MOX-Report No. 47/2025

Penalised Optimal Soft Trees for Functional Data

Gimenez Zapiola, A.; Consolo, A.; Amaldi, E.; Vantini, S.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it https://mox.polimi.it

Penalised Optimal Soft Trees for Functional Data

Alfredo Gimenez Zapiola[1], Antonio Consolo[2], Edoardo Amaldi[3], and Simone Vantini[1]

[1]MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy
[2]DISCo, Università Milano-Bicocca, Milan, Italy
[3]DEIB, Politecnico di Milano, Milan, Italy

August 12, 2025

Abstract

We propose a new tree-based classifier for Functional Data. A novel objective function for Suárez and Lutsko (1999)'s globally-optimised Soft Classification Trees is proposed to adapt it to the Functional Data Analysis setting when using an FPCA basis. It consists of a supervised and an unsupervised term, with the latter working as a penalisation for heterogeneity in the leaf nodes of the tree. Experiments on benchmark data sets and two case studies demonstrate that the penalisation and proposed initialisation heuristics work synergically to increase model performance both in the train and test data set. In particular, including the unsupervised term shows to aid the supervised term to reach better objective function values. The case studies specifically illustrate how the unsupervised term yields adaptiveness to different problems, by using custom criteria of homogeneity in the leaf nodes. The interpretability of the splitting functions at the internal nodes is also discussed.

Keywords: functional data classification · regularisation · soft classification tree · penalised learning· decision tree learning

1 Introduction

In the context of high-dimensional data, Functional Data Analysis (FDA) has consolidated itself as statisticians' go-to framework whenever the different measurements per statistical unit are indexed by some continuous domain, such as space or time. Several classical methods, such as PCA, regression, testing, amongst others, have been extended to such setting (Ramsay and Dalzell 1991). Classification models are no exception, and owing to the relevance of this growing field in statistics, we seek to contribute to this stream of literature through a nonparametric, interpretable and performant model that combines ideas from globally-optimised Decision Trees, where the overall tree is estimated by minimising the objective function over all its decision variables, and the FDA literature.

The most relevant methods for predicting the label of an observation when the data are functions, i.e, classification in FDA, have been reviewed by Wang, Huang, and Cao (2024). In the FDA setting, formulating parametric assumptions on the underlying distribution of the data is particularly critical, since these suppositions can lead to severe misspecification bias and can be difficult to test for, as studied for e.g. by Cuevas, Febrero, and Fraiman (2007)'s work, which has led to the growth of nonparametric FDA (Ferraty 2006). Furthermore, the interpretability of FDA models is a key issue given the infinite-dimensional nature of the objects of analysis.

On the other hand, since Breiman, Friedman, Olshen, and Stone (1984)'s seminal paper, decision tree learning has been widely adopted both by the statistical and machine learning communities, cfr. (Breiman 2001b), altogether with ensemble variants such as Bagging, Random Forests (Breiman 2001a) and Boosting (Hastie, Tibshirani, Friedman, Hastie, Tibshirani, and Friedman 2009). The optimisation problem that has to be solved in order to fit a decision tree is NP-hard (Hyafil and Rivest 1976), and as a result in the

past these models were solved through greedy heuristics, optimise one node of the tree at a time. They are referred to as Classification and Regression Trees (CARTs). Nevertheless, thanks both to an increase in computational power and improved solvers driven by contributions in the operations research literature, a plethora of alternative formulations have been proposed optimise trees globally. These are in turn divided into those with hard splitting rules (Verwer and Zhang 2019; Bertsimas and Dunn 2017); and so-called soft splitting rules, started by Suárez and Lutsko (1999) and further developed by Blanquero, Carrizosa, Molero-Río, and Morales (2021); Blanquero, Carrizosa, Molero-Río, and Romero Morales (2020). In the first case, observations are routed along the tree graph following one single path from the start node until a leaf node; whereas in the latter they can follow different paths to different leaf nodes but in a probabilistic fashion.

In the particular context of CARTs for functional data, we retrieve Belli and Vantini (2022)'s and Maturo and Verde (2023)'s works. Blanquero, Carrizosa, Molero-Río, and Romero Morales (2023) use soft trees focusing in domain selection in the case of regression. In the current work, we build upon optimal (non-greedy) soft classification trees (SCTs), devised by Suárez and Lutsko (1999), expanded and called Optimal Randomised Classification Trees (ORCTs) (Blanquero et al. 2021). In order to tailor SCTs to functional data, we propose a new objective function that contains both a supervised and an unsupervised term, where the latter promotes leaf node homogeneity and also works as a penalisation to mitigate overfitting. We additionally provide a heuristic method that combines bootstrap aggregation and functional LDA (FLDA, see James and Hastie (2002)), yielding a statistically-driven initialisation of the parameters of the tree. Both proposals work in syngery with interior-point methods for optimisation, yielding more stable and better results each time the mathematical program is solved, as we demonstrate through a simulation study. Moreover, to improve interpretability, a suitable dimensionality reduction through

FPCA is performed. Indeed, estimation yields what we name a *splitting function* at each internal node of the tree, which provides a direct interpretability of how the model learns to separate the functional data.

The outline for the current work is the following: in Section 2.1, we first review classification methods in FDA, as well as CARTs and (optimal) soft trees, including ORCTs in detail. In Section 2.2 we delve into the details of our proposal, including the formulation with its penalisation, the initialisation methods we propose, and outline its interpretability through the splitting functions. Next, we turn to its application: in Section 3.2 the algorithm's performance is tested on standard benchmark data sets for functional classification and we show its worth in real-world application in Section in 3.3.2 and 3.3.1. We conclude by commenting our results and outlining possible research directions to further pursue.

2 Methods

2.1 Preliminaries

2.1.1 Notation for Functional Data

Letting $(\Omega, \mathcal{F}, \mathbb{P})$ be a (complete) probability space, we denote as functional datum a realisation of a real-valued process $\mathcal{X}$ defined over a compact subset of $\mathbb{R}^d$, viz. $X(s) = X(\omega, s), \;\; s \in \mathcal{I}$ with $\mathcal{I} \subset \mathbb{R}^d$. In the FDA literature, (cfr. (Ferraty 2006)), it is typically assumed either that (functional) realisations $X(s)$ belong to the $L^2(\mathcal{I}; \mathbb{R})$ H-space or that they are continuous functions: $\mathbb{E}[\int_{\mathcal{I}} X(s) d\mathbf{s}] < +\infty$ (sometimes even $\int_{\mathcal{I}} X(s) d\mathbf{s} < +\infty \; \mathbb{P}\, a.s.$) or $X(s) \in \mathcal{C}(\mathcal{I}, ||.||_\infty; \mathbb{R}) \;\; \mathbb{P}\, a.s.$, respectively. In a given data set, functions are discretely sampled, since it is unrealistic to gather infinite measurements per statistical unit. In the present work we assume they are observed on a fine enough grid, such that standard

smoothing techniques may be applied, such as splines or nonparametric techniques, see (Ramsay and Silverman 2005).

2.1.2 Classification of Functional Data

Given a training data set of N functional data observed at p points of their domain, with observed labels, viz.

$$\left\{ \left(\{X_i(s_j)\}_{j=1}^{p}, y_i \right) \right\}_{i=1}^{N} \tag{1}$$

where $y_i \in \{1, \ldots, K\}$, K denoting the number of different labels present in the data set is given. The objective is to fit a model that is able to predict correctly the label of a newly observed function $\{X_{new}(s_j)\}_{j=1}^{p}$.

Several of the available models are adaptations of algorithms for multivariate data, applied to a suitable dimension reduction of the infinite-dimensional functional data set, but *ad hoc* methods have also been proposed for the FDA setting. The latter in turn can be divided into two groups, namely distance-based approaches, where notable examples are centroid classifiers (Delaigle and Hall 2012, 2013), nearest-neighbours (Galeano, Joseph, and Lillo 2015) (Venturini, Muñoz, and González 2014); and those which utilise so-called functional depths, building onto Cuevas, Febrero, and Fraiman (2006)'s seminal paper, for e.g. Hlubinka, Gijbels, Omelka, and Nagy (2015)'s. Their popularity is attested by their inclusion of the popular Python package *Scikit-fda* (Ramos-Carreño, Torrecilla, Carbajo Berrocal, Marcos Manchón, and Suárez 2024). The second group exploits Reproducing Kernel Hilbert spaces, which allow for a more efficient computation for dot products in Hilbert spaces, being directly applicable to distance calculations (Wang et al. 2024) , cfr. contributions by Berrendero, Cuevas, and Torrecilla (2018) *et* Sang, Kashlak, and Kong (2023).

Since we build upon SCTs, in this research we have chosen the other possibility, *id*

est, to firstly perform a suitable dimensional reduction on the functional data set and algorithmically deal with multivariate data. The key idea is representing each functional datum as a linear combination of a family $\mathcal{G}$ of basis functions or *bases*:

$$x_i(s) = \sum_{g \in \mathcal{G}} \beta_g^{(i)} \phi_g(s) \quad i = 1, ..., N \tag{2}$$

The most popular *bases* in the FDA literature are Fourier, Splines, Wavelets, step functions, amongst others (Ramsay and Silverman 2005), as well as data-driven *bases*, such as functional PCA (FPCA) or functional canonical correlation analysis (FCCA), cfr. (Kneip 1994)(Rice and Silverman 1991).

In the current work, we consider FPCA. Assuming realisations to be in $L^2(\mathcal{I}; \mathbb{R})\, \mathbb{P}\ a.s.$, the variance-covariance function $C : \mathcal{I} \times \mathcal{I} \to \mathbb{R}^+$ given by $C(s_1, s_2) := \mathbb{E}[\big(X(s_1) - \mu(t_1)\big)\big[\big(X(s_2) - \mu(s_2)\big)\big]$, $s_1, s_2 \in \mathcal{I}$ satisfies the conditions to apply Mercer's theorem, yielding the Karhunen-Loève expansion (Hsing and Eubank (2015) provides details). Letting $\mathbb{E}[X(s)] = \mu(s),\ s \in \mathcal{I}$ be the pointwise expected value of the process, one obtains

$$x_i(s) = \mu(s) + \sum_{j=1}^{\infty} \beta_j^{(i)} \xi_j(s) \quad i = 1, ..., N \tag{3}$$

where $\xi_j(s)$ are the eigenfunctions corresponding to the (infinite-dimensional) spectral decomposition of $C(.\,, .)$.

In practice, though, the covariance function $C(.\,, .)$ is unknown and has to be estimated from the available data. Assuming the sampling grid of the functional data to be fine enough, the sample covariance estimator for multivariate data can be considered a good enough finite rank (that is, finite-dimensional) operator that is converging to the true infinite-dimensional covariance function. Indeed, since the covariance function is a compact operator in a Hilbert space, it can be approximated by a sequence of finite rank operators (Brézis 2011). Since we assume the functions to belong to $L^2(\mathcal{I}; \mathbb{R})$, the sample covariance

estimator is consistent. Hence, eigenfunctions $\xi(s)$ of $C(.,.)$ can be approximated by the (finite dimensional) eigenvectors of the sample (multivariate) covariance matrix $N^{-1}\mathbf{X}^T\mathbf{X}$, where $\mathbf{X}$ is the $N \times p$ matrix of centred observed functions. As typically done in dimension reduction techniques, a finite *basis* is constructed by keeping the first J eigenvectors. Then, Ordinary Least Squares (OLS) estimation is employed to obtain the weights of the linear combination for each functional datum of the sample, as follows:

$$x_i(s) = \sum_{j=1}^{J} \hat{\beta}_j^{(i)} \tilde{\xi}_j(s) + \epsilon_i(s) \tag{4}$$

with $\epsilon_i(s)$ being i.i.d. functions with in $L^2(\mathcal{I}; \mathbb{R})\, \mathbb{P}\ a.s.$ with $\mathbb{E}[\epsilon_i(s)] = 0$. Whereas other *bases* could be employed, we make such choice because (i) estimation can be made without smoothing the functional data, provided the sampling grid is fine enough, (ii) it grants interpretability (see Section 2.2.2), (iii) it allows to perform an initialisation heuristic based on FLDA (Section 2.2.3), (iv) it allows for a quick computation of the distance between functions when such is the homogeneity criterion in the penalisation, as shown Section 2.1.2. Indeed that due to the orthogonality of $\{\tilde{\xi}_j\}_{j=1}^{J}$, the estimated $\hat{\beta}_j^{(i)}, j \in \{1, \ldots, J\}$ correspond to the estimates of the principal component scores of the i-th datum with respect to the j-th principal component:

$$\hat{\beta}_j^{(i)} = \langle \mathbf{x}_i, \xi_\mathbf{j} \rangle_{\mathbb{R}^p} \approx \langle X_i(s)\,, \xi_j(s) \rangle_{L^2(\mathcal{I};\mathbb{R})} = \int_{\mathcal{I}} X_i(s) \xi_j(s) ds \tag{5}$$

Regarding aforementioned CARTs expressly for functional data, Belli and Vantini (2022)'s and Maturo and Verde (2023)'s contributions stand out; and for SCTs Riccio, Maturo, and Romano (2024)'s and Blanquero et al. (2023)'s. However, the first three perform greedy optimisation; and the latter, originally designed for regression, could be adapted for classification, yet using step functions *bases* whose support is an optimising variable yields to awkward integral evaluations at the objective function and lacks the flexibility of non-linear

bases. In the present work, interior point methods for trees proposed by Suárez and Lutsko (1999) for Blanquero et al. (2021)'s soft tree global formulation are exploited the context of functional data. After choosing FPCA for the *basis* expansion, two novelties are proposed: (i) adding an unsupervised term in the objective function that promotes homogeneity between the functions at each leaf node, working as a penalisation (ii) devising a heuristic for the initialisation of the optimising variables based on bootstrap aggregation and FLDA. As a result of (i), the first non-greedy SCT in the FDA setting to our knowledge is created. (ii) aids the solver to yield solution with increased performance and lower variability with respect to a trivial initial solutions, as evinced from the simulation study in Section 3. Moreover, by performing the so-called soft splits with respect to all FPCA scores at each internal node yields as an easy and intuitive way of visualising the splitting functions, enhancing interpretability, cfr. Section 2.2.2.

2.1.3 ORCTs and variants

A classification tree can be viewed as a directed binary graph composed of a set of internal (branch) nodes τ_B, which includes the root, and a set of terminal (leaf) nodes τ_L. Each internal node applies a binary splitting rule that routes an input vector $\mathbf{z} \in \mathbb{R}^J$ along one of its two outgoing arcs. Starting from the root, the input vector follows these rules until it reaches a leaf node, where the model assigns a class label $y \in \{1, \ldots, K\}$.

In Soft Classification Trees (SCTs), each branch node employs a link function to determine the routing probability. For an input vector $\mathbf{z}_i$ and a branch node $t \in \tau_B$, the probability of moving to the left child is

$$p_{it} = p_{it}\big(\mathbf{z}_i; (\mathbf{a}_t, m_t)\big) = F\left(\sum_{j=1}^{J} a_{jt} z_{ij} - m_t\right) \tag{6}$$

where $a_{jt} \in \mathbb{R}$ and $m_t \in \mathbb{R}$ arex the decision variables and $F(\cdot)$ denotes the link function, which is usually taken as the logistic distribution function $F(v) = \{1 + \exp(-v)\}^{-1}$. Accordingly, the probability of choosing the right child is $1 - p_{it}$.

Due to the application of distribution functions, every input vector reaches each leaf node with non-negative probability. Thus, the probability that input vector $\mathbf{z}_i$ falls in leaf node $t \in \tau_L$ is

$$P_{it} = P_{it}\big(\mathbf{z}_i; (\mathbf{a}_t, m_t)\big) = \prod_{t_l \in A_{L(t)}} p_{it_l} \prod_{t_r \in A_{R(t)}} (1 - p_{it_r})$$

where $A_{L(t)}$ is the set of ancestors of t whose left branch lies on the path from the root to t, and $A_{R(t)}$ is the set of ancestors whose right branch belongs to that path. Figure 1 illustrates a soft decision tree with depth $D = 2$.

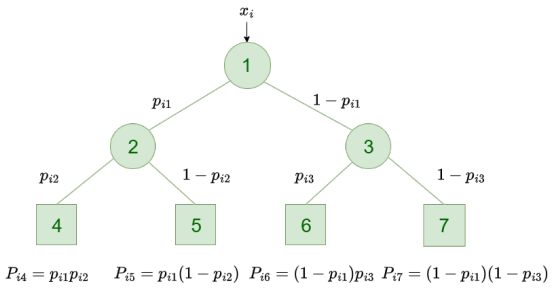

Figure 1: A soft decision tree of depth $D = 2$.

In (Blanquero et al. 2021, 2020), for each leaf node $t \in \tau_L$ and class label $k \in \{1, \ldots, K\}$, the authors define a binary decision variable c_{kt}, which takes the value 1 if all data assigned to node t are classified with label k, and 0 otherwise. To formulate the misclassification error, they introduce, for each observation $(\mathbf{z}_i, y_i)$ with $i \in 1, \ldots, N$ and for each class $k \in \{1, \ldots, K\}$, a parameter $w_{y_i k} \geq 0$ representing the cost incurred when $\mathbf{z}_i$ is classified as class k instead of its true label y_i.

As a result, the task of minimizing the expected misclassification error over the training

set can be expressed as the following continuous nonlinear optimization problem:

$$\min \quad N^{-1} \sum_{i=1}^{N} \sum_{t \in \tau_L} P_{it} \sum_{k=1}^{K} w_{y_i k} c_{kt} \tag{7a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} c_{kt} = 1 \qquad t \in \tau_L, \tag{7b}$$

$$\sum_{t \in \tau_L} c_{kt} \geq 1 \qquad \forall k \in \{1, ..., K\}, \tag{7c}$$

$$a_{jt} \in [-1, 1], \;\; m_t \in [-1, 1] \quad j \in \{1, \ldots, J\}, t \in \tau_B, \tag{7d}$$

$$c_{kt} \in [0, 1] \qquad k \in \{1, \ldots, K\}, t \in \tau_L, \tag{7e}$$

where the set of constraints in (7b) enforces a unique class label assignment for each leaf node, whereas constraints (7c) require that each class label k be assigned to at least one leaf node.

It is worth noting that the integrality constraints on the binary variables c_{kt} have been relaxed in the above formulation, as it has been shown by Blanquero et al. (2021) that the continuous nonlinear problem admits an optimal solution that is also integer.

According to Blanquero et al. (2021), we will refer to the mentioned SCT of depth D obtained by minimizing the misclassification error (7a) subject to (7b)–(7c) as an Optimal Randomized Classification Tree (ORCT), without emphasizing the soft multivariate splits. Blanquero et al. (2020) promote sparsity in SCTs by adding regularization terms involving the ℓ_1 and ℓ_∞ norms to the expected misclassification error. Amaldi et al. (2023) work induce sparsity by exploiting concave approximations of the ℓ_0 norm.

Concerning soft trees for functional data, (Blanquero et al. 2023) extend the formulation proposed in (Suárez and Lutsko 1999; Blanquero et al. 2022) to construct a soft regression tree that performs domain selection of critical intervals for prediction by applying ℓ_1

regularization to the coefficients of functional features.

2.2 Penalised SCT for Functional Data

2.2.1 The homogeneity penalisation

In the current work, we extend the objective function (7) (which in turn extends Suárez and Lutsko (1999)'s formulation) to handle functional data by incorporating an unsupervised term which serves as penalisation that takes into account the expected homogeneity at each leaf node. Following the notation from Section 2.1.3, and denoting d_{jl} the dissimilarity between the j-th and l-th functional data from the available sample (1):

$$N^{-1}\sum_{i=1}^{N}\left(\sum_{t\in\tau_L}P_{it}(\mathbf{z_i};\mathbf{a})\sum_{k=1}^{K}W_{y_ik}C_{kt}\right)+\alpha\,|\tau_L|^{-1}\sum_{t\in\tau_L}\left(\sum_{j=1}^{N}\sum_{l=j+1}^{N}d_{jl}\,P_{jt}(\mathbf{z}_j;\mathbf{a})P_{lt}(\mathbf{z}_l;\mathbf{a})\right) \quad (8)$$

with $\alpha\in\mathbb{R}^+$ a hyperparameter to be set.

The rationale behind such decision is that the distribution of the functional data, conditional on the label, should be the same, and thus with respect to some criterion functions should be homogeneous, pushing their dissimilarity to 0. The algorithm is *ipso facto* divided into a supervised part for classification (misclassification cost) and an unsupervised one for clustering (penalty), with the result being a **semi-supervised** model. These two goals are not opposed: actually, as seen in Section 3, in some cases the clustering term leads to better results in the classification, both in the generalisation of the model (test-set performance) and on the reduced variance of the model with respect to the starting points for the optimisation. We choose to name the d_{jl} as dissimilarity instead of distance since any other suitable functional of the distribution of the functional data may be utilised to help . This also promotes flexibility, since the choice of the dissimilarity is problem-specific (Sections 3.3.2 and 3.3.1).

Note that since the proposed model, apart from the penalisation, minimises expected classification cost, an desirable value of such objective function (7a) would mean better values in a family of standard performance indices of classification, such as accuracy, area under the Receiver Operating Characteristic (ROC), F1-score, amongst others. The justification behind the penalisation, besides from the within-label homogeneity argument exposed in Section 2.2, is to obtain desirable classification metrics not only on the training set, but on data the model has not been calibrated with, referred to as the test set.

The computation of the dissimilarities d_{jl} should be done before the mathematical optimisation, once the training set has been fixed. Whereas the $O(n^2)$ complexity may be worrisome for large datasets, this another instance where using the FPCA *basis* provides an advantage: the L^2 distance between two functions can be computed by the euclidean norm of the difference of FPCA scores (and same for other quantities due to the isomorphism of all spaces of dimension p to $\mathbb{R}^p$, cfr. Brézis (2011) *et* Ramsay and Silverman (2005)). Such procedure is very light computationally and would require the sample size N to be higher than any known FDA application to our knowledge to be a bottleneck in the inference procedure.

2.2.2 Model interpretability

In high-stakes domains, such as medicine, finance, and criminal justice, it is important to know why, given a (new) observation, a class assignment is made. Domain experts are interested in understanding what characteristics of the observation are looked at by the model to predict such label. The advantage of working in the context of soft trees, and deciding to make the dimension reduction through FPCA is that the *bases* functions are re-weighted according to he estimated $\mathbf{a}_t$, $t \in \tau_B$, inheriting the capacity to interpret (splitting) functions as perturbations of the mean. This is explained as follows. At internal

node $t \in \tau_B$, parameters $a_{jt},\ j \in \{1, \ldots, J\}$ and m_t are used to determine the probability for a given observation $\mathbf{z} \in \mathbb{R}^J$ to fall into the left child of such node. Since distribution functions are monotone non-decreasing, a higher scalar product between the observation vector and parameters vector of the fixed node, means a higher probability of descending leftwards from that internal node. One of the main proposals of this article is to set the $\mathbf{z}_i,\ i \in \{1, \ldots, N\}$ as the scores with respect to the estimated FPCA *basis*:

$$\mathbf{z}_i = \begin{bmatrix} \langle \mathbf{x}_i,\, \tilde{\xi}_1 \rangle \\ \vdots \\ \langle \mathbf{x}_i,\, \tilde{\xi}_J \rangle \end{bmatrix} \in \mathbb{R}^J \tag{9}$$

and thus, by linearity of the scalar product, Equation (6) becomes:

$$p_{it}\big(\mathbf{z}_i; (\mathbf{a}_t, m_t)\big) = F(\mathbf{z}_i^\intercal \mathbf{a}_t - m_t) = F(\sum_{j=1}^{J} \Big\{ \langle \mathbf{x_i},\, \tilde{\xi}_j \rangle a_{jt} \Big\} - m_t) = F(\sum_{j=1}^{J} \Big\{ \langle \mathbf{x_i},\, \tilde{\xi}_j a_{jt} \rangle \Big\} - m_t) \tag{10}$$

Since we are using scalar products in $\mathbb{R}^p$, with p sufficiently large to approximate scalar products in $L^2(\mathcal{I}; \mathbb{R})$, we can denote as splitting functions the set $\{f_t(s)\}_{t \in \tau_B}$, given by

$$f_t(s; \mathbf{a}_t) = \sum_{j=1}^{J} a_{tj} \hat{\xi}_j(s) \tag{11}$$

The previous reasoning implies that at node $t \in \tau_B$, the higher the scalar product between function $x_i(s)$ and splitting function $f_t(s)$, the higher the probability of the ith datum of going to the left child of that same node. This results in a direct interpretation feature of our model: at each internal node $t \in \tau_B$, the more correlated (in the L^2 sense) with the splitting function $f_t(s)$ a (newly) observed function is, the higher the probability it will descend along the graph of the tree through the left child of such node. Inspecting such

(estimated) functions grants a direct method to knowing which aspects of the variability of functional data set the model has learnt and is exploiting to perform inference on new functional data. This is demonstrated in the case studies of the current manuscript, viz. Sections 3.3.2 and 3.3.1.

2.2.3 Model fitting: interior point methods

The robustness of the solution of an non-convex optimisation solver method is directly affected by the choice of a starting point, with poor choices leading to failure of convergence, and better choices for reaching of stationary point of better objective function value. For interior point methods, it is common practice to solve the same problem from a range of different initial coordinates, see Nocedal and Wright (1999); Nocedal, Wachter, and Waltz (2009)'s works. Blanquero et al. (2021), for e.g., propose to solve the ORCT problem 20 times, keeping the result with the best objective function value. Given the high non-convexity of objective functions (7a) and (8), the choice of initialisation procedure affects the variability in the solution obtained by the optimisation method, such that it calls upon excogitating an *ad-hoc* initialisation scheme. Ideally, the objective function value should be robust with respect to different starting points, and these should be such that they guide the (Interior Point) solver to attain better objective function values, as done by Nocedal et al. (2009).

Naturally, a trivial initialisation scheme would generate random starting points subject to constraints (7b)–(7c). In the code we provide as supplement to the present manuscript, we develop such scheme, utilising distribution $\mathcal{U}(0,1)$ for $\mathbf{a}_t$, $t \in \tau_B$ and m_t, $t \in \tau_B$; and independent K-dimensional Dirichlet distributions for $\mathbf{c}_t$, $t \in \tau_L$ together with an earth-moving algorithm to ensure (7b). We hereafter refer to this initialisation as the *trivial* one.

A non-trivial initialisation scheme has been provided by Consolo, Amaldi, and Manno (2025) for soft regression trees. Utilising the input data matrix $\mathbf{Z}$, at each $t \in \tau_B$ k-means clustering is used to discover two clusters (left and right). Once the left and right clusters are defined, a logistic regression is fit, whence $\mathbf{a}_t$ and intercepts m_t are retrieved. The observations used at each internal node for the clustering depend on those which descended left or rightwards in the parent node, as a result of the unsupervised cluster assignment, and with all observations being used in the node t_0 at depth 0, i.e. the first one in the tree. We modify such method to output random starting points subject to (7b)-(7c). The full algorithm is provided in code supplementary to the manuscript.

As custom for Interior Point methods, the problem is solved from a number of starting points, keeping the variables associated to the best solution. For each of these initial solutions, Consolo et al. (2025) run the initialisation procedure a number of times, and select the best starting point according to some criterion. They utilise the silhouette score at each leaf node to select the best initial solution to give as input to the solver. In the current setting, where also the labels $y_i, i \in \{1, \dots, N\}$ are available, the Gini index is also utilised. Hereafter we refer to these non-trivial initialisations as *km-silhouette* and *km-gini*.

In the current work, we also propose an *ex novo* procedure, based on FLDA (James and Hastie 2002), which exploits both the scores matrix $\mathbf{Z}$ and the labels vector $\mathbf{y} = \{y_i\}_{i=1}^N$. A first design choice is that at each $t \in \tau_B$, the considered observations, indexed by $I_t \subset \{1, \dots, N\}$, are bootstrapped (i.e. a sample with replacement is taken), and $\left\lfloor \sqrt{J} \right\rfloor$ of the input data's columns are chosen at random. This last decision is aligned with Breiman (2001a)'s Random Forests, as implemented by Pedregosa et al. (2011).

Then, FLDA is fitted on the subsampled columns and a bootstrap sample of the available observations in such node I_t. The unit vector corresponding to the first linear discriminant $\mathbf{w}_t$ is used for the value of $\mathbf{a}_t$. The rationale behind for such setup is that in (F)LDA,

the first linear discriminant maximises the separability of observations with respect to the labels $\mathbf{y}$ (Johnson et al. 2002). The procedure is followed for all $t \in \tau_B$. Note that the $J - \left\lfloor \sqrt{J} \right\rfloor$ columns which were not subsampled at the current node are not included in the FLDA and they are set to zero, whence still $\mathbf{w}_t \in \mathbb{R}^J$.

On the other hand, to obtain m_t, the normalised mean of the FLDA scores with respect to $\mathbf{w}_t$ are computed. That is, the mean of vector $\mathbf{Z}_{(t)} \cdot \mathbf{w}_t$ is calculated and divided by J, where $\mathbf{Z}_{(t)} = \left\{\mathbf{z}_i^\intercal\right\}_{i \in I_t}$ This ensures that the inputs to the logistic distribution function in the calculation of (6) are centred. Such centring, due to the symmetry of the said link function in (6), causes the half of the present observations I_t to go to the left child node of t ($p_{it} > 0.5$ for half of $i \in I_t$), so that as the internal nodes $t \in \tau_B$ are initialised, none of them end up with $I_t = \{\varnothing\}$. As a result, a rule for defining the observations used for the FLDA at each internal node, I_t, is provided. Naturally, $I_{t_0} = \{1, \ldots, N\}$, where t_0 is the first node of the tree.

Finally, for the choice of the c_{kt}, $k \in \{1, \ldots, K\}$, $t \in \tau_L$ the class with the highest number of observations available (according to this initialisation procedure, $I_t,\ t \in \tau_L$) is set as 1 and else 0. As with the *trivial* initialisation, and *ad-hoc* minimum earth moving algorithm is utilised in case constraints are not respected. The Gini index is utilised to select the initial solution after having run this algorithm a specified number of times.

Since the initialisation scheme for the interior point method exploits in turn another model for classification, we deem it to be statistically-driven.

3 Results

This Section is devoted to assess the classification ability of our proposal. On the one hand, we want to evaluate the *POST-FD* classifier across different data sets by means

of a suitable classification metric and of its interpretability. On the other hand, that the critical choices made, viz. the penalisation in (8) and the FLDA-inspired initialisation are justified.

We recall (Section 2.2.3) that when using interior point solvers for non-convex objective functions, a variety of starting points are provided. Given the high-non convexity of (8), good quality initial solutions can enhance model fitting by aiding for the (Interior Point) solver to yield good quality final solutions. That is, better starting points should lead to better objective function values, with less variability. Moreover, since the penalty term in the objective function promotes homogeneity in leaf nodes, the model's ability to avoid overfitting should be improved. Lastly, as outlined in Section 2.2.2, its interpretability features should be helpful to provide qualitative inference upon utilisation.

3.1 Simulation setting

The performance of the proposed classifier is evaluated through numerical experiments carried out on different data sets with different characteristics, as argued by e.g. Friedman (2001) or Baíllo et al. (2011). Firstly, in Section 3.2, we test it on standard benchmark data sets for classification in FDA. Next, we show in Section 3.3 its relevance by applying it in two case studies, where the task of classification is notably harder than for the previous known data sets.

For all experiments, the setup is the following. The classification accuracy on both training and test set are obtained through 5-fold cross validation, ensuring that at each fold the proportion of data belonging to each class is approximately the same (stratified k-fold cross validation). The optimisation problem is solved from 20 starting points, and the distribution across the 5 folds of the training set and test set classification accuracy are the main result under study, with a total of 20×5 runs obtained in the experiment

per training and per test data set, given the set necessary specifications of the tree. The depth is fixed at $D = 2$, which allows both for each class to be represented in at least one leaf node, and to construct an easy way to interpret the estimated tree.

Two hyper-parameters of the model are changed to explore the empirical distributions of the performance measures. The first one is the weight of the penalisation in (8) given by α, which is explored for $\alpha = 0$, i.e. the case without penalisation and 10 different values which are a common choice in the regularised learning setting:

$$\alpha \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 10^{0}, 5, 10^{1}\}$$

These are usually incumbent values for the penalty hyper-parameter, cfr. the implementations by Pedregosa et al. (2011) and Friedman et al. (2010). Values of α are in the abscissa of Figures 3, 4, 6 and 9.

The other hyper-parameter is the algorithm to provide those 20 starting points. We utilise 4 different methods, namely the ones presented in Section 2.2.3: the *trivial* initialisation, *km-silhouette*, *km-gini* and the FLDA-inspired heuristic. In the fashion of the experimental setting used by Consolo et al. (2025), for each of the 20 starting points, the initialisation heuristic is run 40 times, for each of which a single initial solution is selected according to the criterion given by the procedure The type of initialisation varies for each column of plots in Figures 3, 4, 6 and 9. The analysis of the model's interpretability, readily available by the estimated splitting functions (11), is carried out only in the case studies, as the difficulty and importance of these applications elicit more interest.

Another case-specific hyper-parameter of the model is the number J of selected Functional Principal Components (FPCs), which in turn determines the number of variables in the optimisation and thus has an impact on the difficulty to obtain a solution. J was chosen according to standard scree plot evaluations, as done by James and Hastie (2002).

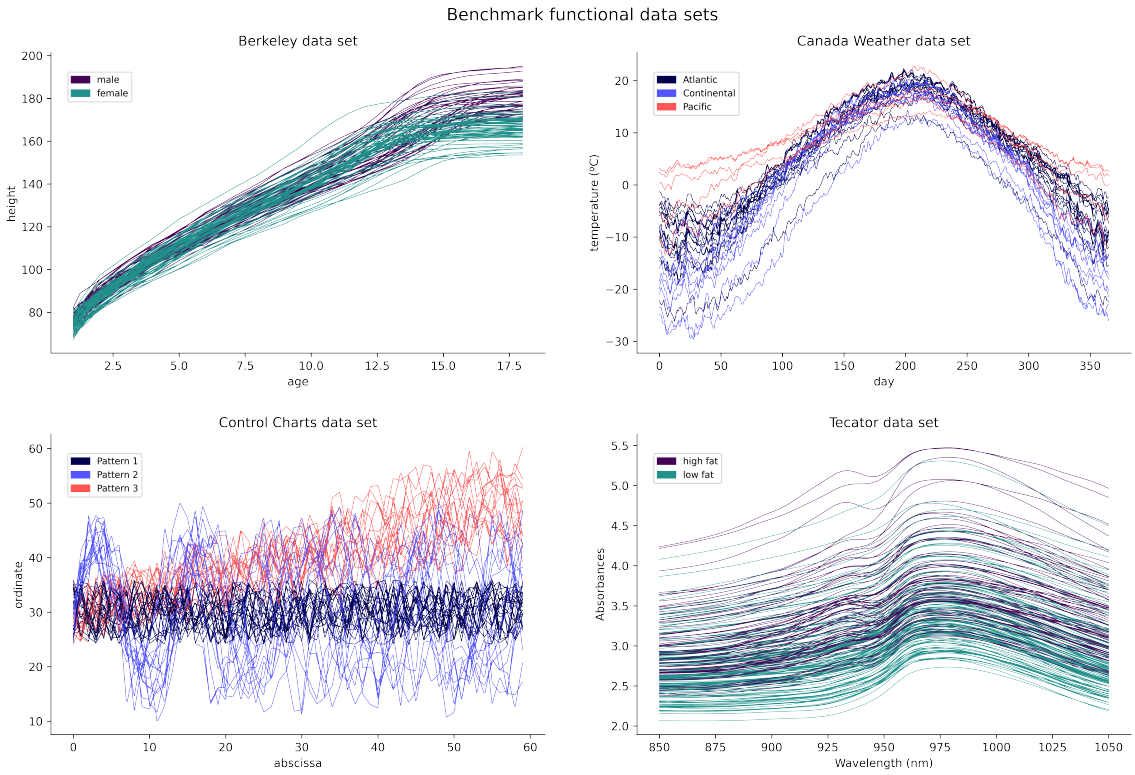

Figure 2: Benchmark data sets for functional classification.

Code for the algorithm and the different initialisation heuristics, as well as the experiments is available in Supplementary Material.

3.2 Benchmark data sets

The first set of experiments are carried out with real-world data which are widely used to illustrate classification techniques in the FDA setting, viz. Berkley Growth (Tuddenham 1954), Canada Weather, Tecator, as pointed out by Ferraty (2006); to which we add Alcock

Table 1: Summary of the benchmark data sets

Name	N	K	Class numerosity	Chosen J
Canada Weather	29	3	(15,19,15)	5
Berkeley Growth	40	2	(23,17)	5
Control Charts	50	3	(19,16,15)	8
Tecator	215	2	(77, 138)	4

et al. (1999)'s Control Charts simulated data set. The first three are provided by Ramos-Carreño et al. (2024)'s *Scikit-fda* Python library, whereas the latter is made available online by Dua and Graff (2019).

Some *ad hoc* modifications are made to render them suitable for our experiments. As done by Ramos-Carreño et al. (2024) for the Canada Weather dataset the *Arctic* class was discarded due to its scarce ($N = 3$) numerosity. For the Control Charts data set, only three classes were kept, subsampling functions at random to retain in total $N = 50$, keeping both the sample cardinality N and the number of different labels K in line with other benchmark data sets. The data, illustrated in Figure 2, are summarised in Table 1.

For the sake of space in this Section we present the results only for Canada Weather. Similar results have been obtained for the other three benchmark data sets and are reported in Supplementary Material. In Figure 3, the distributions induced by 5-fold cross validation and using 20 different starting points, for different values of α and initialisations are displayed. The plots on the first row correspond to the accuracy on the training set: it is immediately recognisable that when a non-trivial initialisation is employed and α grows, the accuracy increases, until it seemingly reaches a saturation point of α before

it starts to descend again. This implies that employing the unsupervised term, there is an improvement for the performance on the training set in the supervised task of classification. Regarding the performances on the test sets, α shows to mitigate overfitting for all cases except that of the trivial initialisation.

Complementarily to these results, the values of the expected misclassification cost (i.e. the first term in objective function (8)) obtained at the end of the optimisation procedure for the same experiments whose accuracies are displayed in Figure 4. It is worth noting that including the unsupervised term ($\alpha > 0$) the first term (viz. the misclassification cost) in (8) achieves better objective function values. Therefore, the idea of including an unsupervised term in the objective function not only mitigates the overfitting, but also leads to better solutions (and as a consequence also a better performance on the training set) in the optimisation of the problem.

Concerning the choice of the initialisation method, it is evinced from Figure 3 that the FLDA initialisation shows the best results on benchmark data sets both on training and test set accuracies (column 4, $\alpha = 10^{-2}$ for Canada Weather in Figure 3), with better values than *km-gini*, *km-silhouette* and of course *trivial* procedures. Indeed, not only the median and mean classification metrics are better, but the variability in their values with respect to different starting points is lower. Inspecting that same figure, the k-means based initialisations, the change in the accuracies as the weight of the unsupervised term varies (as $\alpha > 0$ increases) is more pronounced. A possible explanation could be that since the chosen basis expansion is FLDA, suggesting as initial splits the subsampled and bootstrap-aggregated FLDA directions is in direct alignment with the optimisation task. Methods *km-silhouette* and *km-gini*, which perform similarly to each other, do not possess such a strong link with the FLDA *basis*, rely more on the synergy between the supervised and the unsupervised term: the latter homogenises leaf nodes, and since functions of the same class

should be homogeneous according to some criterion, the misclassification cost (supervised term) is lowered by including the unsupervised term ($\alpha > 0$).

Another aspect of importance is the actual performance of the *POST-FD* classifier, once the 20 different solutions have been obtained after optimisation. As mentioned in Section 2.2.3, the solution with the best objective function value is the one that is utilised for inference. Hence, the accuracy values present at each boxplot in Figure 3 contain the values of all 20 solutions for each of the 5 folds, whereas in practice only the accuracy corresponding to the best objective function value would be the one the model attains. To illustrate this effect, Table 2, shows the accuracies on training and test set of the best solution, averaged across the 5 folds, as both the initialisation method (each column) and α vary. The accuracies, both in the training and test set are higher than when considering all the 20 solutions per fold (Figure 3). As expected by the inspection of Figure 3, the best accuracy is obtained with the FLDA initialisation, with $\alpha = 10^{-3}$. That happens because the objective function is the penalised misclassification cost in just the training set, and the unsupervised term helps mitigate overfitting. As a reference, the 5-fold mean of the accuracies of a quadratic FLDA classifier were $(100\%, 79\%)$.

3.3 Case studies

3.3.1 Knee injury data

Tengman, Grip, Stensdotter, and Häger (2015) investigate reduced dynamic knee stability in one-leg hops on patients who had suffered an anterior cruciate ligament (ACL) injury. In the past, several authors have studied the problem with tools from the FDA setting, see for e.g. Hébert-Losier et al. (2015)'s, and more recently Abramowicz et al. (2018)'s publications. The latter proposes a functional-on-scalar regression model, where the regressed

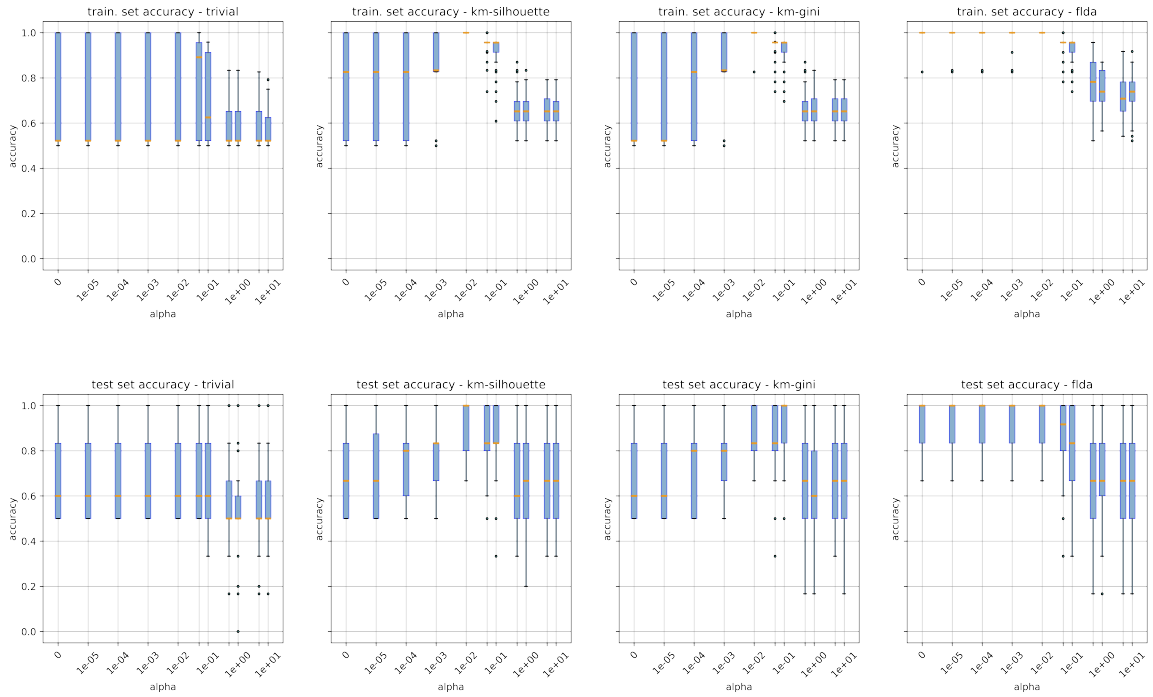

Figure 3: Canada Weather: 5-fold cross validation accuracy in train and test set with different initialisation and α values.

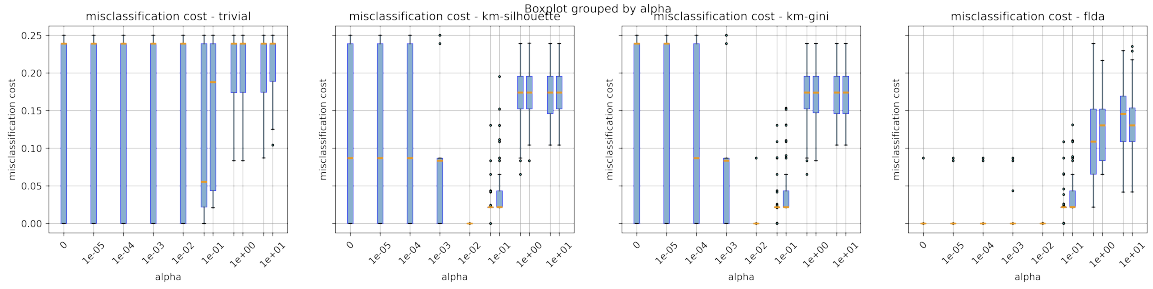

Figure 4: Canada Weather: 5-fold cross validation of only the first term in (8) (expected misclassification cost). If the proposed unsupervised term is included ($\alpha > 0$), better objective function values for the supervised term are obtained.

Table 2: 5-fold mean of the accuracies of the fitted trees with best objective function values

alpha	**trivial**	**km-silhouette**	**km-gini**	**flda-gini**
0	(1.000, 0.893)	(1.000, 0.893)	(1.000, 0.893)	(1.000, 0.860)
10^{-5}	(1.000, 0.860)	(1.000, 0.893)	(1.000, 0.893)	(1.000, 0.860)
10^{-4}	(1.000, 0.893)	(1.000, 0.893)	(1.000, 0.893)	(1.000, 0.827)
10^{-3}	(1.000, 0.893)	(1.000, 0.893)	(1.000, 0.893)	**(1.000, 0.927)**
10^{-2}	(1.000, 0.827)	(1.000, 0.860)	(1.000, 0.860)	(1.000, 0.860)
0.05	(0.957, 0.827)	(0.957, 0.827)	(0.957, 0.827)	(0.957, 0.860)
10^{-1}	(0.957, 0.900)	(0.957, 0.827)	(0.957, 0.860)	(0.931, 0.833)
0.5	(0.655, 0.567)	(0.620, 0.627)	(0.630, 0.580)	(0.664, 0.627)
10^{0}	(0.612, 0.473)	(0.612, 0.667)	(0.629, 0.667)	(0.655, 0.660)
5	(0.604, 0.447)	(0.595, 0.620)	(0.621, 0.767)	(0.621, 0.560)
10	(0.612, 0.547)	(0.603, 0.733)	(0.620, 0.667)	(0.586, 0.553)

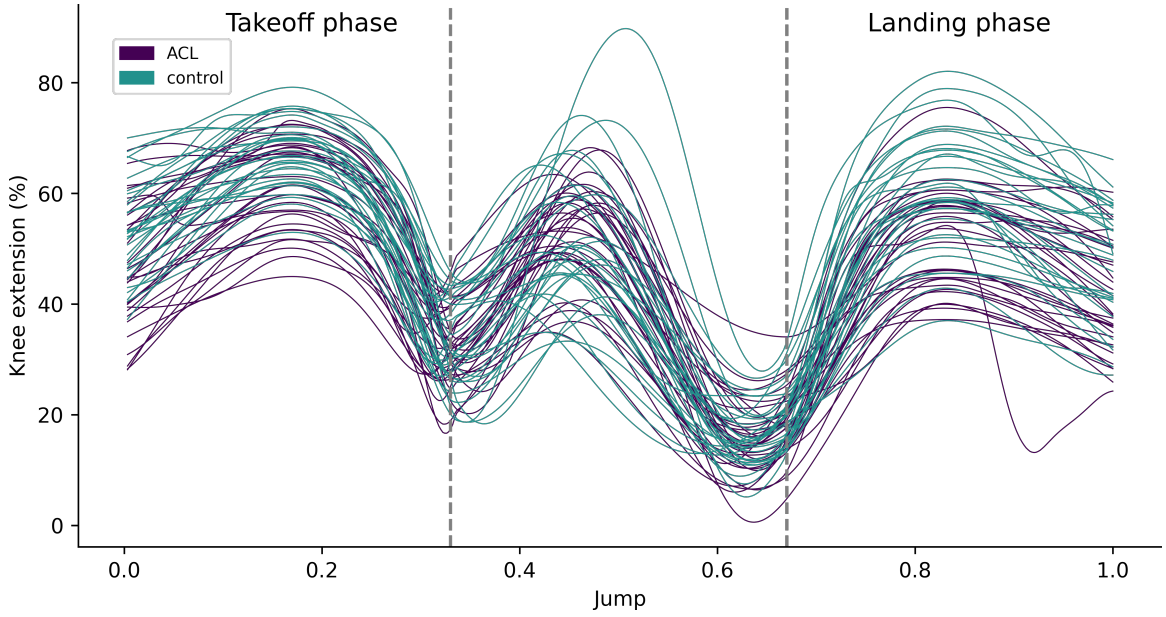

Figure 5: Knee jump data set.

variables are the functions describing one-leg jumps of different patients, so that in the abscissa there is the time of the jump ($\mathcal{I} = [0, 1]$) and on the ordinate the percentage of knee extension. In particular, they perform interval-wise permutation tests for the significance of a variety of factors. In this work, we focus on the group each patient belongs to (patients who suffered an ACL injury 23 ± 2 years ago, which we denote ACL, the control group who have never suffered such injury, denoted *control*; hence $K = 2$ classes), since in that study it was determined to be a statistically significant factor, whence the sense to perform classification using it as a label.

Since the above-mentioned interval-wise procedure not only detects significance but also on which part of the domain the factor is significant, we subdivide the data set into two. By taking two subsets of the dominion of the whole jump, two classification data sets are obtained by focusing on the takeoff phase ($\mathcal{I}_1$) and the landing phase ($\mathcal{I}_2$). This choice is

illustrated in Figure 5. We also consider the functional data set analysed by Abramowicz et al. (2018), where the ACL patients are subsampled at random so that the numerosity of both labels is the same, avoiding the nuances of imbalanced classification: $N = 62$ with $\big|\{i : y_i = control\}\big| = \big|\{i : y_i = ACL\}\big| = 31$

The data, as provided in the *fdahotelling* R package (Stamm 2017), have already been smoothed with a B-spline basis with 150 equally spaced knots. Each functional datum is evaluated on an equally-spaced grid of $p = 298$ points throughout $\mathcal{I}$. FPCA is performed on both (sub) data sets, keeping $K = 5$ throughout. Concerning the dissimilarity, motivated by the fact that the model fit in Abramowicz et al. (2018) estimates the functional regressor for the factor $\mathbb{I}\{y_i = control\}$ as a constant function, the L^2 distance is chosen.

Figure 6 shows the results for the jumping phase $\mathcal{I}_1$. Those for the landing phase are available in Supplementary Material. A first remark is that just as in Section 3.2, the adoption of a non trivial initialisation strategy yields better results both in the training and testing set. Similarly to the benchmark data experiments, the train set accuracy increases as α increases until a saturation point is reached, which justifies the inclusion of the unsupervised term in (8). Results are less pronounced regarding the performance in the test set. Yet there still exists some $\alpha > 0$ such that the distribution of the test set accuracy is better than without the unsupervised term, in particular for values $\alpha \in \{10^{-1}, 0.5, 10^{0}\}$ in Figure 6.

We recall that the accuracy values correspond to all 20 solutions per fold, whilst in practice only those corresponding to the best objective function value would be used for inference, as seen in Section 3.2, Table 2.

Another important aspect to analyse is the interpretability. As explained in Section 2.2.2, an advantage of choosing the FPCA as *basis* is that the splitting functions inherit their interpretation as perturbations of the (functional) mean of the original data sample.

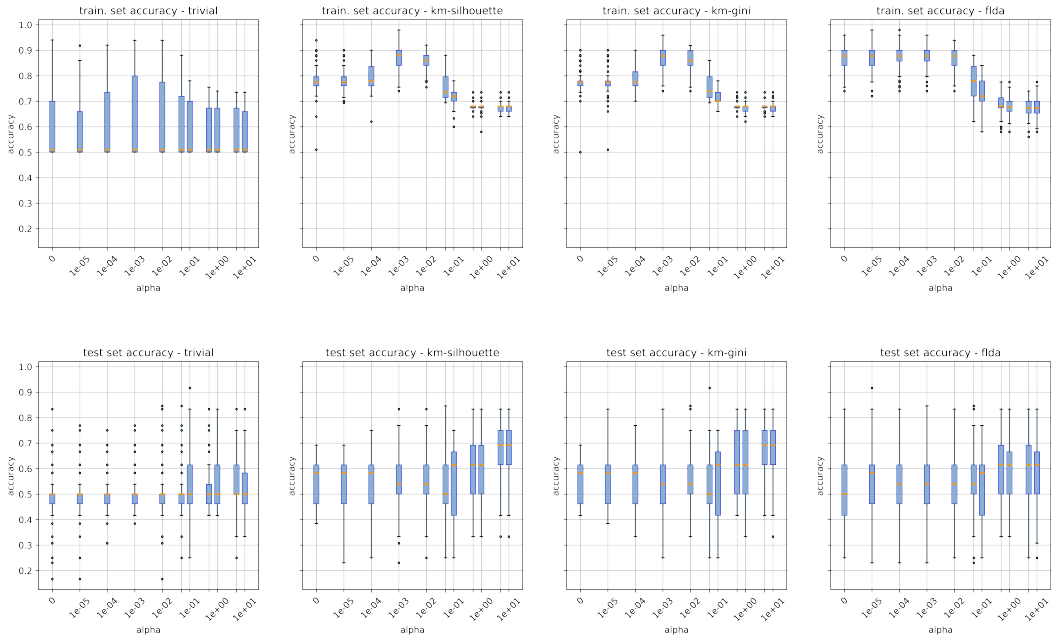

Figure 6: Results for the jumping phase classification.

Figure 7 displays, for the root node t_0, the estimated splitting function (marked by the "+" sign) added to the sample mean (continuous line), corresponding to the best (in terms of objective function value) of the 20 runs for the tree with $\alpha = 10^{-3}$. It shows that functions that tend to have an overall constant level above the mean will have a higher value as input to the link function (6) and hence a higher probability of descending leftwards of such node. This is in direct agreement towards Abramowicz et al. (2018)'s estimation of the regressor of $\mathbb{I}\{y_i = control\}$ being mostly a positive constant function in the jumping phase.

3.3.2 Aneurisk data

The AneuRisk project (cfr. Sangalli et al. (2014)'s data description), has the goal of finding factors that indicate the formation of cerebral aneurysmata. The data were collected at

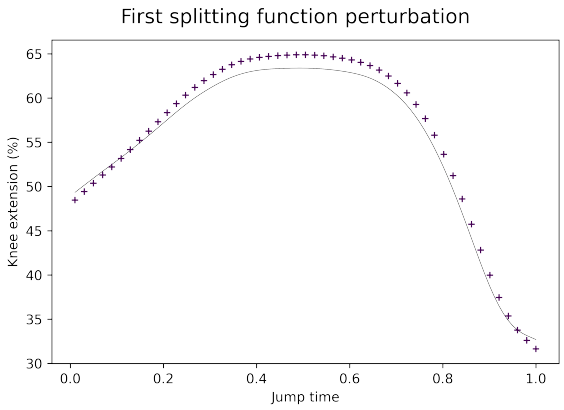

Figure 7: Estimated splitting function with $\alpha = 10^{-1}$ in the jumping phase

Ospedale Niguarda Ca' Grande Milano between 2002 and 2005 from 65 patients deemed at risk for an aneurysm in the Internal Carotid Artery (ICA). Geometrical measurements such as radius, curvature, wall shear stress, amongst others concerning the artery were made, and they have in turn been preprocessed and registered by Passerini et al. (2012) and (also explored in detail by means of FDA tools) by Sangalli et al. (2009). For the present case study, we follow the same setting as Pini et al. (2018), *videlicet* only 50 of the original 65 patients are analysed, and only the radius along the last 5cm of the ICA. Moreover, two classes of patients ($K = 2$) are considered: high-risk patients, due to the presence of an aneurysm in the skull, and low-risk patients, with either an aneurysm outside the skull or none at all. The numerosity per class type coincides at 25 each. The data are displayed in Figure 8.

Besides the well-established relevance of the AneuRisk project, the choice of such application for the present work is motivated by the fact that both Passerini et al. (2012) and Pini et al. (2018) demonstrate that the two groups present significance whilst testing for differences in the functional distributions of the radius of the ICA, thus justifying the

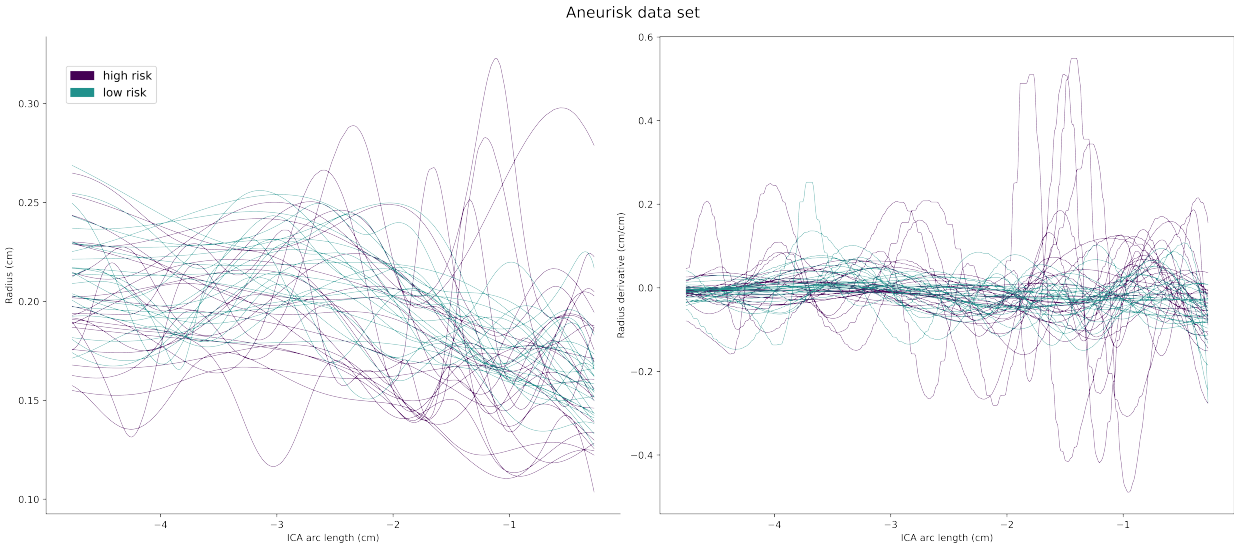

Figure 8: The AneuRisk data set: radius and the derivative of radius of the last 5cm of the ICA

treatment of these data as a classification data set.

Upon inspection of the derivative (cfr. Figure 8), the (absolute) variation across both groups could be hypothesised to be different across groups. Therefore, for data $\{l, j \in \{1, \ldots, N\} : l \neq j\}$ we chose, after assuming that x_i belong to Sobolev space $\in H^1(\mathcal{I})$ (so that the first derivative lies in L^2)

$$d_{jl} = \int_{-5}^{0} |x'_j(r) - \bar{x}'_j| dr \; - \int_{-5}^{0} |x'_l(r) - \bar{x}'_l| dr \tag{12}$$

where $x'_j(r)$ denotes the derivative of the jth datum, and $\bar{x}'_j$ is the mean value of the derivative of such datum, i.e. $\bar{x}'_j = \int_{-5}^{0} x'_j(r) dr$. The calculations of such integrals were approximated through operations in $\mathbb{R}^J$, as outlined in Section 2.2, to avoid the computational bottleneck.

Results can be visualised in Figure 9. As in the previous case study and benchmark experiments, the inclusion of the unsupervised term (when $\alpha > 0$) is justified. Indeed, the

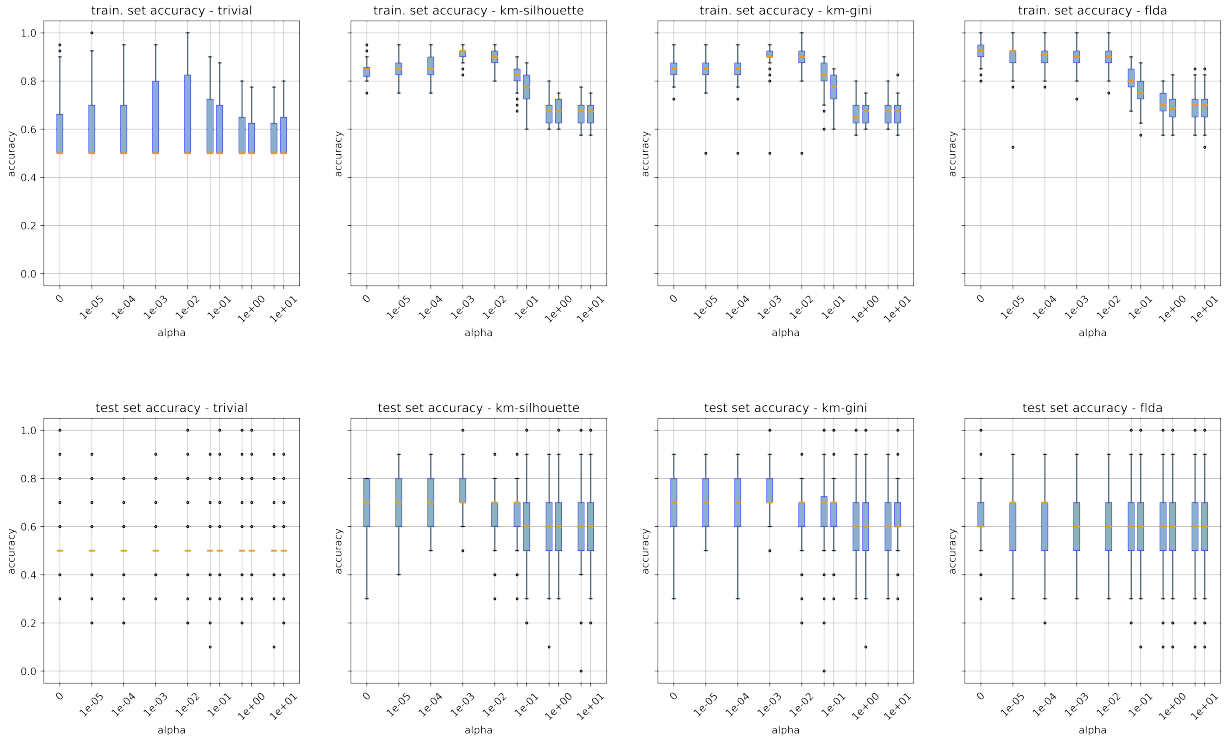

Figure 9: Results for the classification problem with the AneuRisk data

same pattern of an increase in the performance in the train set as α increases is observed. Regarding the test set, again the effect of the unsupervised term as a penalisation that mitigates overfitting is verified: there exist values of $\alpha > 0$ for which there are in mean higher accuracies in the test set. Regarding the different initialisation heuristics, in this case the *k-means-Gini*, seems to perform best (see in particular when $\alpha = 10^{-3}$), since both training and test set accuracies are higher than for all other values of α. Similar results happen for *km-silhouette*, which is to be expected given that the only difference between these two methods is the selection criterion for each starting point. The improvement yielded by the unsupervised term $\alpha > 0$ seem less pronounced in this case, yet the means in test set accuracy for $\alpha \in \{10^{-5}, 10^{4}\}$ are higher than the case without the unsupervised term

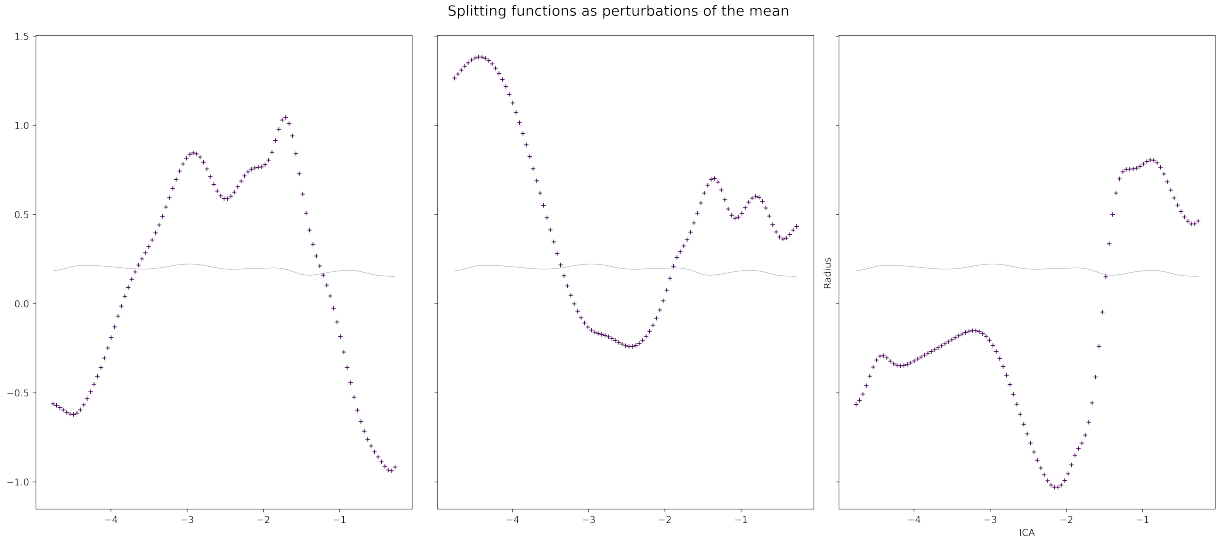

Figure 10: Estimated splitting functions for classifying the AneuRisk data, $\alpha = 10^{-3}$

($\alpha = 0$). A possible explanation could be that since the separation is aided by promoting homogeneity in the variation of the derivative, the alignment between the optimisation task and the FLDA initialisation (remarked in Section 3.2) becomes feebler.

Figure 9 also displays high variability in the train and test set accuracies, with some values below 30% as $\alpha > 10^0$ seems to be too large. Such values would be filtered out by taking the best (in terms of objective function value) run for each fold, as outlined in the benchmark data experiments.

Concerning the interpretability of our soft tree for functional data, Figure 10 displays the splitting functions of the three first nodes of the trees as perturbations of the mean. Interestingly, the the third one is directly associated to the finding by Pini et al. (2018) regarding the therein called Clinical Question 2: patients at lower risk are those who posses ICA with enough variation in the sense that just before reaching the brain, the radius becomes wider to avoid slowing down blood flow upon entrance to the Circle of

Willis. The other splitting functions could be of aid to domain experts, and the absence of a clear interpretation could provide grounds for pruning the tree, i.e., to decrease its depth.

4 Conclusion

We have proposed a new objective function for globally-optimised Soft Classification Trees that contains both a supervised and an unsupervised term, designed to perform classification of Functional Data utilising an FPCA *basis*. The second term penalises heterogeneity in the leaf nodes of the tree, requiring the choice of a dissimilarity statistic between functions that allows adaptation for case-specific modelling. Fitting is performed through Interior Point methods, and we have also proposed novel initialisation heuristics that are statistically driven in order to obtain better objective function values. The model features interpretable splitting functions at each internal node of the tree.

We have evaluated our proposed model with a simulation study on benchmark data sets for classification in the FDA setting, as well as two case studies. The results have demonstrated that the inclusion of the unsupervised term works synergically with the non-trivial initialisations, leading to better objective function values and mitigating overfitting. Indeed, for all benchmark data sets and in both the case studies, scenarios where $\alpha > 0$ (i.e. the unsupervised term) led to better in training and test sets. Another positive effect was aiding the optimisation to converge to objective function values of the supervised term which are better thanks to the inclusion of the unsupervised term. In addition, the adaptiveness thanks to the choice of the dissimilarity, as well as the interpretability of the yielded splitting functions have been shown to be coherent to other works that worked with the same cases studies.

Future work aimed at improving the *POST-FD* classifier could be to guide the choice

of the model's hyperparameters. Common to most regularised algorithms to our knowledge, the weight α of the unsupervised (penalty) term could be trivially chosen by standard hyperparameter tuning techniques (cfr. Yang and Shami (2020)), yet further insights regarding the correct balance between the supervised and the unsupervised term in (8) could be pursued. A possibility would be to modify the Interior Point algorithm to dynamically update the α value, depending on the objective function value through, for e.g. trust region methods (Nocedal and Wright 1999). Regarding the choice of depth D, inspecting the estimated splitting functions could be a possibility, but drawing insights from the pruning methodologies for tree learning as done by Mingers (1989) might be worth exploring.

Another possible extension could be towards non-differentiable dissimilarities. Indeed, the statistic that Pini et al. (2018) use to significantly separate (by means of a permutation test) the two groups of the AneuRisk data set (Section 3.3), which is non-smooth due to the sup operator, would have been an interesting choice, yet since Interior Points methods rely on the gradient, alternatives would be needed. Lastly, (asymptotic) theoretical results concerning the classification capability of the model, as well as analytical knowledge regarding the role of the homogeneity penalisation introduced in the unsupervised term could be fruitful.

Acknowledgements

We wish to thank the U-motion laboratory at Umeå University and Charlotte K. Hager for providing the knee flexion data. We also appreciate Alessia Pini for her suggestions regarding interesting data for the case studies; as well as Aymeric Stamm for his help with the usage of the Aneurisk data.

SUPPLEMENTARY MATERIAL

Github respository for POST-FD: Github repository containing source code and all simulation results (https://github.com/alfredo-g-zapiola/POST-FD)

References

Abramowicz, K., C. K. Häger, A. Pini, L. Schelin, S. Sjöstedt de Luna, and S. Vantini (2018). Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics 45*(4), 1036–1061.

Alcock, R. J., Y. Manolopoulos, et al. (1999). Time-series similarity queries employing a feature-based approach. In *7th Hellenic conference on informatics*, pp. 27–29.

Amaldi, E., A. Consolo, and A. Manno (2023). On multivariate randomized classification trees: l0-based sparsity, vc dimension and decomposition methods. *Computers & Operations Research 151*, 106058.

Baíllo, A., A. Cuevas, and R. Fraiman (2011). *Classification methods for functional data*, pp. 259–293. Oxford University Press.

Belli, E. and S. Vantini (2022). Measure inducing classification and regression trees for functional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal 15*(5), 553–569.

Berrendero, J. R., A. Cuevas, and J. L. Torrecilla (2018). On the use of reproducing kernel hilbert spaces in functional classification. *Journal of the American Statistical Association 113*(523), 1210–1218.

Bertsimas, D. and J. Dunn (2017, July). Optimal classification trees. *Machine Learning 106*(7), 1039–1082.

Blanquero, R., E. Carrizosa, C. Molero-Río, and D. R. Morales (2021). Optimal randomized classification trees. *Computers & Operations Research 132*, 105281.

Blanquero, R., E. Carrizosa, C. Molero-Río, and D. R. Morales (2022). On sparse optimal regression trees. *European Journal of Operational Research 299*(3), 1045–1054.

Blanquero, R., E. Carrizosa, C. Molero-Río, and D. Romero Morales (2020). Sparsity in optimal randomized classification trees. *European Journal of Operational Research 284*(1), 255–272.

Blanquero, R., E. Carrizosa, C. Molero-Río, and D. Romero Morales (2023). On optimal regression trees to detect critical intervals for multivariate functional data. *Computers & Operations Research 152*, 106152.

Breiman, L. (2001a). Random forests. *Machine learning 45*, 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science 16*(3), 199–231.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). Classification and regression trees–crc press. *Boca Raton, Florida 685*.

Brézis, H. (2011). *Functional analysis, Sobolev spaces and partial differential equations*, Volume 2. Springer.

Consolo, A., E. Amaldi, and A. Manno (2025). Soft regression trees: A model variant and a decomposition training algorithm. *arXiv preprint arXiv:2501.05942*. Version 2, last revised 27 Jan 2025.

Cuevas, A., M. Febrero, and R. Fraiman (2006). On the use of the bootstrap for estimating functions with functional data. *Computational statistics & data analysis 51*(2), 1063–1074.

Cuevas, A., M. Febrero, and R. Fraiman (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics 22*(3), 481–496.

Delaigle, A. and P. Hall (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology 74*(2), 267–286.

Delaigle, A. and P. Hall (2013). Classification using censored functional data. *Journal of the American Statistical Association 108*(504), 1269–1283.

Dua, D. and C. Graff (2019). UCI machine learning repository.

Ferraty, F. (2006). *Nonparametric functional data analysis*. Springer.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Galeano, P., E. Joseph, and R. E. Lillo (2015). The mahalanobis distance for functional data with applications to classification. *Technometrics 57*(2), 281–291.

Hastie, T., R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman (2009). Boosting and additive trees. *The elements of statistical learning: data mining, inference, and prediction*, 337–387.

Hébert-Losier, K., A. Pini, S. Vantini, J. Strandberg, K. Abramowicz, L. Schelin, and C. K. Häger (2015). One-leg hop kinematics 20 years following anterior cruciate ligament rupture: data revisited using functional data analysis. *Clinical Biomechanics 30*(10), 1153–1161.

Hlubinka, D., I. Gijbels, M. Omelka, and S. Nagy (2015). Integrated data depth for smooth functions and its application in supervised classification. *Computational Statistics 30*, 1011–1031.

Hsing, T. and R. L. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*, Volume 997. Wiley Online Library.

Hyafil, L. and R. L. Rivest (1976). Constructing optimal binary decision trees is np. cop. pr. *Inf. Proc. Lett.*, 3 (1), 79,87.

James, G. M. and T. J. Hastie (2002, 01). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B: Statistical Methodology 63*(3), 533–550.

Johnson, R. A., D. W. Wichern, et al. (2002). Applied multivariate statistical analysis.

Kneip, A. (1994). Nonparametric estimation of common regressors for similar curve data. *The Annals of Statistics*, 1386–1427.

Maturo, F. and R. Verde (2023). Supervised classification of curves via a combined use of functional data analysis and tree-based methods. *Computational Statistics 38*(1), 419–459.

Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine learning 4*(2), 227–243.

Nocedal, J., A. Wachter, and R. A. Waltz (2009). Adaptive barrier update strategies for nonlinear interior methods. *SIAM Journal on Optimization 19*(4), 1674–1693.

Nocedal, J. and S. J. Wright (1999). *Numerical optimization*. Springer.

Passerini, T., L. M. Sangalli, S. Vantini, M. Piccinelli, S. Bacigaluppi, L. Antiga, E. Boccardi, P. Secchi, and A. Veneziani (2012). An integrated statistical investigation of internal carotid arteries of patients affected by cerebral aneurysms. *Cardiovascular Engineering and Technology 3*(1), 26–40.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Pini, A., A. Stamm, and S. Vantini (2018). Hotelling's t2 in separable hilbert spaces. *Journal of Multivariate Analysis 167*, 284–305.

Ramos-Carreño, C., J. L. Torrecilla, M. Carbajo Berrocal, P. Marcos Manchón, and A. Suárez (2024, May). scikit-fda: A Python Package for Functional Data Analysis. *Journal of Statistical Software 109*(2), 1–37.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis*. Springer, New York.

Ramsay, J. O. and C. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology 53*(3), 539–561.

Riccio, D., F. Maturo, and E. Romano (2024). Randomized spline trees for functional data classification: Theory and application to environmental time series. *arXiv preprint arXiv:2409.07879*.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological) 53*(1), 233–243.

Sang, P., A. B. Kashlak, and L. Kong (2023). A reproducing kernel hilbert space framework for functional classification. *Journal of Computational and Graphical Statistics 32*(3), 1000–1008.

Sangalli, L. M., P. Secchi, and S. Vantini (2014). Aneurisk65: A dataset of three-dimensional cerebral vascular geometries.

Sangalli, L. M., P. Secchi, S. Vantini, and A. Veneziani (2009). A case study in exploratory functional data analysis: geometrical features of the internal carotid artery. *Journal of the American Statistical Association 104*(485), 37–48.

Stamm, A. (2017). fdahotelling: an r package for hotelling's t squared in hilbert spaces. `https://github.com/astamm/fdahotelling`. Consulté le 26 mai 2025.

Suárez, A. and J. F. Lutsko (1999). Globally optimal fuzzy decision trees for classification and regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence 21*(12), 1297–1311.

Tengman, E., H. Grip, A.-K. Stensdotter, and C. K. Häger (2015). Anterior cruciate ligament injury about 20 years post-treatment: a kinematic analysis of one-leg hop. *Scandinavian journal of medicine & science in sports 25*(6), 818–827.

Tuddenham, R. D. (1954). Physical growth of california boys and girls from birth to eighteen years. *University of California publications in child development 1*(2).

Venturini, G. M., A. Muñoz, and J. González (2014). Generalizing the mahalanobis distance via density kernels. *Intelligent Data Analysis 18*(6_suppl), S19–S31.

Verwer, S. and Y. Zhang (2019, Jul.). Learning optimal classification trees using a binary linear program formulation. *Proceedings of the AAAI Conference on Artificial Intelligence 33*(01), 1625–1632.

Wang, S., Y. Huang, and G. Cao (2024). Review on functional data classification. *Wiley Interdisciplinary Reviews: Computational Statistics 16*(1), e1638.

Yang, L. and A. Shami (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing 415*, 295–316.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

46/2025 Mirabella, S.; David, E.; Antona, A.; Stanghellini, C.; Ferro, N.; Matteucci, M.; Heuvelink, E.; Perotto, S.
On the Impact of Light Spectrum on Lettuce Biophysics: A Dynamic Growth Model for Vertical Farming

45/2025 Caliò, G.; Ragazzi, F.; Popoli, A.; Cristofolini, A.; Valdettaro, L; De Falco, C.; Barbante, F.
Hierarchical Multiscale Modeling of Positive Corona Discharges

44/2025 Brivio, S.; Fresca, S.; Manzoni, A.
Handling geometrical variability in nonlinear reduced order modeling through Continuous Geometry-Aware DL-ROM

43/2025 Tomasetto, M.; Manzoni, A.; Braghin, F.
Real-time optimal control of high-dimensional parametrized systems by deep-learning based reduced order models

41/2025 Torzoni, M.; Maisto, D.; Manzoni, A.; Donnarumma, F.; Pezzulo, G.; Corigliano, A.
Active digital twins via active inference

42/2025 Franco, N. R.; Manzoni, A.; Zunino, P.; Hesthaven, J. S.
Deep orthogonal decomposition: a continuously adaptive neural network approach to model order reduction of parametrized partial differential equations

40/2025 Tentori, C.A.; Gregorio, C.; ...; Ieva, F.; Della Porta, M.G.
Clinical and Genomic-Based Decision Support System to Define the Optimal Timing of Allogeneic Hematopoietic Stem-Cell Transplantation in Patients With Myelodysplastic Syndromes

37/2025 Spreafico, M.; Ieva, F.; Fiocco, M.
Causal effect of chemotherapy received dose intensity on survival outcome: a retrospective study in osteosarcoma

Gimenez Zapiola, A.; Boselli, A.; Menafoglio, A.; Vantini, S.
Hyper-spectral Unmixing algorithms for remote compositional surface mapping: a review of the state of the art

35/2025 Perotto, S.; Ferro, N.; Speroni, G.; Temellini, E.
Anisotropic recovery-based error estimators and mesh adaptation for real-life engineering innovation