33, 1771–1781.