ARTICLE IN PRESS

Journal of Multivariate Analysis xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Data driven orthogonal basis selection for functional data analysis

Rani Basna [a,1], Hiba Nassar [b,*,1], Krzysztof Podgórski [c,1]

[a] Department of Internal Medicine and Clinical Nutrition, University of Gothenburg, Sweden
[b] Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[c] Department of Statistics, Lund University, Sweden

ARTICLE INFO

Article history:
Received 28 February 2021
Received in revised form 1 October 2021
Accepted 20 October 2021
Available online xxxx

AMS 2020 subject classifications:
62R10
62-08

Keywords:
Functional data analysis
Machine learning
Splines

ABSTRACT

Functional data analysis is typically performed in two steps: first, functionally representing discrete observations, and then applying functional methods, such as the functional principal component analysis, to the so-represented data. While the initial choice of a functional representation may have a significant impact on the second phase of the analysis, this issue has not gained much attention in the past. Typically, a rather *ad hoc* choice of some standard basis such as Fourier, wavelets, splines, etc. is used for the data transforming purpose. To address this important problem, we present its mathematical formulation, demonstrate its importance, and propose a data-driven method of functionally representing observations. The method chooses an initial functional basis by an efficient placement of the knots. A simple machine learning style algorithm is utilized for the knot selection and recently introduced orthogonal spline bases - splinets - are eventually taken to represent the data. The benefits are illustrated by examples of analyses of sparse functional data.

1. Introduction

In most current implementations of the functional data (FD) methods, the effects of the initial choice of an orthonormal basis that is used to analyze data have not been investigated. As a result, some standard bases such as trigonometric (Fourier), wavelet, or polynomial bases are chosen by default. Sometimes the machine learning methods are used to choose the number of elements of the basis but not the basis itself. No formal criteria are developed to give a researcher indication which of the bases is preferable for this initial representation of the data. On the other hand, it is both theoretically and practically observed that the choice of the basis affects efficiency in retrieving the stochastic structure of a studied model. One can point, in this context, to a classical result: the Karhunen–Loève (KL) expansion, see [11]. The basis associated with this expansion has the optimality in the average mean square error sense. However, this basis is typically the target of the analysis and, unfortunately, one cannot start representing data using this basis. Thus it becomes important to initially represent the data in such a way that the eigenfunctions of the KL expansion will be both efficiently and closely approximated by the elements of the space in which the data are represented. Without any prior knowledge of the functional character of the data, this initial representation needs to be learned from them. We propose to use the flexibility of splines that arises from considering the locations of their knots as free parameters.

* Corresponding author.
E-mail address: hibna@dtu.dk (H. Nassar).
[1] All authors contributed equally.

https://doi.org/10.1016/j.jmva.2021.104868

Please cite this article as: R. Basna, H. Nassar and K. Podgórski, Data driven orthogonal basis selection for functional data analysis, Journal of Multivariate Analysis (2021) 104868, https://doi.org/10.1016/j.jmva.2021.104868.

This work advances the idea initially presented in [14]. The approach of fitting a curve with free-knot splines has been studied before using regression splines by many researchers, we refer to [4,5,18] among others. However, to the best of our knowledge, no previous study has examined the benefits of searching for the knots to represent efficiently functional data. The main difference between the regression model case and the functional data is that in the latter we do not have any specific parametric model in mind, and the initial representation of the data (pre-processing) becomes a non-parametric problem. This affects the way one examines the accuracy as no fitting vs. overfitting methods cannot be utilized. Additional advantage of the proposed method is that we eventually obtain orthogonal spline basis that have not been used in the past. As a result, we obtain the convenient bases that can be compared to the Fourier or wavelet bases both in the efficiency and convenience of spectral representations.

The efficiency can be best argued using sparse functional data. The sparsity here refers to the case of clear locality of the main features in the data. For example, in the liquid chromatography tandem mass spectra (LC-MS/MS), the data are made of local peaks that are used for the identification of metabolomes or proteins. LC-MS/MS data sets are typically enormous and high dimensional in their nature. If one wants to represent them utilizing their sparsity is of the utmost importance. The methods developed in this work target this efficiency which should manifest best in an analysis of such data. We present this through simulations of a model that mimics the sparsity of the data, while empirical studies involving LC-MS/MS metabolomics data are left for some future research.

In practice, the basis for data representation is often chosen by mathematical convenience. Usually, Fourier basis is used to describe periodic data, while spline bases are used for functional information without any strongly cyclic variation, see [16]. Sometimes the number of the basis elements needs to be assessed based on the accuracy of the data representation. Alternatively, large (high-dimensional) basis are used to assure that the representation is sufficiently accurate. In both cases, one faces inefficiencies leading to the unnecessarily high dimensionality of the problem at hand. In our approach, we acknowledge the benefit of using spline basis in FDA but we proceed differently from the regularization approach. We propose algorithmic search for efficient knot patterns and use them for representation of sparse data. The used construction has been recently proposed in [12]. In the process, we incorporate machine learning algorithms for the choice of basis reducing the mean square error (MSE) uniformly for all samples and study its efficiency against other choices of the basis. The optimality criterion is utilized, both in the learning algorithms and in comparison studies. This criterion allows for comparison performances of different basis in a given problem. After efficiently learning from the data about knot placements, we utilize the new construction of the orthonormal spline basis, termed splinets and introduced in [12] as an efficient orthogonalization of the *B*-splines. It has been shown that the splinets bring convenience of the orthogonality while preserving the optimality properties similar to those featured by the *B*-splines. The method is first tested and illustrated on an artificial example that mimics sparsity in the data since the method would be particularly beneficial to account for sparsity. However, we also demonstrate that there is also a gain when we apply the method to a real FD set. Namely, we apply the methodology to provide an efficient way of representing the classical functional wine spectra dataset.

The paper is organized as follows. We start in Section 2 with a presentation of the structural scheme of the approach and motivate it by comparing it to the traditional approach and through a mathematical result on its robustness with respect to the choice of a functional model. This is followed by Section 3 where setting mathematical fundamentals of the initial basis and its role in the dimension reduction in the context of FDA using the spectral decomposition of the covariance operator. The statistical problem of empirically driven basis choice is formulated in mathematical terms and its basic properties are presented. In the next section, we discuss the data driven basis choice approach based on the knots for spline bases. There is also a brief account of the recently proposed orthogonal spline bases, splinets, which has been implemented in an R-package splinets. We refer to [12] for further details on splinets and [15] for a presentation of the package. A motivating example by which we illustrate the main features and benefits of the proposed approach for sparse data in Section 5. Then in Section 6, we present an approach which produces machine learning data-driven knot selection for the orthogonal splinet basis. There we also demonstrate the effectiveness of the obtained spline basis by comparing its performance to the one obtained from the Fourier basis. In Section 7 we illustrate our approach on a classical real functional dataset, namely the Wine spectra dataset.

2. General motivation for the data driven bases

FD are not observed as continuous objects, but high-frequency sampling and mathematical efficiency enable us to see these data as samples of curves, surfaces, or anything else varying over a continuum. The fundamental step in FDA is to convert this discrete recorded data to a truly functional form, which allows each function to be evaluated at any value of its continuous argument. In order to utilize the topology of such data for dimension reduction, one must perform data conversion. Typically, one represents a functional object as a linear combination suitable basis functions. For this purpose, one of the standard bases such as trigonometric, wavelet, or polynomial is typically chosen. Then the efficiency is accomplished by using smoothing through regression or roughness penalty for estimating the coefficients of the basis expansions. All such analyses are preceded by the initial choice of a functional basis used to analyze data, which is hardly objective and is more often driven by mathematical convenience than by the nature of the data itself. On the other hand, it is both theoretically and practically observed that the choice of the basis affects efficiency in retrieving the functional structure of a studied model.

The importance of the choice of the basis is particularly evident if there is a possible mismatch between the character of the data and the chosen basis. The method we propose in this work uses the placement of the knots and associated spline bases to produce a low-dimensional method that is robust with respect to the type of the data involved. In Fig. 1, we illustrate the main difference between the proposed method (blue box to the left) and the traditional approach of pre-processing of discrete data into functional ones. In our approach we use machine-learning style of placing knots to represent the data efficiently first on the training data subset and then verifying the significance of the knot choice on the validation data subset, as described in detail in Section 6.

The robustness of our method is quite evident on examples of sparse functional data as will be seen in Section 5. It follows from the robustness of the method with respect to the character of the data as a proper knot placement results in the efficiency of representation of a function by corresponding bases. One can explore this robustness by developing a mathematical theory through formalization of the concept. In this work, we aim only at the introduction and illustration of the method leaving more complete theories out of its scope. However, to give an example of mathematical results and to motivate the robustness of the method, we present the following theorem with the proof of it. In its formulation, we consider the zero order splines that are used in our methods to select knots and for which the argument, which is placed in the Appendix, is very simple.

Theorem 1. *Let $\mathcal{L}$ be a set of functions of bounded variation on $[0,1]$ with values in $[0,1]$. Further consider two orthogonal spline bases of the zero order $\mathcal{F}$ and $\mathcal{F}_f$, $f \in \mathcal{L}$, where the first is based on $n = 2\ell + 2$, $\ell \in \mathbb{N}$ equally spaced knots and the second is based on n knots obtained from the canonical Jordan decomposition of f into the difference two non-decreasing and bounded functions f_+ and f_- by taking the knots using the generalized inverses*

$$\boldsymbol{\xi} = sort\left(\{0, f_+^{-1}(y_1), \ldots, f_+^{-1}(y_\ell), f_-^{-1}(z_1), \ldots, f_-^{-1}(z_\ell), 1\}\right),$$

where for $i \in \{1, \ldots, l\}$:

$$y_i = f_+(0) + \frac{i}{l+1}(f_+(1) - f_+(0)),\ \ z_i = f_-(0) + \frac{i}{l+1}(f_-(1) - f_-(0)).$$

Then the following bounds hold

$$\sup_{f\in\mathcal{L}} \|f - \sum_{e\in\mathcal{F}} \langle e, f\rangle\, e\|_\infty \geq 1, \quad \sup_{f\in\mathcal{L}} \|f - \sum_{e\in\mathcal{F}} \langle e, f\rangle\, e\,\| \geq (4n)^{-1/2},$$

$$\sup_{f\in\mathcal{L}} \|f - \sum_{e\in\mathcal{F}_f} \langle e, f\rangle\, e\,\|_\infty \leq \frac{2TV(f)}{n}, \quad \sup_{f\in\mathcal{L}} \|f - \sum_{e\in\mathcal{F}_f} \langle e, f\rangle\, e\,\| \leq \frac{2TV(f)}{n},$$

where $\|\cdot\|_\infty$, $\|\cdot\|$ are the sup- and L_2 norm, respectively, while $TV(f) = (f_+(1) - f_+(0)) + (f_-(1) - f_-(0))$ is the total variation of f.

Remark 1. In the above result, we observe that if one uses the basis that accounts for the properties of a function the approximation gains the robustness on the type of a function one has as an input. This type of gain becomes important if the basis $\mathcal{F}$ is not representing efficiently f. In our situation, it occurs, for example, for the sparse functions. However, this efficiency extends beyond this case. For example it applies also when the data has a lot of local features spread irregularly over its domain.

Remark 2. This simple result can be related to Fig. 1, where the basis $\mathcal{F}$ corresponds to the one used in the traditional pre-processing of the data shown in the right-hand-side column, while $\mathcal{F}_f$ through the dependence on f imitates a data driven choice of the basis. In the above result we do not explore a theory of stochastic models as it would require a specific argument for each model. For example, for the Karhunen–Loéve, see the next section, one can develop bounds by considering the Jordan decomposition of each eigenfunction. Our intention in this work is not to develop such a theory but only signal fundamental principles of the data driven selection of the functional basis and show an example of practical implementation of these principles.

3. Initial basis selection problem

In this section, we present a mathematical formulation of initial basis selection and provide the formal assessment of the effect the initial basis has on the dimension reduction and the bias–variance trade-off in the final analysis of the data.

To be more specific, let us consider functional observations $x_k(t)$, $k \in \{1, \ldots, n\}$ that are random elements of $L^2 = L^2[0,1]$, i.e., the space of square integrable functions on the unit interval. The collection of such FD is denoted by $\mathcal{X}$. For the Hilbert space L^2, we use inner product $\langle\cdot,\cdot\rangle$ for an integral of the product of its two functional elements and which generates the norm $\|\cdot\|$. We call L^2-valued data functional observations. We use upper case and lower case letters in the context of FD in a similar manner as in the classical statistical convention, i.e., X is yet not observable random element, while $x = x(\cdot)$ stands for its particular observed functional realization, i.e., a functional outcome of a random

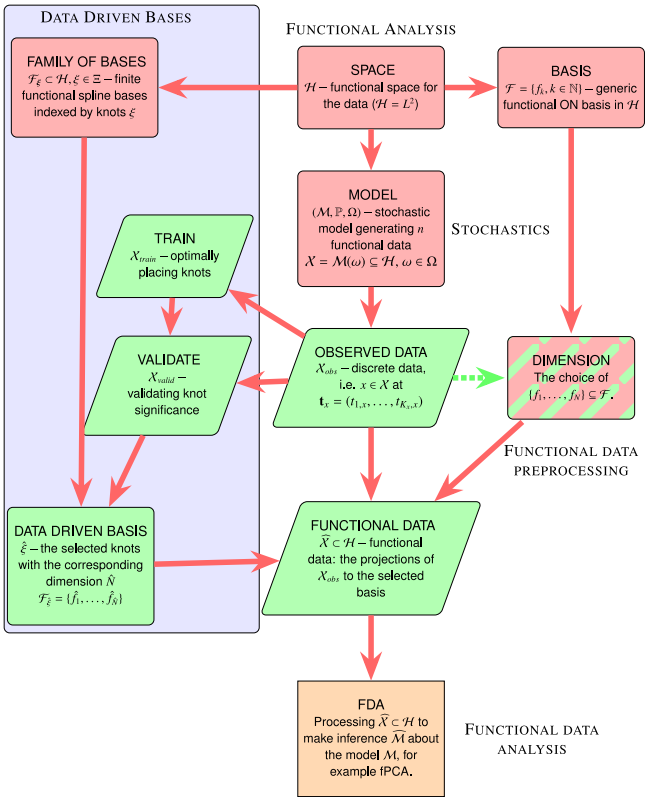

Fig. 1. Comparison of the traditional pre-processing of the functional data (right column) vs. the one based on data driven functional bases (left bluish box). The green boxes represent parts of the processing where the data explicitly is involved, the reddish boxes involve only purely theoretical framework. In the traditional functional data pre-processing the choice of the dimension is typically set rather high and is not based on the data. However, occasionally it is driven by the data which is indicated by the green-dashed arrow and the green lined box. Additionally, $\widehat{\cdot}$ is put over a symbol whenever data are used to define the corresponding object. The data driven bases are further detailed in Section 6 and in Algorithm 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

experiment carried out according to the probability model for X. The mean function of the functional data model X is defined by

$$m(t) = E[X(t)]$$

and the covariance function by

$$\sigma(s, t) = Cov(X(s), X(t))$$

for $s, t \in [0, 1]$, provided the appropriate expectations exist. Finally, the covariance operator on L^2 is defined as

$$\mathbf{K}h = \int_0^1 \sigma(\cdot, t)h(t)\, dt.$$

The Karhunen–Loève expansion, see [11], demonstrates that the basis associated with this expansion is characterized by the optimality in the average mean square error sense, for more details see [9].

Theorem 2 (*Karhunen–Loéve*)**.** *For a zero-mean $X(t)$, if $Cov(X(s), X(t)) = \sigma(s, t)$ is a continuous function, then there exist a non-increasing square summable sequence of non-negative numbers λ_k, an orthonormal basis e_k, $k \in \mathbb{N}$ in $L^2[0, 1]$ and a sequence of uncorrelated, zero-mean variance-one random variables Z_k such that*

$$Z_k = \int_0^1 X(t)e_k(t)\, dt/\sqrt{\lambda_k}, \quad X(t) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} Z_k\, e_k(t), \tag{1}$$

where the convergence is in the mean squared value and is uniform in t. Moreover, the covariance function of the process is represented in the uniform convergence over $[0, 1]^2$ *as*

$$\sigma(s,t) = \sum_{k=1}^{\infty} \lambda_k e_k(s)\bar{e}_k(t).$$

In the spirit of the main data analysis paradigm, for a given FD set, it is computationally effective and optimal to work with the basis of the eigenfunctions $\{e_k\}$. However, in the classical functional data, the basis $\{e_k\}$ is the target of the statistical analysis, is not known a'priori and thus cannot be used for the initial representation of the data. Instead, the data must be represented by some other functional basis. Consider, for example, the classical smoothing problem, where for a given data we want to fit a smooth function. Using the *B*-splines together with a regularization method, for example, the Lasso method, one may selectively choose a subspace of the spline space by shrinking parameters to zero, see [5,7]. Such a basis can be chosen for each FD sample but a choice valid for all samples is not obvious and can significantly affect the accuracy and efficiency of the analysis. For example, choosing an efficient basis for each sample and then conglomerate the bases into one leads to an unnecessarily high dimensional basis and produces computational challenges, see [1].

Let us formulate the problem in mathematical terms. We assume that our data are functions belonging to $\mathcal{E} = \overline{\text{lin}}\{e_k, k \in \mathbb{N}\}$ that are sampled at some discrete times. Here and in what follows, $\overline{\text{lin}}(\cdot)$ stands for the closure in L^2 of a linear space spanned over the argument of the operation. In general, $\mathcal{E}$ can be a proper sub-Hilbert space of L^2. For example, in the case of the so-called sparse functional data, for which there are regions where functions are equal to zero, the eigenfunctions e_k's will be zero on these regions and thus $\mathcal{E} \subsetneq L^2$. Consider another orthonormal basis $\mathcal{F} = \{f_k, k \in \mathbb{N}\}$ in which the data will be represented, i.e.,

$$X(t) = \sum_{i=1}^{\infty} \langle X, f_i\rangle f_i(t). \tag{2}$$

We have

$$\langle X, f_i\rangle = \sum_{k=1}^{\infty} \sqrt{\lambda_k} Z_k \langle e_k, f_i\rangle,$$

and due to the obvious practical limitations, one can only consider the above for a finite number, say I, of f_i's. In this set-up, the criticality of the initial choice of the basis lies both in the fact that we consider only the finite elements of it and use them to represent the data X but also because the basis functions f_i, $i \in \{1, \ldots, I\}$, may not well approximate e_k's, i.e., the projections $\hat{e}_{k,I}$ of e_k to $\mathcal{F}_I = \{f_1, \ldots, f_I\}$ given by

$$\hat{e}_{k,I} = \sum_{i=1}^{I} \langle e_k, f_i\rangle f_i$$

may poorly approximate e_k's. This would result also in a poor approximation of the original data. In fact, we have the following result about the mean square error of approximation of the data within basis $\mathcal{F}_I$.

Proposition 1. *We consider X given in the K-L model (1). Let $I \in \mathbb{N}$ and X_I be the projection of X to $\mathcal{F}_I$. Then*

$$E[\|X - X_I\|^2] \le \sum_{k=1}^{\infty} \lambda_k \|e_k - \hat{e}_{k,I}\|_2^2.$$

Proof. We note that

$$X_I = \sum_{k=1}^{\infty} \sqrt{\lambda_k} Z_k \hat{e}_{k,I}$$

and the result follows from the triangle inequality and the fact that the variables Z_k's in (1) are uncorrelated. □

Remark 3. From this result, it is quite clear that the accuracy of the representation of FD in some functional basis improves if $\hat{e}_{k,I}$ approximates the eigenfunctions e_k well. For a fixed basis $\mathcal{F}$, it can be achieved by increasing the number I. However, in order to reduce the dimensionality, it can be beneficiary to seek a basis $\mathcal{F}$ that yields a good approximation $\hat{e}_{k,I}$ of e_k even for small I. The eigenvalues λ_i's weigh into the quality of the approximation, thus choosing the basis $\mathcal{F}_I$ so that $\hat{e}_k$ is close to e_k for large values of λ_k (small values of k due to the ordering of λ_k's would be preferable). However, since neither eigenvalues nor eigenfunctions are known, in the approach below, we consider choosing the basis $\mathcal{F}_I$ that minimizes the empirical version of $E[\|X - X_I\|^2]$.

Since $\mathcal{F} = \bigcup_{I=1}^{\infty} \mathcal{F}_I$ spans L^2, it is clear that $E[\|X - X_I\|^2]$ decreases with the increase of I and one can improve the quality of data representation by selecting I sufficiently large. In particular, one can consider empirically driven $\hat{I}$ as the

minimizer of the total empirical mean square error

$$AMSE(\mathcal{X}, \mathcal{F}_I) = \sum_{i=1}^{n} \|x_i - x_{i,I}\|^2/n.$$

In this set-up, one can use some statistical or machine-learning criteria to select $\hat{I}$ such that $\mathcal{F}_{\hat{I}}$ efficiently approximate data without overfitting. The machine-learning approach was, for example, advocated in [17]. Alternatively, one could use noise data $\mathcal{W}$ to see how the reduction of *AMSE* for $\mathcal{X}$ compares with the one obtained for $\mathcal{W}$ and stop the procedure when the improvement is similar. In either approach, the choice $\hat{I}$ is based on the data making it random and thus $\mathcal{F}_{\hat{I}}$ becomes random and data driven. The problem with this approach is that $\hat{I}$ may need to be very large to capture the features in the data. This is particularly evident if a lot of elements of $\mathcal{F}$ are needed to represent the eigenfunctions corresponding to large eigenvalues. This increase produces variance inflation, as discussed next.

If $\mathcal{F}_{\hat{I}}$ is chosen for representing the functional data, then the eigenfunctions will be determined by the vectors $\mathbf{v}_k = [\langle e_k, f_1\rangle \dots \langle e_k, f_I\rangle]^\top$ which satisfy the spectral equation

$$\lambda_k \mathbf{v}_k = \boldsymbol{\Sigma}\mathbf{v}_k \tag{3}$$

for the $I \times I$ matrix $\boldsymbol{\Sigma} = [Cov(\langle X, f_\ell\rangle, \langle X, f_k\rangle)]_{\ell,k=1}^{I}$. In practice, $\boldsymbol{\Sigma}$ will be obtained as an estimate from the I-dimensional multivariate problem with n observations given by $\mathbf{x}_i = [\langle x_i, f_1\rangle \dots \langle x_i, f_I\rangle]^\top$, $i \in \{1, \dots, n\}$. However, if I is large relatively to n, then the consistency of estimation of the eigenvector–eigenvalue pairs is no longer valid, see [10], which is not surprising as the number of entries to be estimated in $\boldsymbol{\Sigma}$ is on the order $I^2/2$ for large I and all these entries are involved in solving (3). Thus to avoid this estimation variance inflation associated with large I, it is practically important to find the initial basis that allows for a small value of I to be sufficient for accurate estimation. □

4. The data driven choice of the orthogonal spline basis

As pointed above, the initial choice of the basis for data representation is of high importance and, typically, it can be only chosen by exploring the data themselves. In what follows, we present a formal set-up in which the problem can be formulated and then present its particular realization that involves spline bases.

It is assumed that there exists a family $\{\mathcal{F}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi}\in\Xi}$ of finite functional and orthonormal bases that are parameterized by a finite dimensional parameter $\boldsymbol{\xi}$. In our implementation of this approach, the parameter is associated with knots of spline bases. Our data driven choice of the basis is defined as $\widehat{\mathcal{F}} = \mathcal{F}_{\widehat{\boldsymbol{\xi}}}$, where

$$\widehat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi}\in\Xi}{\operatorname{argmin}}\, AMSE(\mathcal{X}, \mathcal{F}_{\boldsymbol{\xi}}).$$

We note that this data driven choice of $\widehat{\mathcal{F}}$, while appealing, is difficult to study analytically. For example, Proposition 1 is no longer valid even if we consider the conditional expectation $E\left[\|X - \hat{X}_{\hat{I}}\|^2 \,\middle|\, \widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}\right]$. This is because conditioning on the random $\widehat{\boldsymbol{\xi}}$ may introduce the dependence between variables Z_i's in the K-L representation of X. Realistically, one has to resort to either Monte Carlo or machine learning methods (or both). The fundamental prerequisite for the method is the identification of $\{\mathcal{F}_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi}\in\Xi}$ and this may depend on conceptualization of the targeted FD. The spline bases, their dependence on the location of knots, and capacity of representing sparse data by considering support subsets of the argument space were the main inspiration for the presented approach. The flexibility of the splines due to the knots placements has been a subject of the studies in the past, see [5,13] among many others. What is new in the proposed approach is tying it to the efficiency of FDA or more precisely to the accuracy of the functional principal component decomposition.

The proper orthogonal spline basis that are characterized by sparsity understood as local support sets are fundamental for the analysis of sparse functional data. Such basis obtained by appropriate orthogonalization of the *B*-splines has been introduced in [12]. Here is their very brief account.

A *B*-spline is a smooth function that consists of polynomial pieces that have the same degree, connected smoothly at join points $\xi_0 < \xi_1 < \cdots < \xi_{n+1}$, referred to as knots. The *B*-splines are sensitive to the choice of the knots' position, which is behind our main idea of the basis selection since the choice of the knots can be data-driven. Once the knots are set, *B*-splines can be effectively evaluated in a recursive way for any degree using the Cox–de Boor formula [3]. The *B*-splines have interesting properties that characterize them. Namely, all *B*-splines are positive, differentiable up to a certain level (the spline order) and have minimal compact intervals for their supports. But except for the case of order zero, the *B*-splines are not orthogonal. Different orthogonalization methods appeared in the literature but we are using the structured orthogonalization that creates basis systems referred to as splinet.

The splinet is prioritized over other orthonormal spline systems as it preserves locality and computational efficiencies of the original splines. For a more detailed explanation, we refer the reader to [12]. Given the knots, these spline basis are obtained from the *B*-splines through efficient dyadic orthogonalization that is performed in a self-similar fashion and thus preserving the locality entertained by the *B*-splines. In Fig. 2, we see the ability to represent the local detail of two splinets spread over two different knot placements.

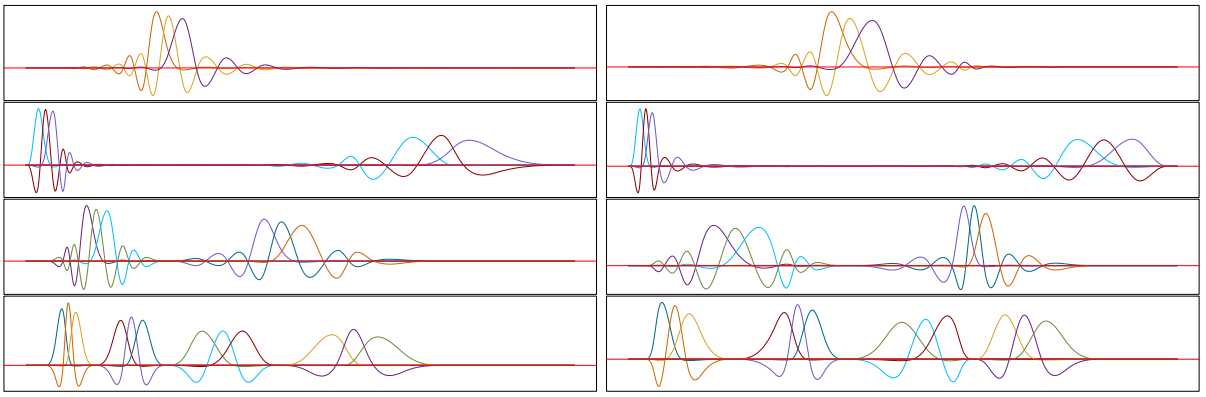

Fig. 2. Two splinets (efficient orthogonal spline bases of order three) spread over two different knot placements. The locality and sensitivity to the knot placement is evident.

5. A motivating sparse data example

To illustrate the main point of our work and features of the proposed approach, we consider (1) that follows the general finite dimensional set-up. Namely, for a certain generic orthonormal basis (f_ℓ), $\ell \in \{1, \ldots, I\}$, the eigenfunctions (e_k), $k \in \{1, \ldots, K\}$, are defined through

$$e_k = \sum_{\ell=1}^{I} a_{k\ell} f_\ell$$

where the $K \times I$ matrix $\mathbf{A} = (a_{k\ell})$ satisfies

$$\mathbf{A}\mathbf{A}^{\perp} = \mathbf{I}_K,$$

where $\mathbf{I}_K$ is the $K \times K$ identity matrix. Further λ_k, $k \in \{1, \ldots, K\}$ are non-increasing eigenvalues corresponding to e_k.

As discussed in Sections 3 and 4, there are several issues with functionally representing the original data through a certain basis. First, the basis may be known, i.e., f_ℓ's are known but the number of elements in the basis used for the representation is not properly chosen, second, how this effect is amplified when the estimation of the covariance operator is factored into the eigenvalue–eigenfunction decomposition of the data. Finally, if the basis is chosen a'priori with no input from the data in the selection process, the estimation in the functional principal component introduces a significant variance noise in the eigenvalues estimation. In this work, we illustrate how using a data-driven orthonormal basis can improve efficiency in representing FD. In the sequel, we compare FD representation in splinets built on knots chosen using our method with FD representation in splinets built on equally spaced knots. It should be pointed out that what we observe in our study for splinets basis build on equally spaced knots is also present for other types of non-empirically driven basis, the Fourier basis being a classical case, see [14].

In what follows, we discuss all these issues. In a simple specification of the above model, we take $I = 9$ and for f_ℓ, $\ell \in \{1, \ldots, 9\}$, we take the third order orthogonal splines that are elements of a splinet spanned on irregularly placed knots, see [12]. Due to the sparsity of the B-splines that are inherited by the splinet, the model serves as a toy example of a sparse data generator. The splinet (orthonormal functional basis) is presented in Fig. 3, the top graph. We take $K = 4$ eigenfunctions defined through the matrix $\mathbf{A}$ given in

$$\mathbf{A} = \begin{bmatrix} 2^{-\frac{1}{2}} & 0 & 0 & 0 & 2^{-\frac{1}{2}} & 0 & 0 & 0 & 0 \\ 0 & 2^{-\frac{1}{2}} & 0 & 0 & 0 & 2^{-\frac{1}{2}} & 0 & 0 & 0 \\ 0 & 0 & 2^{-\frac{1}{2}} & 0 & 0 & 0 & 2^{-\frac{1}{2}} & 0 & 0 \\ 0 & 0 & 0 & 3^{-\frac{1}{2}} & 0 & 0 & 0 & 3^{-\frac{1}{2}} & 3^{-\frac{1}{2}} \end{bmatrix}$$

that leads to the normalized eigenfunctions e_k, $k \in \{1, \ldots, 4\}$, shown in Fig. 3 the middle graph. The four corresponding eigenvalues are

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 0.5, 0.3, 0.01).$$

Ten samples from (1) in which we assumed Z_i's to be the standard normal variables, the case of a Gaussian model, are shown in the bottom graph of Fig. 3.

In this simple example, one can illustrate how critical is the initial choice of a basis used for the representation of the data. For example, if one chooses the basis $\mathcal{B}_1 = \{f_3, \ldots, f_9\}$, then the mean square error of using this basis for the data

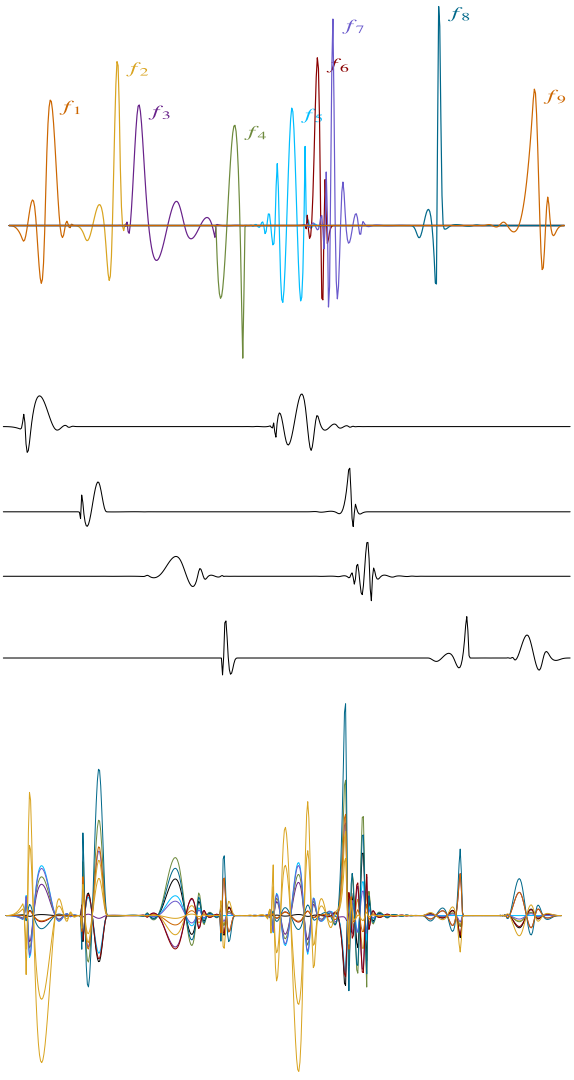

Fig. 3. The basis $\{f_\ell\}$, $\ell \in \{1, \ldots, 9\}$ giving a complete representation of the data (top). The eigenfunctions $\{e_k\}$, $k \in \{1, \ldots, 4\}$ (middle) for the functional model (1). Ten samples from the model are shown in the bottom graph.

representation is

$$E\left\|\sum_{k=1}^{K}\sqrt{\lambda_k}Z_k\ \left(e_k(t)-\hat{e}_k(t)\right)\right\|^2 = E\left\|\sum_{k=1}^{K}\sqrt{\lambda_k}Z_k\ \left(a_{k1}f_1+a_{k2}f_2\right)\right\|^2 = \sum_{k=1}^{K}\lambda_k\left(a_{k1}^2+a_{k2}^2\right) = 0.85.$$

which leads to major inaccuracies. On the other hand, similar computations for the basis $\mathcal{B}_2 = \{f_1, \ldots, f_7\}$ lead to the error 0.046 which is rather negligible as compared with the total mean squares norm of the data which is $\lambda_1^2+\cdots+\lambda_4^2 = 1.3401$. Of course, choosing the full basis $f_1, \ldots, f_9$ will yield the exact representation of the data. However, in practice, such an exact representation rather should not be expected. Firstly, it is typically assumed that the data are from infinite dimensional space so that there are infinite many e_k's in (1), while an analyst will have only a finite number of the basis elements available. The further complication of the problem is that the mean square error can only be estimated and further confounded by the noise in the data.

In a more realistic setup, the observed data are also involving an observational noise with some given standard deviation σ_0. Thus in a simple model that accounts for the noise, we consider

$$X(t) = \sum_{k=1}^{K}\sqrt{\lambda_k}Z_k\, e_k(t) + \sigma_0 dB(t), \tag{4}$$

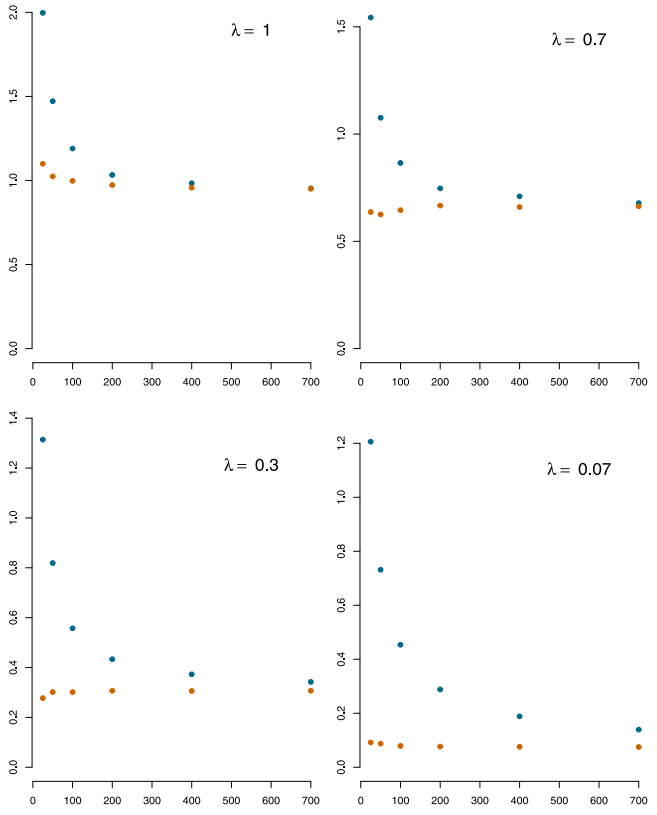

Fig. 4. Mean of the four eigenvalues for the estimated covariance matrix obtained from 200 Monte Carlo simulations. Two different cases of the initial basis selection, the case f_ℓ, $\ell \in \{1, \ldots, 9\}$ (orange-lower dots), and 200 splinets OB_ℓ, $\ell \in \{1, \ldots, 200\}$ (blue-upper dots) as a function of the sample size, which ranges from 25 to 700.

where B is a Brownian motion independent of Z_k's. Consequently, for any f_ℓ, $\ell \in \{1, \ldots, I\}$, we have the following projection of the data into the basis (f_ℓ):

$$\langle \mathbf{X}, f_\ell \rangle = \sum_{k=1}^{K} \sqrt{\lambda_k} a_{k\ell} Z_k + \sigma_0 \int_0^1 f_\ell dB(t).$$

We observe that the noise component $\epsilon_\ell = \sigma_0 \int_0^1 f_\ell dB(t)$ has the variance σ_0^2 one and constitutes uncorrelated random variables. This means that instead of the covariance $\boldsymbol{\Sigma}$, the coefficients of the data expansions in the basis (f_ℓ) have the covariance

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma} + \sigma_0^2 \mathbf{I}_I, \tag{5}$$

where $\mathbf{I}_I$ is the $I \times I$ identity matrix. In this situation, an efficient basis is beneficent not only because of its computational convenience but it allows also to reduce the estimation error if the sample size is small comparing to the dimension of the basis.

Next, we study the effect of the sample size on precision of the eigenvalue/eigenfunction estimation based $\hat{\sigma}$ in two cases of the initial selection of the orthogonal basis: f_ℓ, $\ell \in \{1, \ldots, 9\}$, and 200 splinets OB_ℓ, $\ell \in \{1, \ldots, 200\}$ created on equally spaced knots. We consider the case of noisy data with the variance $\sigma_0^2 = 0.1$ and 6 different sample sizes: 25, 50, 100, 200, 400, 700. For each sample size, we run a Monte Carlo study with $MC = 200$ to compute the eigenvalues of the covariance matrix (5) for the two initial choices of a basis. Fig. 4 shows that the four eigenvalues of the estimated covariance matrix make good estimators for the actual eigenvalues in the case of f_ℓ, $\ell \in \{1, \ldots, 9\}$ as the initial basis despite the sample size, in contrast to the case of 200 splinets OB_ℓ, $\ell \in \{1, \ldots, 200\}$ where eigenvalues of the covariance matrix become good estimators for the actual eigenvalues only when the sample size is fairly big.

Moreover, the Monte Carlo study also shows that the mean square errors (MSE) of the estimated eigenvalues for small sample sizes are significantly better if the 'correct' basis of f_ℓ, $\ell \in \{1, \ldots, 9\}$ is chosen instead of 200 splines with equally spaced knots, see Table 1. Fig. 4 shows that using 200 splines with equally spaced knots fails to estimate the first four eigenvalues of the covariance matrix in the cases of a small or medium sample size. The results become obviously much worse when one considers a smaller number than 200 of equally spaced basis splines.

Table 1
MSE of the estimated eigenvalues λ_i, $i \in \{1, \ldots, 4\}$, using two bases: f_ℓ, $\ell \in \{1, \ldots, 9\}$ (Basis 1) and 200 equally spaced splinets (Basis 2).

Sample size	Basis 1				Basis 2			
	MSE(λ_1)	MSE (λ_2)	MSE(λ_3)	MSE(λ_4)	MSE(λ_1)	MSE(λ_2)	MSE(λ_3)	MSE (λ_4)
25	0.78	0.20	0.10	0.04	2.58	0.94	1.10	1.32
50	0.57	0.14	0.07	0.03	1.14	0.31	0.29	0.46
100	0.43	0.11	0.06	0.02	0.69	0.14	0.09	0.16
200	0.37	0.08	0.04	0.02	0.48	0.08	0.04	0.06
400	0.31	0.05	0.03	0.01	0.35	0.05	0.02	0.02
700	0.29	0.04	0.02	0.01	0.31	0.04	0.02	0.01

Our illustrative example emphasizes the importance of choosing the initial basis. Of course, in reality, it is not possible to know a priori the basis that generates the model. Now, we turn to a method of selecting a data-driven orthogonal basis.

6. Data driven choice of the knots

The common degree of polynomials and the placement of knots define the spline basis. While order 3 is the most popular generic choice, often the degree of the involved polynomials may be decided by the nature of the data. For example, if one considers the Brownian bridge B, then the samples are continuous but not differentiable and the first order splines (piecewise linear functions) are a natural choice. On the other hand, if one chooses the Laplace bridge $L = B \circ \Gamma$, where Γ is a gamma process, which is an example of a pure jump process, using the zero order splines may be the most natural to represent process samples. Overall, the choice of the spline order is a separate issue and we do not discuss it in any further detail here. Our focus is on the choice of knots and subsequently on their number.

As presented in the diagram of Fig. 1, at the center of the proposed methodology is data driven knot (DDK) selection. While the formal argument for such a DDK method is signaled in Theorem 1, practical implementations can vary depending on a functional model, a type of data, or a purpose of the analysis. For example, if one assumes the observational noise in the data, the addition of suitably placed knots can stop if there is an indication that added knots stop representing some features in the data and instead are capturing a pure noise. In such a case, the behavior of knot selection for pure noise data should be quantified. It should be pointed that in the functional data pre-processing one does not fit any parameters and thus strictly speaking the concept of overfitting is not applicable. However, if there is abundance of the data, one can mimic the machine learning methods by dividing the data into learning and validation parts and stop the addition of the knots if on validation set it is detected that the specific local features are no longer captured by the added knots. This can be achieved through, for example, a significance type criterion as described next.

The average mean-square error (AMSE) reduction by adding a knot can be considered as a sample from a distribution on the positive real half-line. In the training set one chooses a knot that brings the largest AMSE reduction. However, if this knot represents a feature in the data, then the reduction for this knot in the AMSE sample computed based for the validation sample should also be rather large, i.e. to be in the tail of this validation AMSE reduction distribution. In the future, we plan to investigate such a machine-learning inspired approach.

In this work however, we propose a simplified machine-learning style technique for the placement and validation of the knots. The general idea of using the piecewise constant function and partitioning the domain goes back to [6] and is nowadays exploited by the regression tree method and its modifications. Using piecewise constant functions (0-order splines) for the knot selection is dictated by its simplicity and numerical efficiency and other order of splines can be utilized for the knot selection as well although algorithms needed for the purpose are computationally more demanding. The chosen knots are used to build orthogonal splines basis functions $f_k(t)$, $k \in \{1, \ldots, K\}$ that are used in basis function expansion to convert the data from discrete recorded data into a functional one. We use a standard random split of the data to training and validation parts, although other machine learning techniques such as cross-validation, bagging, etc. can be utilized for the purpose as well. The random partitioning preserves the original discretization for each sample. More specifically, a functional data set $\mathcal{X}$, is partitioned randomly to $\mathcal{X}_{train}$ and $\mathcal{X}_{valid}$ using $60\% - 40\%$ split. Our method is based on selecting data driven knots (DDK) and its initial implementation in R language is available as a package (under continuous development) at the GitHub page: https://github.com/ranibasna/ddk.

The method of adding knots is based on the mean square error of approximating the FD. The method is iterative and resembles the regression tree building, see [8, Chapter 9]. Namely, for any FD set $\mathcal{X} = \{x_i \in L^2, i \in \{1, \ldots, n\}\}$, we start with the set of best least square constant predictors

$$x_i^{(0)} = \langle x_i, \mathbf{1} \rangle \mathbf{1} = \int x_i \cdot \mathbf{1}.$$

The constant functions over the entire domain [0, 1] can be viewed as 0-order splines with no internal knot points, and its one dimensional basis is given by the constant function $\mathbf{1}$. We set the initial set of knots to an empty set, i.e. $\mathcal{K}^{(0)} = \emptyset$,

the initial basis $\mathcal{B}^{(0)} = \{\mathbf{1}\}$, and the projection to the space spanned by $\mathcal{B}^{(0)}$ is given by $\mathbf{P}^{(0)}x = \langle x, \mathbf{1}\rangle \mathbf{1}$. The average mean square error (AMSE) per function of the approximations of x_i's by the optimal constant functions is given by

$$AMSE(\mathcal{X}, \mathcal{B}^{(0)}) = \frac{1}{n}\sum_{i=1}^{n} \| x_i - \mathbf{P}^{(0)}x_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} \| x_i - \langle x_i, \mathbf{1}\rangle \mathbf{1}\|^2.$$

The method at the first step, $s = 1$, finds a knot $\xi \in [0, 1]$ such that the optimal approximation of x by a linear combination of the 0-order splines with the set of knots $\mathcal{K}^{(0)} \cup \{\xi\}$ yields the smallest AMSE. In other words, denote by $\mathcal{B}^{(1)}(\xi)$ the orthonormal basis of piecewise constant functions over the intervals given by these knots. The new knot ξ_{new} is chosen as

$$\xi_{new}^{(1)} = \underset{\xi\in(0,1]}{\operatorname{argmin}}\, AMSE(\mathcal{X}, \mathcal{B}^{(1)}(\xi)). \tag{6}$$

Then the new, enlarged by one function, basis $\mathcal{B}^{(1)} = \mathcal{B}^{(1)}(\xi_{new}^{(1)})$ is uniquely defined by the new set of knots $\mathcal{K}^{(1)} = \mathcal{K}^{(0)} \cup \{\xi_{new}^{(1)}\}$. In the recurrent process, at the step s, we start with a sequence of knots $\mathcal{K}^{(s-1)}$. Those knots divide the interval I into s smaller intervals. The search for a new optimal knot is done by using (6) on each smaller interval. As a result we get a set of s nominated knots $\{\zeta_1, \zeta_2, \ldots, \zeta_s\}$ (one knot in each smaller interval). The new knot ξ_{new} is chosen as

$$\xi_{new}^{(s)} = \underset{\xi\in\{\zeta_1,\zeta_2,\ldots,\zeta_s\}}{\operatorname{argmin}}\, AMSE(\mathcal{X}, \mathcal{B}^{(s)}(\xi)). \tag{7}$$

At the end of the iteration step s, the algorithm returns the set of knots $\mathcal{K}^{(s)} = \mathcal{K}^{(s-1)} \cup \{\xi_{new}^{(s)}\}$ and the corresponding orthonormal basis of piecewise constant functions $\mathcal{B}^{(s)}(\xi)$.

Remark 4. It is important to point out that the evaluation of the nominated knots has to be done only for two smaller intervals (those who has been created by the split from the previous step) since all the other has been computed in the previous steps of the algorithm and are preserved in our implementation.

6.1. Validation

The number of selected knots plays an essential role. A large number of knots may result in capturing a noise rather than some systematic features in the data. In contrast, a small number may result in missing some local features in the data. Here, we utilize the validation set to test if a newly selected knot proposed by the training stage as the optimal one has also a significant improvement on the validation set. That is to say, if the value of the reduction in the average mean square error is 'significant' as opposed to randomly selected knots. In the sparse data context, the goal is to stop the iterations before the algorithm starts picking knots on the flat regions.

To address this problem, a validation index using a stopping threshold has been proposed motivated by traditional machine learning algorithms. Namely, we developed stopping criteria that automate the convergence time of the above-explained knots selection method. For the stopping threshold, we suggest two criteria: the step difference between consecutive AMSE values and the relative step difference between consecutive AMSE values. Both criteria are achieved by using a predefined step size threshold. The step size at iteration s is defined as the absolute difference between two sequential average mean square error values, $|AMSE_s - AMSE_{s+1}|$, where AMSE_s is the value of the AMSE at the step s. The step size threshold θ is a bound on the value of a step the AMSE function takes.

The validation procedure is running simultaneously with the training one. At the iteration step s, the training algorithm delivers the new set of knots $\mathcal{K}^{(s)}$ and a new set of pieces-wise constant basis $\mathcal{B}^{(s)}$. In the next step, $AMSE_s = AMSE(\mathcal{X}_{valid}, \mathcal{B}^{(s)})$ is computed and two criteria for the stopping: the step difference between consecutive AMSE values and the relative step difference between consecutive AMSE values.

By the first criterium, if the algorithm at iteration i reduces the AMSE value by a value smaller than the threshold θ the iteration stopped and no more knots are selected. In other words, we stop the iteration if the following condition is met

$$| AMSE_s - AMSE_{s-1} | < \theta. \tag{8}$$

Similarly, for the second criterion, the algorithm stops at iteration s if the relative step difference between consecutive AMSE values is smaller than the threshold θ. The relative step difference is the ratio of the absolute difference to a reference value

$$| AMSE_s - AMSE_{s-1} | < \theta \, | AMSE_s | . \tag{9}$$

It is clear that in mathematical terms, the increase of the number of knots is no longer needed when the knots start to fit a pure noise instead of the actual functional structure. In the future development, we plan to expand these ideas to a full-blown statistical test that will analyze the behavior of the above criteria when they are applied to data sampled

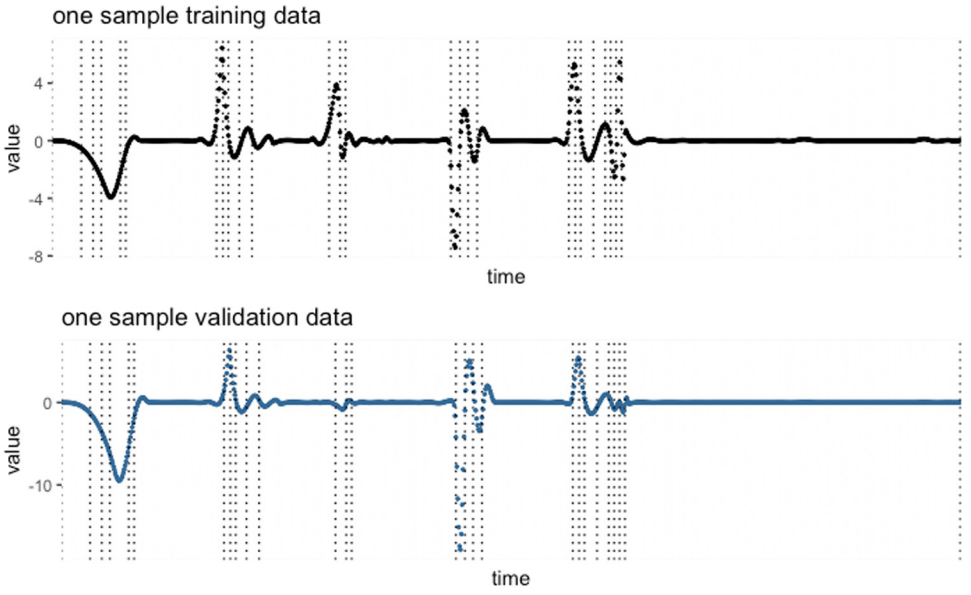

Fig. 5. A sample from the train set *(top)* and from the validation set of the motivating example model of Section 5. The vertical dashed lines are at the 25 data-driven knots selected with validation. The knots emerged from using criterion (8) with predefined step size $\theta = 0.01$.

from dB, where B is a Brownian motion: the algorithm should stop adding knots when the hypothesis that one deals with a pure noise cannot be rejected. A full description of the DDK algorithm can be found in Algorithm 1.

Algorithm 1: DDK algorithm

Input: I - knot search interval, θ - validation hyperparameter, $\mathcal{X}_{train}$ - training data, $\mathcal{X}_{valid}$ - validation data

- Find the first global knot ξ using Eq. (6) on $\mathcal{X}_{train}$;
- Add split to I at the location of ξ and add ξ to $\mathcal{K}^{(0)}$, the initial set of knots;
- Set $s = 1$

while *condition* (8) *or* (9) *on* $\mathcal{X}_{valid}$ *is satisfied* **do**

- Find the new optimal placement using Eq. (7);
- Add split to I at the location of the new selected knot ξ_s.
- $s = s + 1$

end

Output: $\mathcal{K}^{(s)}$

For our motivated example of Section 5, we run the validation process using the first criterion with predefined step size $\theta = 0.01$. The threshold value θ is chosen such that the reduction produced by more iteration is very small and the contribution is not significant compared to the computational cost one has to pay.

Fig. 6 shows that this criterion is met at the 25th iteration and one can observe that after the 25th iteration very little reduction can be achieved. The selected θ triggered a stopping at iteration number that represents the elbow of the AMSE reduction curve. The intuition behind such a choice is that increasing the number of knots will naturally improve the fit (explain more of the variation). However, at the same time, this leads to overfitting, and the elbow point reveals this threshold. Fig. 5 displays the outcome of the validation process and the selected knots from the training and validations represented in dashed lines.

Fig. 6 shows both the AMSE evaluation on $\mathcal{X}_{train}$ and $\mathcal{X}_{valid}$. It is evident that in the training phase each knot selection imposes a reduction on AMSE with different step sizes. On the validation data, we observe similar behavior as in the training data. Overall, Fig. 6 explains the mechanism of stopping the iteration process.

6.1.1. Computational complexity

There are many different aspects of computational complexity in the proposed approach. Probably the most innovative is using an efficient orthogonalization of the splines leading to the orthogonal basis of splines - splinet. The efficiency of the orthogonalization algorithm, see [12], and the obvious computational advantages of the orthogonalized functional basis reduce the computational complexity for the functional PCA.

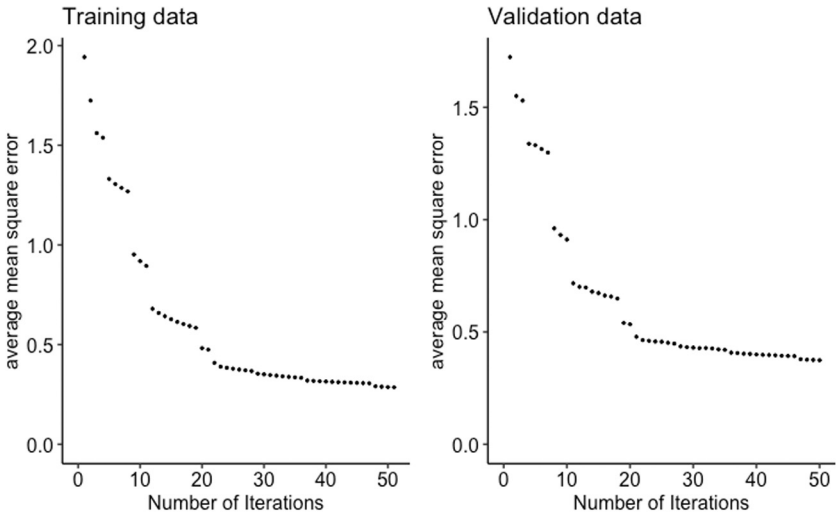

Fig. 6. Reduction in AMSE in the motivating example of Section 5. *Left*: reduction in AMSE achieved after each additional knot selection during training. *Right*: reduction achieved after each additional knot selection on the validation data.

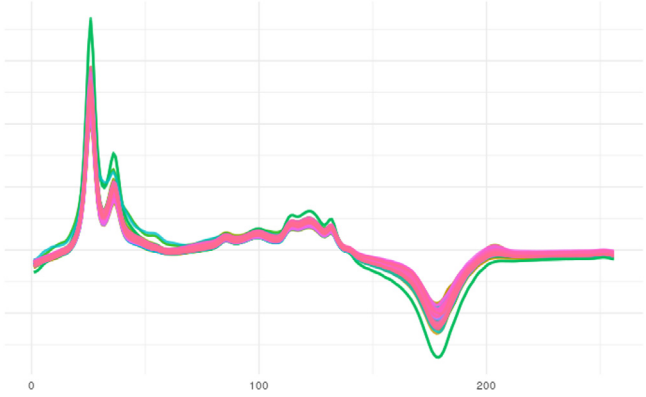

Fig. 7. 30 samples from the wine training data set. On the horizontal axis we have 256 integers corresponding to 256 wave numbers between 4000 and 400 cm^{-1}, and on the vertical axis the absorbance of the wine samples at these wave numbers.

Another computational benefit is using the zero-order splines for the knot selection through the machine learning methodology. Here, the algorithm benefits from the locality and orthogonality piecewise constant basis so that each new knot requires removal of only one base function (the constant over an interval that includes the new knot) and replaces it with two new functions that remain orthonormal to all the other basis functions from the previous step.

As a result, the algorithm is quite computationally efficient and fast. The computational complexity is at most of order $O(n \cdot m)$, where n is the number of observations and m is the number of iterations. Moreover, the algorithm spends most of the time (which depends on the dataset size and the size of its sampling argument) in the first iteration, but after this step, it only updates one basis function at a time and utilizes all the previous steps computations.

7. Application – Wine dataset

For the purpose of testing the efficiency of DDK methods we choose the classical functional wine spectra dataset. The wine dataset is provided by Professor Marc Meurens of the Université Catholique de Louvain, [2]. The data set consists of results collected from an experiment measuring the alcohol concentration in wine samples by mid infrared spectroscopy. The training and validation sets contain 94 and 30 spectra, respectively (i.e., the data have been split up to approximately 75% training and 25% validation), with 256 spectral variables that are the absorbance ($\log 1/T$) at 256 wave numbers between 4000 and 400 cm^{-1} (where T is light transmittance through the sample thickness). Fig. 7 shows 30 samples from the training dataset.

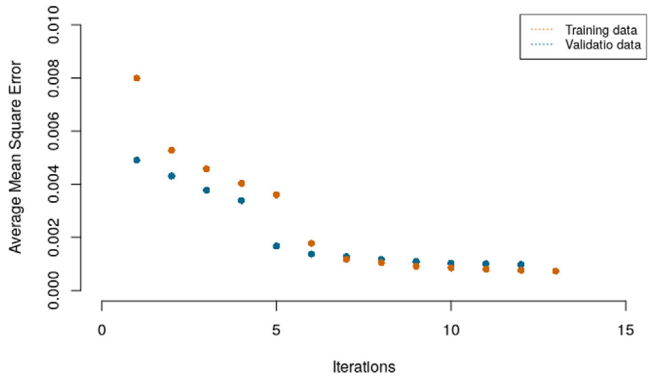

Fig. 8. Reduction in AMSE in the Functional Wine data. *(Orange-bottom)* reduction in AMSE achieved after each additional knot selection during training. *(Blue-top)* reduction achieved after each additional knot selection on the validation data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7.1. The DDK decomposition

In the sequel, we will apply the knot selection algorithm to the Wine dataset. Our focus is to retrieve local features near locations where the data curves have some curvatures.

Applying DDK algorithm on the training dataset for selecting knots and simultaneously validating our choice on the validation dataset at every iteration. At each iteration, a new knot is selected and the average mean square error will reduce in both the training dataset and the validate dataset, as can be seen in Fig. 8. As in Section 6.1, we use the first criterion to determine the optimal number of knots. We choose θ to be 0.0001. As described in Section 6.1 the θ represents a predefined stopping threshold that produces a stopping breakpoint for the algorithm. Indeed, when looking at Fig. 8, it is clear that θ induced a stopping point in which the number of the selected knots is at the "elbow" of the reduction in AMSE curve i.e. the point after which the distortion/inertia start decreasing in a linear fashion, in the reduction of the average mean square error in the validation data set. Thus for the given data, we decide for 8 knots.

After selecting the knots, we use the function `project()`, from the `Splinets` R-package, to project data into splinets of the third order with the selected knots. In the next step, we perform the spectral decomposition of data by estimating the eigenvalues λ_i's and the corresponding eigenfunctions $f_i(t)$. We present the estimated eigenvalues in the decreasing order and the first four eigenfunctions scaled by the square roots of their corresponding eigenvalues in Fig. 9. The vertical dashed lines refer to selected knots.

Fig. 9 (Bottom-Left) shows, for a sample in the data, the difference between the original wine data, the projected into splinets over the knots selected using the data-driven developed approach, and the functional data spectrally decomposed and reconstructed using the first eigenfunction. It is apparent that the projection using the DDK method gives a decent smooth 1D fit of the original data and efficiently avoids overfitting the original curve.

At the last step of this study we compare between $\hat{f}_1$, the projected data into splinets build over the DDK and $\hat{f}_2$, the projected data into splinets build over the same number of equally spaced knots. The average mean square errors between the original data and the projected data are

$$AMSE(f - \hat{f}_1) = 0.007242398, \quad AMSE(f - \hat{f}_2) = 0.01908081.$$

One can observe a significant improvement of the quality (by approximately 65%) of the averaged mean square error. Moreover, One needs around 16 knots (double number) to achieve a relatively close average mean square of the error between the original data and the projected data into splinets build over the 16 equally spaced knots. Consequently, this will enlarge the subspace spanned to be twice the dimensions of the subspace resulted from the DDK method.

We then compare between the projected data into the first eigenfunction $\tilde{f}_1$, the first two eigenfunctions $\tilde{f}_2$ and the first three eigenfunctions $\tilde{f}_3$. The average mean square errors between the original data and the projected data are

$$AMSE(f - \tilde{f}_1) = 0.01428159, \quad AMSE(f - \tilde{f}_2) = 0.00810629, \quad AMSE(f - \tilde{f}_3) = 0.007637792.$$

It is worth noting that the average mean square error between the original data and the projected data into the first two eigenfunctions is so close to the average mean square errors between the original data and the projected data into splinets build over the DDK.

In the conclusion of this study, we report that one can obtain more samples of the wine spectra by using a simple low-dimensional model

$$\tilde{Y}_u(t) = \mu(t) + \sqrt{\lambda_1} Z_1 f_1(t) + \sqrt{\lambda_2} Z_2 f_2(t) + \sqrt{\lambda_3} Z_3 f_3(t),$$

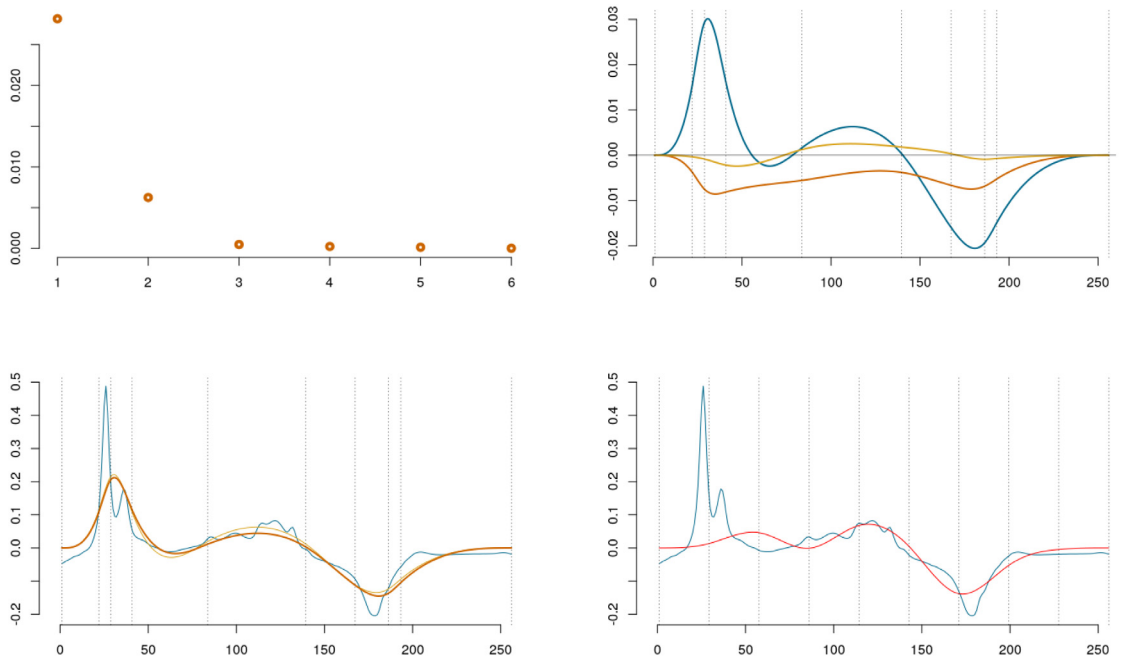

Fig. 9. *Top-Left:* The eigenvalues ordered in a decreasing manner; *Top-Right:* the first three eigenfunctions scaled by the square roots of their corresponding eigenvalues; *Bottom-Left:* One sample data (Blue curve), the projected data into splinets build over the knots chosen with DDK (yellow curve) with the location of the selected knots (vertical dashed lines), and the data decomposed using only the first eigenfunctions (orange curve); *Bottom-Right:* One sample data (Blue curve), the projected data into splinets build over equally spaced knots (yellow curve) with the location of the knots (vertical dashed lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in which the distribution of the vector standardized uncorrelated random variables (Z_1, Z_2, Z_3) can be fitted by taking the inner products of the data with the eigenfunctions and standardized obtained data. Our inspection of this empirical distribution showed that it does not deviate in a drastic manner from the standard vector normal distribution. Thus in further simplification of the model, one can obtain samples from it using samples of standard normal variables.

8. Conclusions

The mathematical foundations for an empirically driven basis selection for representation of functional data analysis have been presented. The framework was then implemented through the knot selection for orthogonal spline basis. The proposed method of the data-driven orthonormal basis decomposition has been tested both through numerical simulations and on functional wine data. The Monte Carlo simulations show clear advantages over the Fourier based method, in particular, when smoothed splines are used. The accuracy is not only exhibited in smaller errors but also in the reduced variability of the error. The improvement is greater, as expected, for the data that shows some local detail. The obtained results suggest that the method has a great potential to improve the functional analysis of the sparse data as shown through the classical functional wine data. The benefits of the proposed methodology for the sparse data which was also demonstrated through Monte Carlo simulations. To provide more analytical insight into the gains more studies are needed. It is expected that due to the heavily non-linear effect of data driven choice of the initial basis, any analytical approach needs to be assisted by Monte Carlo/machine learning methodology.

CRediT authorship contribution statement

Rani Basna: Conceptualization, Formal analysis, Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Hiba Nassar:** Conceptualization, Formal analysis, Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Krzysztof Podgórski:** Conceptualization, Formal analysis, Data curation, Investigation, Methodology, Software, Writing – original draft, Writing - review & editing.

Acknowledgments

The financial support of the Swedish Research Council (VR) Grant DNR: 2020-05168 is gratefully acknowledged by the third author.

Appendix

Proof of Theorem 1. Consider $f \in \mathcal{L}$ taking value one at a symmetric interval at the center of $[1/n, 2/n]$ and zero otherwise. It is clear that the projection $\hat{f}$ of such a function to the splines of the zero order (piece-wise constant functions) will be constant at $[0, 1/n]$, then having value zero above $2/n$. When the length of the interval over which the function f is constant converges to zero the value over this interval of $\hat{f}$ will converge to zero thus implying the first lower bound in the result.

The second part follows by setting the value one over $[0, 1/(2n)]$ for the above choice of f and zero otherwise. It can be easily shown that $\hat{f}$ is equal to $1/2$ over the same interval and thus $\|f - \hat{f}\|^2 = 1/(4n)$ proving the second lower bound.

We turn now to the upper bounds for the function driven basis selection. It is clear that the constant value over $[\xi_i, \xi_{i+1}]$ that correspond to the projection (in the L_2-sense) is equal to $\int_{\xi_i}^{\xi_{i+1}} f(x)\,dx/(\xi_{i+1} - \xi_i)$, which in turn, by the mean value theorem, is equal to $f(\xi_i')$ for some $\xi_i' \in [\xi_i, \xi_{i+1}]$ (here we assume that f is continuous or otherwise the argument needs to be refined through a continuous approximation). Using the Jordan decomposition and the definition of $\boldsymbol{\xi}$ we have for $x \in [\xi_i, \xi_{i+1}]$

$$|f(x) - f(\xi_i')| \le |f_+(x) - f_+(\xi_i')| + |f_-(x) - f_-(\xi_i')| \le TV(f)/(\ell + 1) = 2TV(f)/n$$

and thus the first upper bound is shown.

For the second upper bound, using the above inequality we observe

$$\|f - \hat{f}\|^2 = \sum_{i=0}^{n} \int_{\xi_i}^{\xi_{i+1}} \left(f(x) - f(\xi_i')\right)^2 dx \le 4\left(TV(f)\right)^2/n^2,$$

which yields the result. □

References

[1] A.M. Aguilera, M. Aguilera-Morillo, Comparative study of different B-spline approaches for functional data, Math. Comput. Modelling 58 (7–8) (2013) 1568–1579.
[2] N. Benoudjit, E. Cools, M. Meurens, M. Verleysen, Chemometric calibration of infrared spectrometers: selection and validation of variables by non-linear models, Chemometr. Intell. Lab. Syst. 70 (1) (2004) 47–53.
[3] C. De Boor, A Practical Guide To Splines, vol. 27, Springer-Verlag, New York, 1978.
[4] D. Denison, B. Mallick, A. Smith, Automatic Bayesian curve fitting, J. R. Stat. Soc. Ser. B Stat. Methodol. 60 (2) (1998) 333–350.
[5] I. DiMatteo, C.R. Genovese, R.E. Kass, Bayesian curve-fitting with free-knot splines, Biometrika 88 (4) (2001) 1055–1071.
[6] J.H. Friedman, Multivariate adaptive regression splines, Ann. Statist. (1991) 1–67.
[7] J. Guo, J. Hu, B.-Y. Jing, Z. Zhang, Spline-lasso in high-dimensional linear regression, J. Amer. Statist. Assoc. 111 (513) (2016) 288–297.
[8] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer-Verlag, New York, 2009.
[9] T. Hsing, R. Eubank, Theoretical Foundations of Functional Data Analysis, with an Introduction To Linear Operators, John Wiley & Sons, Chichester, 2015.
[10] I.M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Statist. 29 (2) (2001) 295–327.
[11] K. Karhunen, Über lineare Methoden in der Wahrscheinlichkeitsrechnung, Ann. Acad. Sci. Fennicae. Ser. A. 37 (1947) 1–79.
[12] X. Liu, H. Nassar, K. Podgórski, Splinets–efficient orthonormalization of the B-splines, 2019, arXiv preprint arXiv:1910.07341.
[13] N. Molinari, J.-F. Durand, R. Sabatier, Bounded optimal knots for regression splines, Comput. Statist. Data Anal. 45 (2) (2004) 159–178.
[14] H. Nassar, K. Podgórski, Empirically driven orthonormal bases for functional data analysis, in: Numerical Mathematics and Advanced Applications ENUMATH 2019, Lecture Notes in Computational Science and Engineering 139, Springer, 2021, pp. 1–12.
[15] K. Podgórski, Splinets–splines through the taylor expansion, their support sets and orthogonal bases, 2021, arXiv preprint arXiv:2102.00733.
[16] J. Ramsay, B. Silverman, From functional data to smooth functions, Funct. Data Anal. (2005) 37–58.
[17] F. Yao, H.-G. Müller, J.-L. Wang, Functional data analysis for sparse longitudinal data, J. Amer. Statist. Assoc. 100 (470) (2005) 577–590, arXiv:https://doi.org/10.1198/016214504000001745.
[18] S. Zhou, X. Shen, Spatially adaptive regression splines and accurate knot selection schemes, J. Amer. Statist. Assoc. 96 (453) (2001) 247–259.