SORT 47 (2) July-December 2023, 295-322 DOI: 10.57645/20.8080.02.10

Classification of probability density functions in the framework of Bayes spaces: methods and applications

Ivana Pavlů[1], Alessandra Menafoglio[2], Enea G. Bongiorno[3] and Karel Hron[4]

Abstract

The process of supervised classification when the data set consists of probability density functions is studied. Due to the relative information contained in densities, it is necessary to convert the functional data analysis methods into an appropriate framework, here represented by the Bayes spaces. This work develops Bayes space counterparts to a set of commonly used functional methods with a focus on classification. Hereby, a clear guideline is provided on how some classification approaches can be adapted for the case of densities. Comparison of the methods is based on simulation studies and real-world applications, reflecting their respective strengths and weaknesses.

MSC: *62R10.*

Keywords: *Probability density functions, Bayes spaces, classification, functional data analysis.*

1. Introduction

Classification, i.e., assigning observations to classes based on a set of features, is one of the most common tasks of mathematical statistics with a strong practical motivation. Banks and insurance companies evaluate their potential customers and divide them into groups to avoid excessively risky behaviours. Hospitalized patients can be classified into different risk groups based on their symptoms and/or their physical attributes. Dis-

[1] Department of Mathematical Analysis and Application of Mathematics, Palacký University Olomouc, Czech Republic; ivana.pavlu@upol.cz

[2] MOX - Department of Mathematics, Politecnico di Milano, Italy.

[3] Dipartimento di Studi per l'Economia e l'Impresa, Università degli Studi del Piemonte Orientale, Novara, Italy.

[4] Department of Mathematical Analysis and Application of Mathematics, Palacký University Olomouc, Czech Republic.

Received: October 2022
Accepted: May 2023

tributions of financial income amongst population can be used to distinguish differences between regions. In another context, soil particle-size distributions can be classified into one of potential sampling localities.

Most classification methods originate in the multivariate setting (Hartigan, 1975; Hastie, Tibshirani and Friedman, 2009). Here, in many cases, observed data are relative, meaning that the relevant information is contained in proportions between components rather than in their absolute values interpreted separately. However, this fact is not taken into account when using standard multivariate methods, which can, in a sense, overlay the geometrical properties resulting from the relative structure (Filzmoser, Hron and Templ, 2012). Instead, compositional data analysis (Aitchison, 1986; van den Boogaart, Egozcue and Pawlowsky-Glahn, 2014) offers a comprehensive methodology for dealing with such type of data, which can be further extended into the (virtually infinite-dimensional) case described hereinafter.

One common way of portraying distributional data is in the form of probability density functions (PDFs), unit integral non-negative functions defined over a bounded or unbounded supporting interval (Egozcue, Díaz-Barrero and Pawlowsky-Glahn, 2006; Hron et al., 2016; van den Boogaart et al., 2014). The proportions between the amounts of probability corresponding to certain subdomains of the support then represent counterparts of the proportions between components in the multivariate compositional case (Egozcue et al., 2013). Despite being functions, PDFs cannot be straightforwardly processed using standard functional data analysis (FDA) methods (Ramsay and Silverman, 2005; Horváth and Kokoszka, 2012) due to their relative properties (Hron et al., 2016; van den Boogaart et al., 2014). Most of the standard functional tools are developed for functions belonging to the L^2 space, the usual space of square-integrable functions. However, the L^2 geometry should not be blindly used for PDFs, as it does not preserve the compositional properties of distributional data – unlike the so-called Bayes space $\mathcal{B}^2$ (van den Boogaart, Egozcue and Pawlowsky-Glahn, 2010; 2014) considered in this work.

Due to the recent interest of the scientific community in functional distributional data, the analysis of PDFs including the Bayes space approach has been at the forefront. Although a lot of work has been done on classification for functional data from the L^2 perspective (Ferraty and Vieu, 2006; Jacques and Preda, 2014; James and Hastie, 2001; Nourollah Mousavi and Sørensen, 2017; Ramsay and Silverman, 2005; and more), at the moment there is a lack of a comprehensive methodology for dealing with classification of PDFs. This work aims to fill this gap, by focusing on the framework of supervised classification (i.e. when class labels of training data are known). In particular, classical and recent classification methods for functional data are discussed when extended or adapted to the PDFs setting through the Bayes approach. This paper thus provides a clear guideline on how also other possible classification methods can be adapted to the case of PDFs. Moreover, the selected methods are intentionally chosen with different theoretical foundations for a twofold purpose: first, to cover the most common supervised classification FDA approaches in the PDF context, and second, to assess whether and which effects are emphasized with the different approaches.

The structure of the paper is a follows. Section 2 offers a concise summary of the Bayes space background as well as a brief recall of the spline representation of PDFs in this setting. Section 3 introduces one common way of reducing the dimensionality of functional data and summarizes a selection of five classification methods with their reinterpretation for PDFs, i.e.: functional logistic regression, functional principal component regression, functional linear discriminant analysis, an approach based on small-ball probability and the functional k-nearest neighbours. In Sections 4 and 5, these methods are applied to both simulated and real-world data sets. Finally, Section 6 offers some additional comments and conclusions.

2. Representation of PDFs in Bayes spaces

The Bayes spaces $\mathcal{B}^2$ serve as a unifying framework for working with functional distributional data, expressed often in the form of PDFs. Their geometric structure is based on the generalization of the Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001) which is commonly used in the context of compositional data analysis (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015). Formally, the Bayes space $\mathcal{B}^2(I)$ is set to be a space of positive functions with square integrable logarithm carrying relative information (also known as functional compositions), defined on the bounded interval $I = [a,b]$ (Egozcue et al., 2006). This is the most common choice of the domain in FDA (often further restricted to $[0,1]$), however, the Bayes space theory can be developed also for possibly unbounded domains (van den Boogaart et al., 2010). Similarly, one can consider different reference measures - in this work, the uniform measure on I strictly plays the role of reference measure. To avoid confusion, only integrable densities are discussed in further text, although Bayes spaces cover the non-integrable densitites as well. In $\mathcal{B}^2$ one can define an equivalence relation, based on the scale invariance principle. Indeed, by rescaling a density function, its relative information – as defined by the proportions between the measure of intervals contained in I – does not change. As a consequence, one can consider non-negative unit-integral density functions (PDFs) as representatives of equivalence classes defined by the relation $f =_{\mathcal{B}} g$ if $f = cg$, for a given $c > 0$, making it a subspace of $\mathcal{B}^2$.

In $\mathcal{B}^2$, operations of sum of two functions, multiplication of a function by a scalar and inner product of two functions are replaced by perturbation $\oplus$, powering $\odot$ and inner product $\langle\cdot,\cdot\rangle_{\mathcal{B}}$ in Bayes Hilbert space, respectively. These are defined for a uniform reference measure for densities f and g in $\mathcal{B}^2$ and a real constant α, by

$$(f \oplus g)(t) =_{\mathcal{B}} f(t)\cdot g(t), \quad (\alpha \odot f)(t) =_{\mathcal{B}} f(t)^{\alpha}, \tag{1}$$

$$\langle f,g\rangle_{\mathcal{B}} =_{\mathcal{B}} \frac{1}{2(b-a)} \int_I \int_I \ln\frac{f(t)}{f(u)} \ln\frac{g(t)}{g(u)} \mathrm{d}t\,\mathrm{d}u, \tag{2}$$

where $t,u \in I = [a,b]$. The construction of operations (1) and (2) ensures that the resulting function maintains the relative properties of a density as well as the unit-integral

representation (if needed for interpretation purposes). Using (1), one can also define the distance between f and g as

$$d_{\mathcal{B}}(f_1, f_2) =_{\mathcal{B}} ||f_1 \ominus f_2||^2_{\mathcal{B}}, \quad \text{where} \quad (f_1 \ominus f_2)(t) =_{\mathcal{B}} f_1 \oplus [-1 \odot f_2(t)], \quad t \in I. \quad (3)$$

In further text, the subscript in $=_{\mathcal{B}}$ will be dropped - however, it should be always clear whether the equation holds in $\mathcal{B}$ or L^2.

To proceed with the statistical processing of densities, two ways can be pursued. Indeed, methods can be either developed directly in the $\mathcal{B}^2$ setting, or PDFs can be mapped into L^2 to make use of existing FDA methods. The latter approach is appealing due to the existence of an isometric mapping from $\mathcal{B}^2(I)$ into a subspace of $L^2(I)$ of zero-integral functions - frequently denoted as $L^2_0(I)$. This mapping, called the centred logratio (clr) transformation (van den Boogaart et al., 2014)

$$\mathrm{clr}(f)(t) = \ln f(t) - \frac{1}{b-a}\int_I \ln f(u)\mathrm{d}u, \quad t \in I \quad (4)$$

then results in a zero-integral real function on the same domain I. Using these clr-transformed densities (see Figure 1 for an illustration), the standard functional approaches can be considered, thus avoiding the computational inconveniences related to processing directly in Bayes spaces, and maintaining the relative information of the original data.

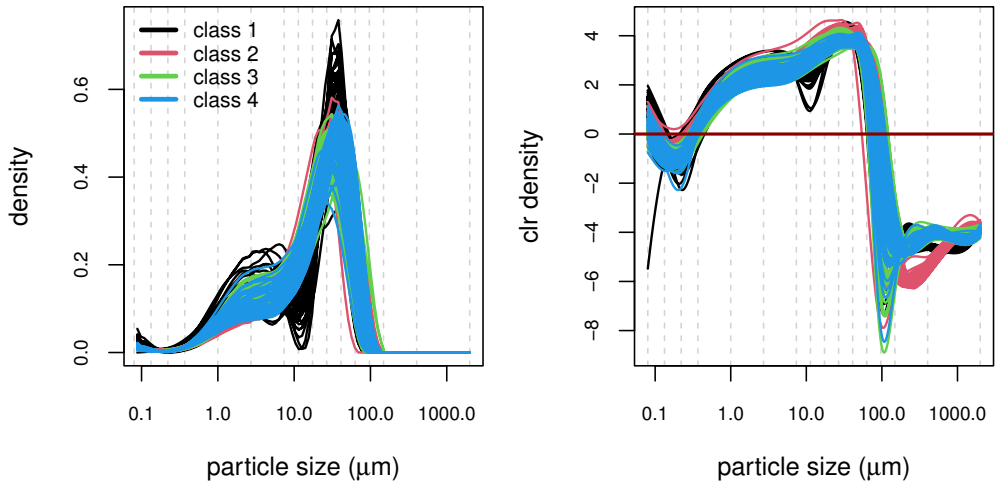

Figure 1. *Example of a four-class functional data set: densities in $\mathcal{B}^2$ (left) with their clr-transformed counterparts (right). Gray dashed lines represent the position of the knots used for their ZB-spline representation. It is obvious that clr transformation enables to highlight variation related to small functional values as compared to the original PDFs. This data set is further discussed in Section 5.2.*

Another important aspect to be mentioned is that, despite the increasing quality of measuring devices and data collection, it is rarely possible to observe functional data

as a whole continuous curve. Instead, the raw data usually consist of a discrete set of measures dispersed over the supporting domain. This obstacle can be overcome during the preprocessing stage. Commonly, for L^2 data, the B-spline representation (De Boor, 1978; Machalová, Hron and Monti, 2016) is used to smooth the discrete data by piece-wise polynomial functions, with Fourier bases and wavelets serving as examples of other basis choices. Using a common spline basis – defined by I, a sequence of knots and the order of polynomial k – each observation is uniquely described as a linear combination of basis functions. Commonly, knots are placed more densely in areas where data are more varying to capture this variability. In the context of clr densities, compositional splines (ZB-splines, Machalová et al. (2016, 2021)) were introduced where, as opposed to the usual B-splines, the basis functions are already constructed as zero-integral curves. Intuitively, each function constructed by using the compositional spline basis is indeed also a clr density with a zero integral and the basis itself is uniquely defined through the same parameters as in the general B-spline setting.

Similarly as for B-splines, the compositional ZB-spline functions $Z_i^k(t)$ are k-degree piecewise polynomials with $k-1$ continuous derivatives defined on $t \in I$. Examples of both the ZB-spline (right) and the B-spline bases (left), defined using cubic basis functions over the same set of 16 non-equispaced knots, are illustrated in Figure 2. When considering the ZB-spline basis, the observations are represented as a linear combination of the basis functions

$$x_k(t) = \sum_{i=-k}^{l-1} z_i Z_i^{k+1}(t), \quad t \in I, \tag{5}$$

where l defines the length of the vector of inner knots and z_i represents the spline coefficient corresponding to the $i-$th basis function. By linearity, each observation derived from (5) is guaranteed to still belong to the clr space. This essential step of the data preprocessing stage will be considered further in the descriptive text as well as during the data processing in both simulated and real data sets.

3. Classification techniques for PDFs

The techniques considered in this work cover (i) three of the most popular and commonly used parametric classification methods available in FDA; (ii) a semi-parametric approach based on the idea of the small-ball probability (see Bongiorno and Goia (2016)); and (iii) the non-parametric method of k-nearest neighbours. This section offers an overview of these methods, when adapted to the case of densities. To start, however, an insight to a functional principal component analysis (FPCA) is offered, as dimensionality reduction is a common and necessary step in data analysis of both multivariate and functional data.

For further reference, a brief summary of notation is offered here. In the following, X is a random element defined on the probability space $(\Omega, \mathcal{F}, P)$, taking values in $\mathcal{B}^2(I)$ and x is its observed value. Assume that Ω is partitioned by G subsets $\Omega_g \subset \Omega$ with $g \in \{0, \ldots, G-1\}$ and let Y be the random variable identifying the group label of ω and

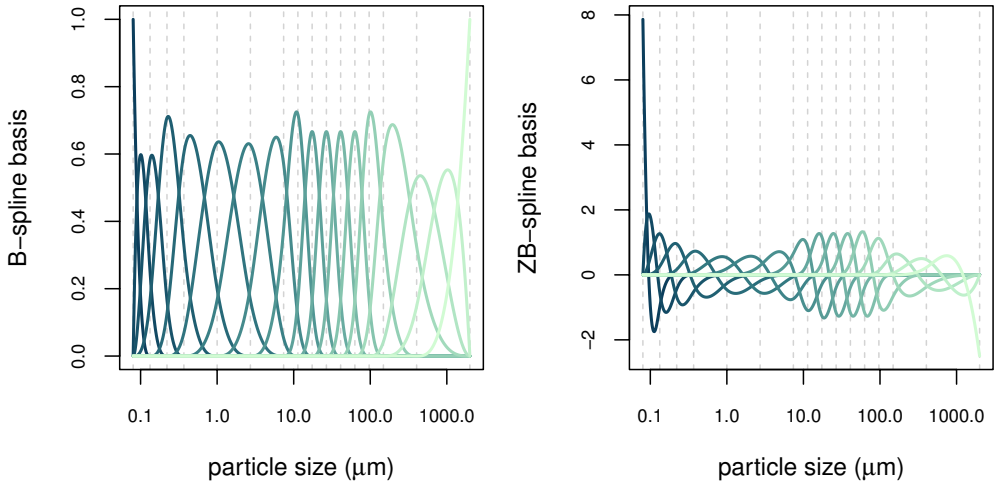

Figure 2. *B-spline (left) and ZB-spline basis (right) corresponding to the data set displayed in Figure 1.*

defined by

$$Y(\omega) = \sum_{g=0}^{G-1} \pi(g) \mathbb{I}_{\Omega_g}(\omega), \qquad \forall \omega \in \Omega, \tag{6}$$

where $\mathbb{I}_A$ denotes the indicator of A, $\pi(g) = P(Y = g) > 0$ and $\sum_{g=0}^{G-1} \pi(g) = 1$. If not specified, functional objects from $\mathcal{B}^2(I)$ will be solely considered further in the text.

3.1. Dimension reduction in functional data analysis

When considering functions in the Bayes space, reducing dimensionality involves projecting the data over a basis of the underlying Hilbert space and keeping a small number of the obtained basis coefficients. As an instance in the PDF setting, it is possible to consider the compositional spline basis as presented in the previous section, or the basis obtained from the functional principal component analysis (FPCA, Ramsay and Silverman (2005); Horváth and Kokoszka (2012); Hron et al. (2016)). The latter allows one, in addition to the dimensionality reduction, to maintain a significant proportion of variability from the original data set and for this reason is introduced in what follows. FPCA can be formulated directly for data belonging to $\mathcal{B}^2(I)$; in this case, which will be of interest for this work, it is named simplicial functional principal component analysis (SFPCA, Hron et al. (2016)).

In particular, for a random function $X(t) \in \mathcal{B}^2(I)$ with mean function $\mu(t) = E[X(t)]$ and the covariance operator $\mathbf{\Sigma}(\cdot) = E[\langle X \ominus \mu, \cdot \rangle (X \ominus \mu)]$, it is possible to consider the Karhunen-Loève expansion (Eaton, 1983; Horváth and Kokoszka, 2012)

$$X(t) = \mu(t) \oplus \bigoplus_{j=1}^{\infty} \theta_j \odot \xi_j(t) \tag{7}$$

where $\theta_j = \langle X \ominus \mu, \xi_j \rangle$ satisfies $E[\theta_j] = 0$, $Var(\theta_j) = \lambda_j$ and $E[\theta_j \theta_i] = 0$ for any $i \neq j$, and stands for the scores along the j-th simplicial functional principal component (SFPC), whereas $\{\xi_j(t)$ are the loadings obtained as the orthonormal eigenfunctions of $\mathbf{\Sigma}$ ordered according to the decreasing values of the associated eigenvalues $\{\lambda_j$. As a consequence, $\sum_{k=1}^{d} \lambda_k / \sum_{k=1}^{\infty} \lambda_k$ indicates the fraction of variability of X explained by the first d SFPCs, that is, by the d-dimensional process $X^{(d)}(t) = \mu(t) \oplus \bigoplus_{j=1}^{d} \theta_j \odot \xi_j(t)$ obtained by truncating (7). When working with a sample of N observations $x_1, \ldots, x_N$, the theoretical quantities are replaced with their empirical counterparts, namely the sample mean $\bar{x} = \frac{1}{N} \bigoplus_{i=1}^{N} x_i$ and the corresponding sample covariance operator $\mathbf{V}(\cdot) = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \langle x_i \ominus \bar{x}, \cdot \rangle (x_i \ominus \bar{x})$.

From a computational point of view, it has been shown that the easiest and most straightforward way to perform SFPCA is to transform the densities using (4) and to proceed within the L^2 space (Hron et al., 2016). Using a sample of zero-integral functional observations, the common FPCA setting – which is widely explored in the literature (Ramsay and Silverman, 2005) – is obtained.

We finally remark that, as in FPCA, there is no universal answer for which fraction of explained variability in SFPCA is sufficient for proceeding further with data analysis. It is important to account for the specific features of the data and the scope of the data analysis, to find a dimension d of functional principal components sufficiently high to appropriately describe the data. In this work, SFPCA will come into play in Sections 3.2, 3.3 and 3.5, and the influence of the number of used principal components on quality of classification will be illustrated within the Sections 4 and 5.

3.2. Functional logistic regression

Functional logistic regression (Nourollah Mousavi and Sørensen, 2017) is a classification method designed specifically for binary response, meaning that the problem reduces to the decision between two groups with labels $\{0,1\}$. In particular, in the $\mathcal{B}^2(I)$ setting, the classification rule is based on the probabilities $\pi(g|x) = P(Y = g|X = x)$ with $g \in \{0,1\}$ which provide the probabilities to associate the g-th group with a new observation $x \in \mathcal{B}^2(I)$. For $g = 1$, this conditional probability is given by

$$\pi(1|x) = \frac{\exp\{\alpha + \langle \beta, x \rangle_{\mathcal{B}}\}}{1 + \exp\{\alpha + \langle \beta, x \rangle_{\mathcal{B}}\}} \tag{8}$$

and can be rewritten in terms of the logit transformation of the original probability as

$$\eta(x) = \log\left(\frac{\pi(1|x)}{1 - \pi(1|x)}\right) = \alpha + \langle \beta, x \rangle_{\mathcal{B}}\,. \tag{9}$$

This way, aside from the intercept α, the relationship between X and Y can be described by a functional parameter $\beta(t)$. Embedded within the same B-spline basis as x, the PDF $\beta(t)$ can be decomposed as

$$\beta(t) = \sum_{i=-k}^{l-1} b_i B_i^k(t) = \mathbf{b}^T \mathbf{B}_i^k(t), \tag{10}$$

where $B_i^k(t)$ stands for the i-th B-spline basis function. Additionally, due to the relatively simple form of (9), parameters α and $\beta(t)$ (specifically spline coefficients $\mathbf{b}$ for the latter) can be estimated from the associated conditional likelihood function $L(\alpha,\beta|x_1,\ldots,x_N)$,

$$L(\alpha,\beta|x_1,\ldots,x_N)=\prod_{i=1}^{N}\pi_i^{g_i}(1-\pi_i)^{1-g_i}=\prod_{i=1}^{N}\frac{\exp\{g_i(\alpha+\langle\beta,x_i\rangle_{\mathcal{B}})\}}{1+\exp\{g_i(\alpha+\langle\beta,x_i\rangle_{\mathcal{B}})\}}, \tag{11}$$

where $\pi_i=\pi(1|x_i)$ and $g_i\in\{0,1\}$ is the observed group label for x_i. By computing $\hat{\beta}(t)=\hat{\mathbf{b}}^T\mathbf{B}_i^k(t)$, it is then possible to estimate the predictive probabilities from (8) and, for a given new observation x, the classification rule reduces to choosing $g\in\{0,1\}$ which maximizes the estimated conditional probability $\hat{\pi}(g|x)$.

Note that one can also choose to work directly with clr-transformed PDFs, using a proper ZB-spline basis. Resulting parameter $\beta(t)$ then can offer better interpretable information (Talská, Hron and Matys Grygar, 2021). Similarly, a principal component basis can be used instead of a fixed (Z)B-spline basis when (S)FPCA was performed prior classification.

3.3. Simplicial functional principal component regression

In many cases, it is necessary to classify observations into $G>2$ groups, a possibility which is allowed by the following method. As noticeable from its name, simplicial functional principal component regression (SFPCR) is a generalization of multivariate principal component regression (Varmuza and Filzmoser, 2009) which combines the results of SFPCA with regression, aiming to produce a regression model which can be then used for the class prediction. Here, the $\mathcal{B}^2$ counterparts to scores from (7) play the role of covariates whereas the class indices serve as response. Functional principal component regression has been studied, for data in L^2, in Reiss and Ogden (2007); it is here reformulated for density data in $\mathcal{B}^2$.

The keystone is the structure of response, which is a $N\times G$ matrix. With the assumption that each density belongs exclusively to one of the predefined classes, it is possible to reorder observations in the data set such that the response follows a pattern of matrix $\mathbf{R}$:

$$\mathbf{R}=\begin{pmatrix}
1 & -1 & \ldots & -1\\
\vdots & \vdots & \ldots & \vdots\\
1 & -1 & \ldots & -1\\
-1 & 1 & \ldots & -1\\
\vdots & \vdots & \ldots & \vdots\\
-1 & 1 & \ldots & -1\\
\ldots & \ldots & \ldots & \ldots\\
-1 & -1 & \ddots & 1\\
\vdots & \vdots & \ddots & \vdots\\
-1 & -1 & \ldots & 1
\end{pmatrix}$$

where $\mathbf{R}_{ig} = 1$ if the i-th observation belongs into group g, $\mathbf{R}_{ig} = -1$ otherwise. When using d SFPCs as newly obtained latent variables from the SFPCA step, $[N \times (d+1)]$-dimensional predictor matrix $\mathbf{P}$ is in form

$$\mathbf{P} = \begin{pmatrix} 1 & \langle x_1 \ominus \overline{x}, \zeta_1 \rangle_{\mathcal{B}} & \langle x_1 \ominus \overline{x}, \zeta_2 \rangle_{\mathcal{B}} & \dots & \langle x_1 \ominus \overline{x}, \zeta_d \rangle_{\mathcal{B}} \\ 1 & \langle x_2 \ominus \overline{x}, \zeta_1 \rangle_{\mathcal{B}} & \langle x_2 \ominus \overline{x}, \zeta_2 \rangle_{\mathcal{B}} & \dots & \langle x_2 \ominus \overline{x}, \zeta_d \rangle_{\mathcal{B}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \langle x_N \ominus \overline{x}, \zeta_1 \rangle_{\mathcal{B}} & \langle x_N \ominus \overline{x}, \zeta_2 \rangle_{\mathcal{B}} & \dots & \langle x_N \ominus \overline{x}, \zeta_d \rangle_{\mathcal{B}} \end{pmatrix},$$

where the first column is linked to the intercept and the remaining ones are formed from the scores corresponding to the orthogonal eigenfunctions $\{\zeta_i\}_1^d$, obtained as linear combinations of the ZB-spline basis elements.

The resulting regression model can be then written in a matrix form as

$$\mathbf{R} = \mathbf{PB} + \mathbf{E}$$

with a $(d+1) \times G$-dimensional matrix $\mathbf{B}$ of real coefficients or, using a scalar product for each element of $\mathbf{R}$, as

$$\mathbf{R}_{ig} = \mathbf{P}_{i\cdot} \cdot \mathbf{B}_{\cdot g} = \beta_{0g} + \sum_{j=1}^{d} \left(\beta_{jg} \ \ x_i \ominus \overline{x}, \zeta_j \ \ _{\mathcal{B}} \right) + \varepsilon_{ig} \tag{12}$$

for any $g \in \{0, \dots, G-1\}$ and $i \in \{1, \dots, N\}$.

It is then possible to estimate the $[(d+1) \times G]$-dimensional matrix $\mathbf{B}$ of regression coefficients using the OLS approach leading to

$$\widehat{\mathbf{B}} = (\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1}\mathbf{P}^{\mathrm{T}}\mathbf{R}, \tag{13}$$

where $\widehat{\mathbf{B}}_{\cdot g}$ carries the coefficients for the SFPCs corresponding to class g. Similarly to the previous method, the estimated response can be computed using the regression coefficients estimates in (12). The classification rule is then based on maximizing the estimated response for newly observed density x: the assigned group index corresponds to the column with the highest (positive) row value.

3.4. Functional linear discriminant analysis

The functional adaptation of linear discriminant analysis (Johnson and Wichern, 2007) (FLDA) has been already developed in James and Hastie (2001). A possible application in context of particle size distributions was presented in Pavlů et al. (2022). Although the method could be formulated directly in $\mathcal{B}^2$, we resort for simplicity to the equivalent formulation in L^2 based on clr-transform densities. In the following, $\mathrm{clr}(X_{ig})$, $\mathrm{clr}(x_{ig})$ will thus denote the clr-transformed forms of X_{ig}, x_{ig}, namely the i-th random/observed PDF in the g-th group.

Likewise in the multivariate case, the basic idea of FLDA lies in the concept of reducing the – essentially infinitely-dimensional – observations into a lower dimensional

discriminant space. Considering a d-dimensional discriminant space together with the concept of compositional splines (Machalová et al., 2021), each considered random functional observation $\mathrm{clr}(X_{ig})$ can be rewritten as

$$\mathrm{clr}(X_{ig})(t) \approx \mathbf{Z}(\boldsymbol{\nu}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_g + \boldsymbol{\gamma}_{ig})(t) + \varepsilon_{ig}(t), \tag{14}$$

$$g = 0,\ldots,G-1, \quad i = 1,\ldots,N_g, \quad \sum_{g=0}^{G-1} N_g = N, \quad t \in I$$

where $\mathbf{Z}$ stands for the common system of compositional spline basis functions, the argument in parentheses, $\boldsymbol{\nu}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_g + \boldsymbol{\gamma}_{ig}$, represents the decomposition of the spline coefficients and N_g stands for the number of observations from group g. In this form, it is possible to distinguish three main parts which form the spline coefficients and recognize their meaning:

- $\boldsymbol{\nu}_0 \in \mathbb{R}^d$ represents the main effect which is common for all observations regardless their (true) class;
- $\boldsymbol{\Lambda}\boldsymbol{\alpha}_g$ stands for the effect of the group, which is of the main interest in this case. While $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times d}$ serves as a form of standardization for the group effect (using $\boldsymbol{\Lambda}^T\mathbf{C}^{-1}\boldsymbol{\Lambda} = \mathbf{I}$), $\boldsymbol{\alpha}_g \in \mathbb{R}^d$ can be understood as representatives of the classes in the low-dimensional discriminant space, so-called *centroids*;
- $\boldsymbol{\gamma}_{ig} \in \mathbb{R}^d$ represents the individual effect of each separate observation.

In (14), ε_{ig} represents the random functional error. However, for an easier use, discretized observations $\mathrm{clr}(\mathbf{x}_{ig})$ of the clr-transformed densities will be considered, which can be easily obtained by numerical evaluation of functions over a certain grid of points $T = (t_1,\ldots,t_n) \in I$. This leads to decomposing the discretized $\mathrm{clr}(\mathbf{x}_{ig})$ as

$$\mathrm{clr}(\mathbf{x}_{ig}) = \mathbf{Z}_T(\boldsymbol{\nu}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_g + \boldsymbol{\gamma}_{ig}) + \boldsymbol{\varepsilon}_{ig}, \quad i = 1,\ldots,N_g, \quad g = 0,\ldots,G-1, \tag{15}$$

where $\mathbf{Z}_T = (\mathbf{Z}(t_1),\ldots,\mathbf{Z}(t_n))^\top$. In the standard L^2 setting, the vector $\boldsymbol{\varepsilon}_{ig}$ of measurement errors would follow a multivariate normal distribution $\mathrm{N}(\mathbf{0},\sigma^2\mathbf{I})$. Instead, given the zero integral constraint induced by the clr transformation, the random error has a singular normal distribution (Kwong and Iglewicz, 1996; Pavlů et al., 2022), i.e. $\boldsymbol{\varepsilon}_{ig} \sim \mathrm{N}(\mathbf{0},\Sigma = \sigma^2\mathbf{V}\mathbf{V}^\top)$ with $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_{N_g} - \mathbf{1}_{N_g\times N_g}/N_g$ being an idempotent matrix. The model, covering all necessary conditions, can be formulated as follows (James and Hastie, 2001),

$$\mathbf{x}_{ig} = \mathbf{Z}_T(\boldsymbol{\nu}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_g + \boldsymbol{\gamma}_{ig}) + \boldsymbol{\varepsilon}_{ig}, \quad i = 1,\ldots,N_g, \quad g = 0,\ldots,G-1, \tag{16}$$

$$\boldsymbol{\gamma}_{ig} \sim \mathrm{N}(\mathbf{0},\boldsymbol{\Gamma}), \qquad \boldsymbol{\varepsilon}_{ig} \sim \mathrm{N}(\mathbf{0},\sigma^2\mathbf{V}\mathbf{V}^\top), \quad \sum_{g=0}^{G-1}\boldsymbol{\alpha}_g = \mathbf{0}, \quad \boldsymbol{\Lambda}^\top\mathbf{B}^\top\mathbf{C}^{-1}\mathbf{B}\boldsymbol{\Lambda} = \mathbf{I}.$$

Using the additional assumption of independence of the observations, the parameters of the joint likelihood function of $\mathrm{clr}(\mathbf{x}_{ig})$ are estimated using the *EM* algorithm (Dempster,

Laird and Rubin, 1977); details in James and Hastie (2001); Pavlů et al. (2022)). The final estimates of $\hat{\boldsymbol{\alpha}}_g$ can also be used for graphical representation of the group centres in the discriminant space with h components in the form of centroids - an example of such visualisation can be seen in Figure 3.

The final decision is then made based on the classification rule which minimizes the criterion

$$\underset{g\in\{0,\ldots,G-1\}}{\operatorname{argmin}} \left(||\hat{\boldsymbol{\alpha}}_{\mathbf{x}_{ig}} - \hat{\boldsymbol{\alpha}}_g||^2 - 2\ln\frac{N_g}{N} \right). \tag{17}$$

The ratio $\frac{N_g}{N}$ here corresponds to the estimation of prior probability of group g being represented in the original data set. Note that, up to the constant, (17) essentially minimizes the Euclidean distance between the linear discriminant of $\operatorname{clr}(\mathbf{x}_{ig})$ and the class centroids. The index g which attains the minimization of the given criterion is then chosen as the estimated class for $\operatorname{clr}(\mathbf{x}_{ig})$.

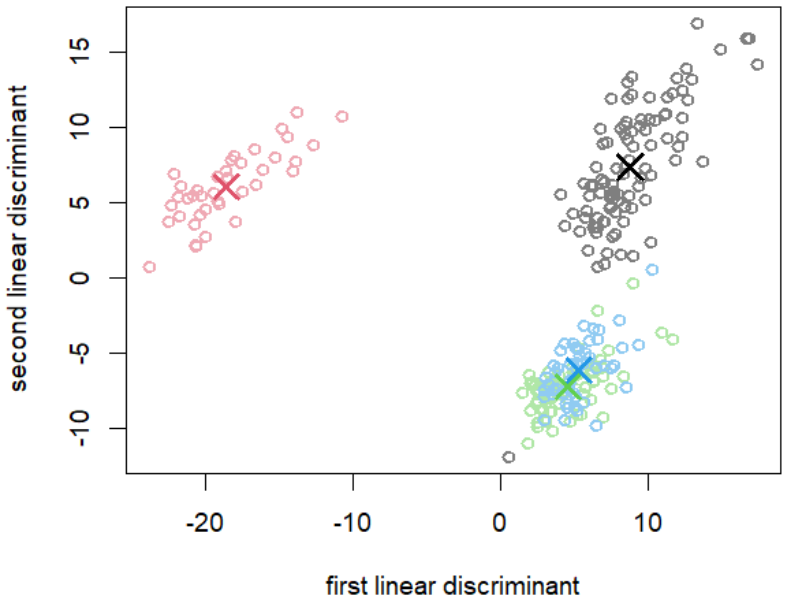

Figure 3. *Clr-transformed particle size distributions from Figure 1 displayed in the two-dimensional linear discriminant space.*

3.5. Small-ball probability approach

The method, introduced in detail in Bongiorno and Goia (2016), was already formulated for the general Hilbert space setting; therefore, it is only briefly summarized here. The main idea is based on the so-called *small-ball probability* (SMBP), which can be defined

for x in any Hilbert space H (in the following, $H = \mathcal{B}^2$) as the limit behaviour of

$$\varphi(x,h) = P(||X - x|| < h), \quad h > 0. \tag{18}$$

for $h \to 0$. Equation (18) essentially captures how the probability law of X is concentrated around x. Given Y as in (6), $\varphi(x,h)$ can be rewritten as the mixture

$$\varphi(x,h) = \sum_{g=0}^{G-1} \frac{N_g}{N} \varphi(x,h|g), \tag{19}$$

where

$$\varphi(x,h|g) = P(||X - x|| < h | Y = g).$$

Additional conditions on the eigenvalues decay of the covariance operator of X described in Bongiorno and Goia (2016) ensure that, when h tends to zero, it is possible to find $d = d(h)$ such that the small-ball probability factorizes as

$$\varphi(x,h) \sim f_d(\theta_1, \ldots, \theta_d)\phi(d,h), \tag{20}$$

where $f_d(\theta_d)$ is the PDF of the scores along the first d functional principal components evaluated at $\theta_d = (\theta_1, \ldots, \theta_d)^T$ with $\theta_j = \langle x, \xi_j \rangle$, ξ_j are the SFPCs (as appearing in (7)) and $\phi(d,h)$ stands for the volume of the d-dimensional ball of radius h.

Classification itself is based on a slight modification of the Bayes classification rule – specifically, the new observation x is assigned to the g-th group if, for small values of h, the posterior probability $P(Y = g | \, ||X - x|| < h)$ is maximal over $\{0, \ldots, G-1\}$. Thanks to factorization (20) and for a large enough d, it is possible to rewrite and simplify the classification rule in the following way: observation x is assigned to class g if

$$\frac{P(G = g | \, ||X - x|| < h))}{P(G = g' | \, ||X - x|| < h)} = \frac{(N_g/N)\varphi(x,h|g)}{(N_{g'}/N)\varphi(x,h|g')} \sim \frac{N_g f_{d|g}(\theta_d^{(g)})}{N_{g'} f_{d|g'}(\theta_d^{(g')})} > 1, \tag{21}$$

for any $g' \neq g$ and h tending to zero, where $\theta_d^{(g)}$ is the vector containing the first d principal components computed for the g-th group.

The resulting classifier is

$$\rho(x,d) = \underset{g=0,\ldots,G-1}{\operatorname{argmax}} \frac{N_g}{N} f_d(x|g) \tag{22}$$

that can be approximated by means of a kernel density estimator leading to

$$\hat{\rho}(x,d) = \underset{g=0,\ldots,G-1}{\operatorname{argmax}} \frac{N_g}{N} \sum_{i=1}^{n} \mathbb{I}_{\{g_i^* = g\}} K_{H_g}(||\hat{\Pi}_{g,d}(X_i - x)||), \tag{23}$$

where $K_{H_g}(u) = \det(\mathbf{H}_g)^{-\frac{1}{2}} K(\mathbf{H}_g)^{-1} 2u$ with a kernel function K. In this context, $\mathbf{H}_g$ stands for a symmetric positive semi-definite matrix defining the bandwidth of the kernel, and $\hat{\Pi}_{g,d}$ denotes the projector operator over the newly-obtained subspace spanned by the first d eigenfunctions of the empirical covariance operator $\widehat{\mathbf{V}}_g$. As the index g suggests, $\widehat{\mathbf{V}}_g$ are estimated individually within each $g \in \{0, \ldots, G-1\}$.

3.6. Functional k-nearest neighbours algorithm

Classification based on the k-nearest neighbours (FKNN) (Burba, Ferraty and Vieu, 2009) is a popular tool for any type and origin of data because it commonly works well under very weak assumptions. Given a new PDF x_0, the decision rule of the k-nearest neighbours algorithm is based on looking for the k closest observations of x_0 with respect to a given metric (or semi-metric), and assigning x_0 to the most frequently represented class among its neighbours. Consistently with the Bayes space approach considered here, the $\mathcal{B}^2$ distance between original PDFs is taken, which is equivalent to the L^2 distance between the respective clr-transformed curves. In $\mathcal{B}^2$, the distance between two functional objects is given by (3) with an equivalent formulation in the L^2

$$d(\mathrm{clr}(f_1),\mathrm{clr}(f_2)) = ||\mathrm{clr}(f_1)-\mathrm{clr}(f_2)||^2 = \int_I [\mathrm{clr}(f_1)(s)-\mathrm{clr}(f_2)(s)]^2\,\mathrm{d}s. \qquad (24)$$

The group index g^* associated with the highest frequency of neighbours is then assigned as a group label for x_0. In case there are two or more groups with the same highest occurrence, a secondary criterion may be considered for those groups, e.g., in the form of considering the sum of distances from x_0 to all observations from these groups included within k-nearest neighbors of x_0. The observation x_0 is then assigned to the group with the smallest sum of distances.

4. Simulation study

The data analyses included in the two following sections were performed using the software environment R (R Core Team, 2021) and its packages *robCompositions* (Templ, Hron and Filzmoser, 2011) and *fda*. To implement FLDA in the $\mathcal{B}^2$ case, a code available at http://faculty.mar-shall.usc.edu/gareth-james/Research/Research.html by authors of James and Hastie (2001) was used as a starting point and adapted to the case of clr-transformed PDFs.

Since there is usually no universal rule for the choice of key parameters in the classification methods under consideration, they were evaluated each time for multiple values of parameters. The parameters included in the final comparisons were then chosen to ensure the best results of the method on the given data set. To avoid overparametrized models, the maximum value of parameters was determined prior to the analysis of the misclassification rates. For methods based on SFPCA (FLR, FPCR, SMBP), the fraction of explained variability (FEV) was taken into account - the maximum number of FPCs was chosen such that these FPCs would explain at least 98% of data variability (assuming the rest is caused by noise). For FKNN, $k = \sqrt{n}$ with n denoting the size of the training data set served the same purpose. In case of FLDA, the highest possible dimension of the discriminant space was given by the number of considered groups.

Finally, since FLR is designed for binary classification, it will not be considered in a multi-class case, although it could be used for classification of one group against the rest of the data set (one vs. all scenario).

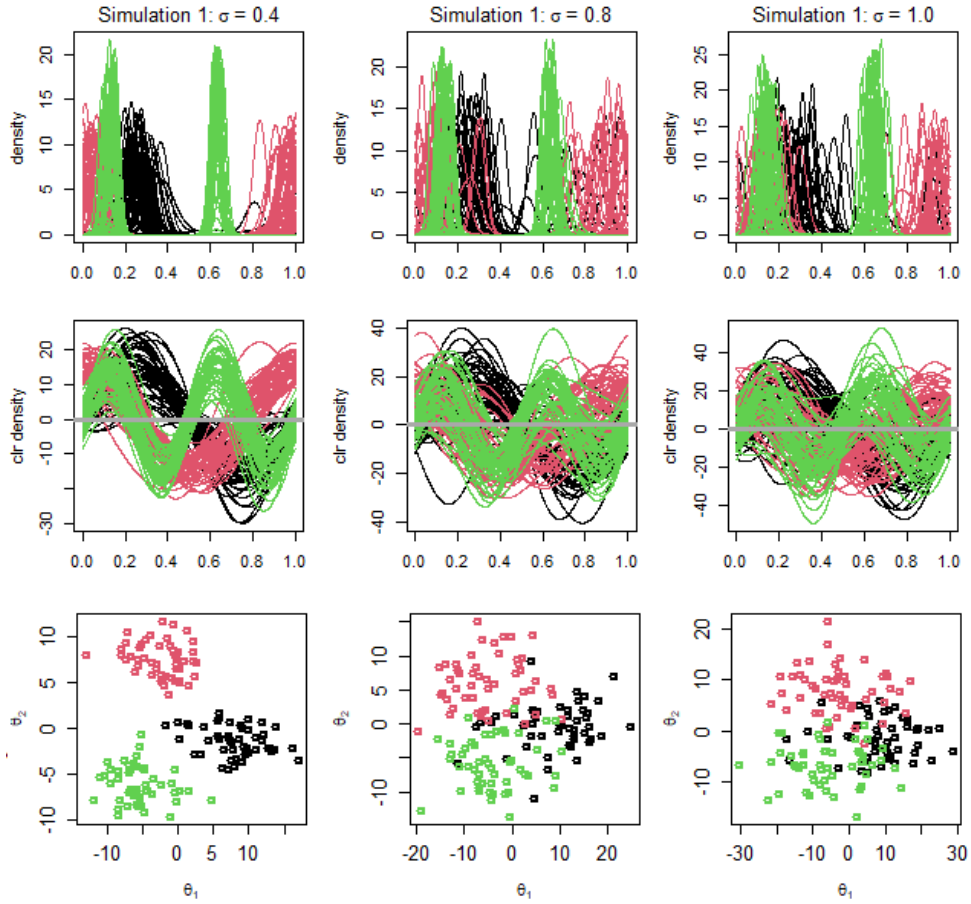

Figure 4. *Simulation 1: Data set (PDFs and clr-transformed PDFs) and the 2-dimensional scores based on the different value of parameter* σ.

4.1. Simulation 1

The first data set of PDFs was designed to provide a favourable scenario for the linear methods, namely FPCR and FLDA, thanks to the convex shape of the low-dimensional score clusters resulting from SFPCA. The data set used for this simulation is defined over a domain $[0,1]$ using the first four non-constant elements of Fourier basis, i.e. the process

$$X = \sum_{n=1}^{2} A_n \cos(2\pi nx) + B_n \sin(2\pi nx), \quad x \in [0,1]. \tag{25}$$

The data in the three proportional clusters were independently generated by sampling a set of coefficients (A_1, B_1, A_2, B_2) from a Gaussian distribution $N(\mu_g, \sigma^2\mathbf{I})$, $g = 0,1,2$ with $\mu_0 = (1,0,0,0)^T, \mu_1 = (0,1,0,0)^T, \mu_2 = (0,0,1,0)^T$. An example of the simulated data set for different choices of the scatter parameter σ is presented in Figure 4.

To get more information from the simulation, the influence of both the size of the training set and the variance was further studied. For the first setting, σ was set to 0.4 to determine the impact of the sample size. To this end, the number of observations n_i in each of classes was set to 10, 20, 50, 100 and 300. The size of the testing set stayed the same throughout the simulation, being equal to 150 (50 observations per group) and generated from the same setting described above. For each sample size and each parameter, $N = 100$ training and testing data sets were produced and used for the evaluation of the quality of classification. Table 1 summarizes the results for the optimal parameters, specifically mean and standard deviation, as well as their robust counterparts (median and median absolute deviation - MAD) of the estimated misclassification error, here defined as the proportion of misclassified observations in the sample.

Table 1. *Simulation 1: Summary of results with changing parameter n_i ($\sigma = 0.4$). Overall, FLDA and FKNN seem to perform best within the given setting.*

n_i	Algorithm	Parameter	Miscl. error - mean	Miscl. error - sd	Miscl. error - median	Miscl. error - MAD
10	FPCR	$d = 2$	0.0162	0.0107	0.0133	0.0099
	FLDA	$d = 2$	0.0085	0.0124	0.0067	0.0099
	SMBP	$d = 2$	0.1030	0.0654	0.0900	0.0544
	FKNN	$k = 4$	0.0055	0.0065	0.0067	0.0099
20	FPCR	$d = 2$	0.0160	0.0103	0.0133	0.0099
	FLDA	$d = 2$	0.0057	0.0075	0.0067	0.0099
	SMBP	$d = 2$	0.0873	0.0548	0.0700	0.0346
	FKNN	$k = 4$	0.0047	0.0056	0.0000	0.0000
50	FPCR	$d = 2$	0.0164	0.0096	0.0133	0.0099
	FLDA	$d = 2$	0.0049	0.0061	0.0000	0.0000
	SMBP	$d = 2$	0.0820	0.0278	0.0067	0.0297
	FKNN	$k = 3$	0.0034	0.0046	0.0000	0.0000
100	FPCR	$d = 2$	0.0149	0.0103	0.1000	0.0099
	FLDA	$d = 2$	0.0037	0.0054	0.0000	0.0000
	SMBP	$d = 2$	0.0757	0.0252	0.0000	0.0297
	FKNN	$k = 9$	0.0019	0.0041	0.0000	0.0000
300	FPCR	$d = 2$	0.0174	0.0102	0.0200	0.0099
	FLDA	$d = 2$	0.0045	0.0059	0.0000	0.0000
	SMBP	$d = 2$	0.0763	0.0218	0.0000	0.0198
	FKNN	$k = 8$	0.0028	0.0041	0.0000	0.0000

To explore the influence of the changing variability of scores, the size of both training and testing data sets was set to 150 observations (50 per group), while the σ was gradually set to 0.2, 0.4, 0.6, 0.8 and 1 - keeping the rest of the simulation the same. Once again, results are displayed in Table 2.

The results of the simulation are not surprising; we can see that, given the shape of the score clusters, the linear methods perform quite well and can compete even with the seemingly most universal FKNN. Different, and rather non-elliptic score clusters could be expected when PDFs were considered in the original $\mathcal{B}^2$ space (Figure 4, upper row). Obviously, small values of PDFs form the main source of variability which results from

Table 2. *Simulation 1: Summary of results with changing parameter σ ($n_i = 50$). Although the results seem quite similar with low σs, the development of the mean misclassification error suggests stronger potential for FLDA and FKNN in case of higher variance.*

σ	Algorithm	Parameter	Miscl. error - mean	Miscl. error - sd	Miscl. error - median	Miscl. error - MAD
0.4	FPCR	$d = 2$	0.0173	0.0097	0.0133	0.0099
	FLDA	$d = 2$	0.0047	0.0067	0.0000	0.0000
	SMBP	$d = 3$	0.0041	0.0054	0.0000	0.0000
	FKNN	$k = 3$	0.0030	0.0043	0.0000	0.0000
0.6	FPCR	$d = 2$	0.0901	0.0250	0.0900	0.0297
	FLDA	$d = 2$	0.0377	0.0157	0.0400	0.0199
	SMBP	$d = 3$	0.0414	0.0246	0.0400	0.0199
	FKNN	$k = 12$	0.0332	0.0144	0.0333	0.0099
0.8	FPCR	$d = 3$	0.1630	0.0318	0.1600	0.0297
	FLDA	$d = 2$	0.0996	0.0243	0.0933	0.0297
	SMBP	$d = 3$	0.1143	0.0288	0.1133	0.0297
	FKNN	$k = 8$	0.0970	0.0261	0.1000	0.0297
1	FPCR	$d = 3$	0.1984	0.0304	0.2000	0.0297
	FLDA	$d = 2$	0.1693	0.0307	0.1667	0.0297
	SMBP	$d = 3$	0.1907	0.0366	0.1867	0.0395
	FKNN	$k = 10$	0.1659	0.0306	0.1667	0.0297

their relative scale. This would, however, be ignored if the original PDFs were classified as L^2 objects. Note that FKNN appears to outperform the other methods in most scenarios based on the mean misclassification error, but results appears to be substantially equivalent to those of FLDA if accounting for the standard deviation of the errors. In this case, while FKNN provides a simple solution to the classification problem, a strength of FLDA relies in the interpretability of the linear discriminants, which can be used to shed light on the data set itself.

4.2. Simulation 2

For this simulation setting, the training data set is produced in the same way as in Section 3.1 in Bongiorno and Goia (2016) (except for the different number L of basis functions). For the sake of completeness, a brief description of the simulation design is provided in what follows.

The idea behind this simulation setting is to deal with scores that are not spherically clustered but present a nonlinear data structure. To this end, the following functional basis expansion

$$X_{ig}(t) = \sum_{l=1}^{L} \sqrt{\beta_l}\, \tau_{igl} \xi_l(t), \quad t \in [0,1],\ i = 1,\ldots,n_g \ \text{ and } \ g = 0,\ldots,G-1 \tag{26}$$

is used, where $\beta_l = 0.7 \times 3^{-l}$ ($l = 1,\ldots,L = 30$), $\xi_l(t)$ is the l-th non-constant element of the Fourier basis (extension of (25)) and $G = 2$. To avoid spherical shaped groups of

scores, uncorrelated but dependent coefficients $(\tau_{igl})_{l=1}^{L}$,

$$\begin{cases} \tau_{ig1} &= \sin(\vartheta_i)\cos(\frac{\pi}{2}\mathbb{I}_{\{g=1\}}) + \sigma\varepsilon_{i,1}, \\ \tau_{ig2} &= \sin(\vartheta_i)\sin(\frac{\pi}{2}\mathbb{I}_{\{g=1\}}) + \sigma\varepsilon_{i,2}, \\ \tau_{ig3} &= \cos(\vartheta_i) + (-k)^g + \sigma\varepsilon_{i,3}, \\ \tau_{igl} &= \sqrt{0.1}\varepsilon_{i,l}, \qquad\qquad 4 \le l \le L, \end{cases}$$

were generated with (ϑ_i) i.i.d. from Beta(5,5) scaled on $[-\pi,\pi]$, $(\varepsilon_{i,l})_{l=1}^{L} \overset{i.i.d.}{\sim} N(0,1)$ representing the Gaussian noise, and $(-k)^g$ (here $k = 0.5$) standing for the vertical translation. This way, it is possible to obtain FPCA scores which replicate the τ's structure (shown for different σ's in the bottom row of Figure 5). This setting ensures that the decay of the eigenvalues of the covariance operator is fast enough to guarantee (20) (see Bongiorno and Goia (2016) for more details). Due to this nonlinearity, one would expect a weaker performance of the (linear) parametric methods, namely FLR, FPCR and FLDA.

Likewise as in the previous simulation study, the impact of the size of the training set and the amount of random noise were tracked. The size of the two groups was set to $n_1 \in \{10, 20, 50, 100, 300\}$ while the noise was determined by parameter $\sigma \in \left\{\sqrt{0.001}, \sqrt{0.005}, \sqrt{0.01}, \sqrt{0.05}, \sqrt{0.1}\right\}$ – its influence on both the coefficients and the resulting data sets can be seen in Figure 5. The size of the testing set remained 100 (50+50).

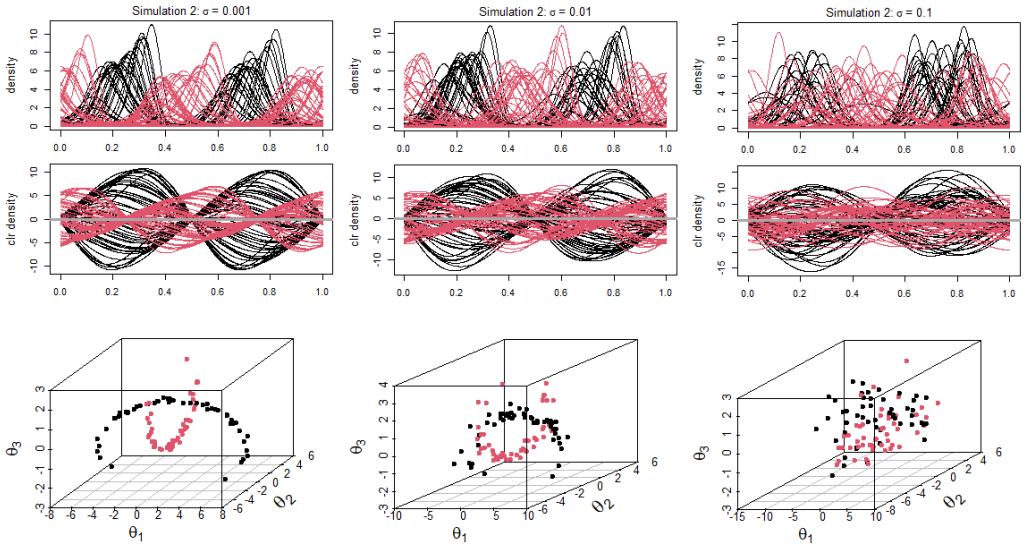

Figure 5. *Simulation 2: Data set (original and clr) and the 3-dimensional scores representation based on the different value of parameter* σ.

Table 3. *Simulation 2: Summary of results with changing parameter n_i. Altogether, SMBP and FKNN seem to perform best based both on the mean/median misclassification error as well as the significantly lower standard deviation/MAD.*

n_i	Algorithm	Parameter	Miscl. error - sd	Miscl. error - mean	Miscl. error - median	Miscl. error - MAD
10	FLR	$d=3$	0.3713	0.0845	0.3300	0.0593
	FPCR	$d=3$	0.3393	0.0812	0.3250	0.5189
	FLDA	$d=1$	0.3243	0.0561	0.3200	0.0593
	SMBP	$d=3$	0.0104	0.0270	0.0000	0.0000
	FKNN	$k=1$	0.0096	0.0213	0.0000	0.0000
20	FLR	$d=3$	0.3152	0.0701	0.3000	0.0445
	FPCR	$d=3$	0.3147	0.0708	0.3050	0.0519
	FLDA	$d=1$	0.3236	0.0670	0.3200	0.0445
	SMBP	$d=3$	0.0020	0.0043	0.0000	0.0000
	FKNN	$k=2$	0.0009	0.0032	0.0000	0.0000
50	FLR	$d=3$	0.3093	0.0452	0.3100	0.0445
	FPCR	$d=3$	0.3088	0.0456	0.3100	0.0445
	FLDA	$d=1$	0.3255	0.0523	0.3200	0.0445
	SMBP	$d=3$	0.0022	0.0005	0.0000	0.0000
	FKNN	$k=2$	0.0006	0.0028	0.0000	0.0000
100	FLR	$d=3$	0.2971	0.0429	0.3000	0.0445
	FPCR	$d=3$	0.2972	0.0438	0.3000	0.0445
	FLDA	$d=1$	0.3091	0.0501	0.3200	0.0445
	SMBP	$d=3$	0.0001	0.0010	0.0000	0.0000
	FKNN	$k=4$	0.0004	0.0020	0.0000	0.0000
300	FLR	$d=3$	0.2941	0.0410	0.3000	0.0445
	FPCR	$d=3$	0.2941	0.0412	0.2950	0.0445
	FLDA	$d=1$	0.3010	0.3200	0.0417	0.0445
	SMBP	$d=3$	0.0000	0.0000	0.0000	0.0000
	FKNN	$k=2$	0.0001	0.0010	0.0000	0.0000

The results obtained for increasing n_i which affects size of the training set (constant $\sigma = \sqrt{0.005}$) are shown in Table 3, while the latter case (constant $n_i = 50$, changing σ) is summarized in Table 4. Altogether it can be observed that the non- and semi-parametric approaches indeed work better than the parametric ones (taking into account the standard deviation of the misclassification error, they can be considered interchangeable). Here, both FKNN and SMBP are designed to catch nonlinear correlation in data and hence outperform linear methods.

5. Real-world applications

In this section, the proposed methods will be used for classification of two real-world data sets of different origin. The first one deals with age distributions of men and women in Upper Austria (used in Hron et al. (2016)) while the second one contains particle size distributions from four measuring sites in the Czech Republic.

Table 4. *Simulation 2: Summary of results with changing parameter σ. SMBP and FKNN demonstrate their strength once again - especially for the cases with lower σs, as the differences between classification results seem to decrease with larger variability in the data.*

σ	Algorithm	Parameter	Miscl. error - sd	Miscl. error - mean	Miscl. error - median	Miscl. error - MAD
0.001	FLR	$d = 3$	0.2990	0.0422	0.2900	0.0445
	FPCR	$d = 3$	0.2985	0.0416	0.3000	0.0445
	FLDA	$d = 1$	0.3080	0.0472	0.3100	0.0445
	SMBP	$d = 4$	0.0007	0.0026	0.0000	0.0000
	FKNN	$k = 2$	0.0002	0.0014	0.0000	0.0000
0.005	FLR	$d = 3$	0.3141	0.0510	0.3150	0.0519
	FPCR	$d = 3$	0.3134	0.0508	0.3100	0.0593
	FLDA	$d = 1$	0.3212	0.0514	0.3100	0.0445
	SMBP	$d = 3$	0.0008	0.0027	0.0000	0.0000
	FKNN	$k = 2$	0.0004	0.0020	0.0000	0.0000
0.01	FLR	$d = 3$	0.3037	0.0540	0.3100	0.0519
	FPCR	$d = 3$	0.3047	0.0530	0.3100	0.0445
	FLDA	$d = 1$	0.3200	0.0594	0.3200	0.0593
	SMBP	$d = 3$	0.0012	0.0036	0.0000	0.0000
	FKNN	$k = 2$	0.0008	0.0027	0.0000	0.0000
0.05	FLR	$d = 3$	0.3503	0.0761	0.3300	0.0593
	FPCR	$d = 3$	0.3511	0.0.0765	0.3300	0.0667
	FLDA	$d = 1$	0.3642	0.0713	0.3550	0.0445
	SMBP	$d = 3$	0.0511	0.0231	0.0500	0.0148
	FKNN	$k = 6$	0.0503	0.0219	0.0500	0.0148
0.1	FLR	$d = 3$	0.3737	0.0620	0.3600	0.0593
	FPCR	$d = 3$	0.3732	0.0610	0.3600	0.0593
	FLDA	$d = 1$	0.3970	0.0569	0.3900	0.0593
	SMBP	$d = 3$	0.1505	0.0349	0.1500	0.0297
	FKNN	$k = 8$	0.1462	0.0370	0.1500	0.0445

5.1. Age distribution

The data set from Hron et al. (2016) contains age distributions from 57 municipalities in Upper Austria (Figure 6). For each district, age distributions for both men and women were observed. Accordingly, the aim is to classify these observations by gender.

All presented methods were evaluated using the 5-fold cross-validation.The best results for this data set were obtained using the following parameters: $d_{\text{FLR}} = d_{\text{FPCR}} = 3$, $d_{\text{FLDA}} = d_{\text{SmBP}} = 2$, $k = 5$. Table 5 then summarizes the results, while Figure 7 displays the 2D score representation of the original data set. From the results, it is evident that both FLR and FPCR as well as the nonparametric method FKNN perform very well – nevertheless, the mean misclassification rate of all presented methods is below or around 5%.

Hereby it is interesting to observe the main source of variability, and consequently also classification for the original versus the clr-transformed PDFs. While when looking at the original PDFs (Figure 6, left) one would guess ages around 30 and 80 as the

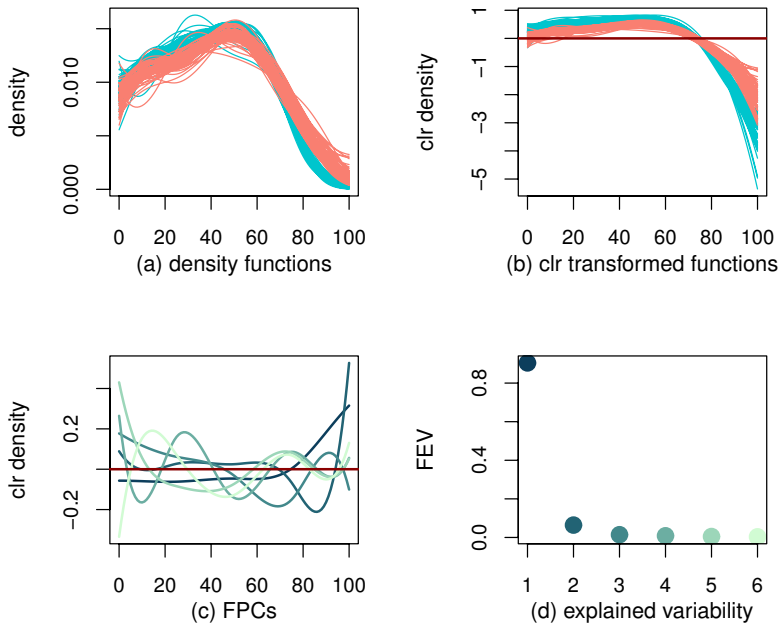

Figure 6. *Age distribution data. In (a) and (b) colors are given according to their classes (blue - men, pink - women).*

main source of information for classification, for clr-transformed PDFs (Figure 6, right) this comes clearly from the oldest age categories. The latter source is quite natural as it corresponds to mean lifetime which forms clearly the most important difference between men and women age distributions.

Table 5. *Classification results - age distributions.*

Algorithm	Parameter	Miscl. error - mean	Miscl. error - sd	Miscl. error - median	Miscl. error - MAD
FLR	$d = 3$	0.0025	0.0102	0.0000	0.0000
FPCR	$d = 3$	0.0037	0.0122	0.0000	0.0000
FLDA	$d = 2$	0.0514	0.0414	0.0455	0.0562
SMBP	$d = 2$	0.0126	0.0220	0.0000	0.0000
FKNN	$k = 3$	0.0048	0.0147	0.0000	0.0000

Given that functional logistic regression performs best in this case, we analyse further the shape of the functional regression parameter, as this might be indicative of the discrimination power of the different portions of the domain of the PDFs. As indicated in Section 3.2, the interpretation of the regression parameter (β) is straightforward in the clr space, because the resulting zero-integral function forms naturally a contrast between positive and negative values which can be, accordingly, assigned to effects of either classes. Specifically, positive values of the regression parameter (in its clr representation) for a specific part of the domain indicate a tendency toward the class $g = 1$, (here indicating women) and vice versa for negative values (Talská et al., 2021). As it

can be observed from Figure 8, the interval $[60, 100]$, corresponding to the oldest age groups, carries significant differences between the two gender groups. The high absolute effect of $\beta(t)$ on interval $[80, 100]$ then confirms that this age group discriminates the most between gender groups, positive values being associated to the group of women and their higher life expectancy. This fact can be possibly used also for weighting of the domain of PDFs (van den Boogaart et al., 2014; Talská et al., 2020) to highlight the effect of the part of domain which contributes at most to classification between both gender groups.

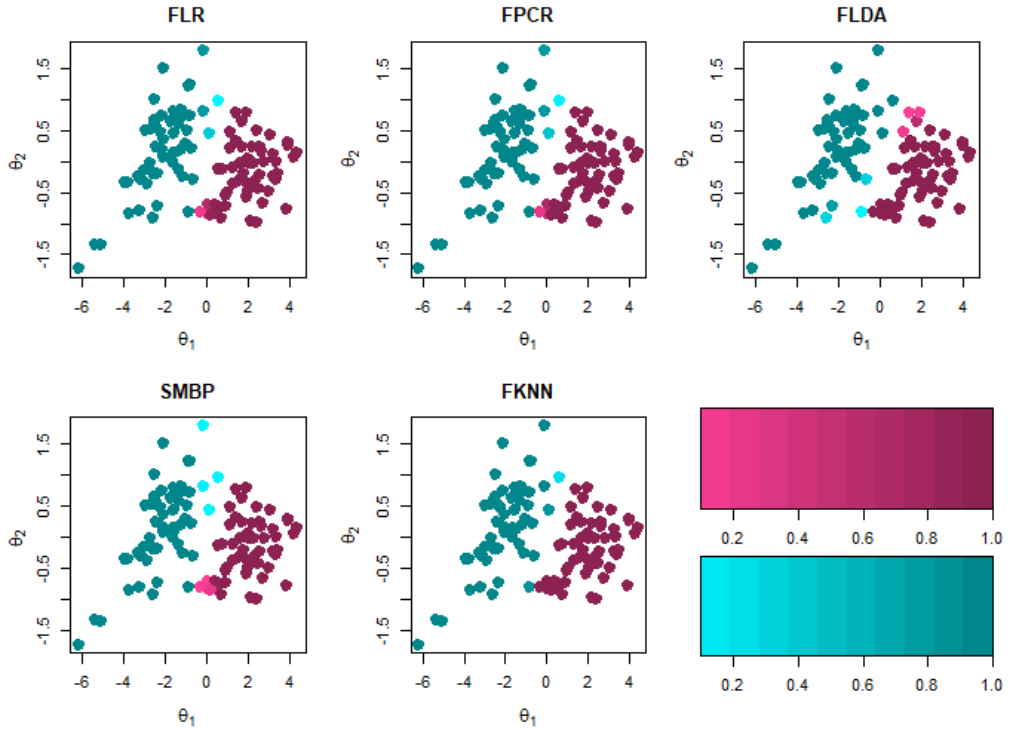

Figure 7. *Age distribution data: visualisation of classification results of each observation during CV individually through their 2-dimensional scores representation. The darker the color, the higher proportion of scenarios where the observation was classified correctly.*

5.2. Particle size distribution example

The data set discussed in this section contains particle size distributions (PSDs) measured at four different locations in the Czech Republic (Dobšice, Brodek u Prostějova, Rozvadovice, Ivaň; further denoted as classes 1-4) with locations playing the role of classes (Šimíček et al., 2021). The original data set consists of 250 vectors (each corresponding to a unique discretized PSD), which were smoothed using the compositional splines. Here, the different classes are represented unevenly, with sample sizes 96, 39, 66, and 49. As the site of origin (and therefore the correct classification) of the measured

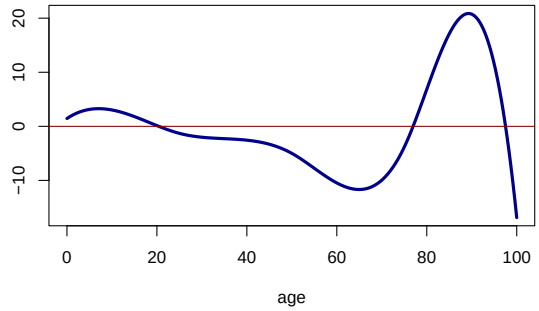

Figure 8. *Estimated functional regression parameter $\beta(t)$ corresponding to functional logistic regression with the age distribution data set.*

samples is known, it is possible to estimate the required parameters and to examine the quality of the classification model via cross validation. The data set is displayed in Figure 9.

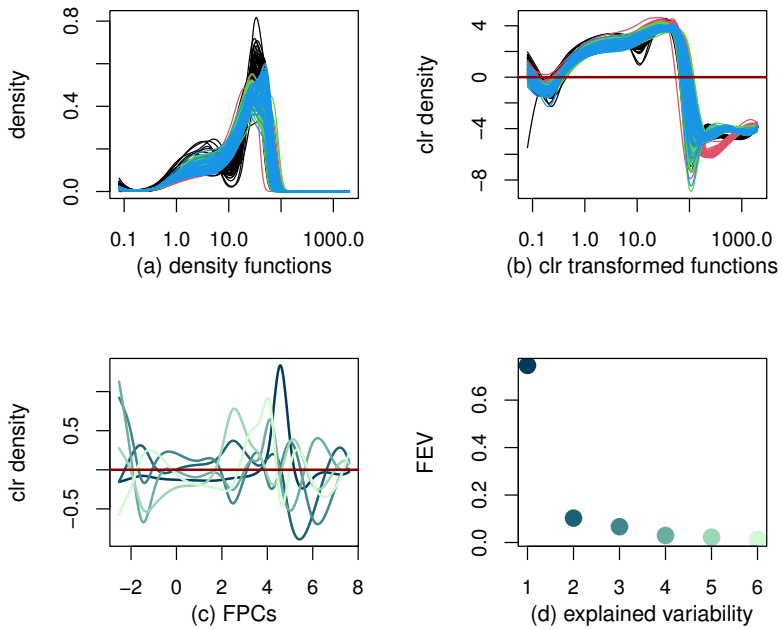

Figure 9. *Particle size distribution data. In (a) and (b) colors are given according to their classes.*

Even from a visual inspection of Figure 9b, it is possible to capture some differences between groups, namely class 1 (black) is deviating from the common behaviour around points 0.3 μm and 10 μm and class 2 (red) in the neighbourhood of 300 μm; nevertheless, the effect of the latter class would not be observable from Figure 9a. On the other side, classes 3 and 4 are practically indistinguishable. This demonstrates once again that a proper representation of the original PDFs is crucial to assess the source of classifica-

tion into given groups. Clearly, the clr transformation highlights the role of small values of PDFs which is fully in line with the relative scale of PDFs. Accordingly, while no differences between groups can be observed for fractions above 100 μm with the original PDFs, this is clearly the opposite case when their clr-transformed counterparts are considered. Moreover, also here it might be beneficial to change the reference measure, e.g. to B-mean as in van den Boogaart et al. (2014) or to a user designed reference, to provide a better insight and prospectively also better classification.

Again, the 5-fold cross-validation was performed for all presented methods resulting in the following optimal parameters: $d_{\text{FPCR}} = 6$, $d_{\text{FLDA}} = 2$, $d_{\text{SmBP}} = 6$, $k = 4$. Overall results are summarized in Table 6, confirming the dominance of FKNN. An interesting effect can be observed from Figure 10, showing the 3D scores representation of the given data set. It seems that, although group 2 (red) is quite clearly different from the rest, both FPCR and SMBP struggle to capture these differences. On the other hand, the task of differentiating between groups 3 (green) and 4 (blue) seems to be difficult for all methods, although FKNN performs best even in this case.

Table 6. *Classification results - particle size distributions.*

Algorithm	Parameter	Miscl. error - mean	Miscl. error - sd	Miscl. error - median	Miscl. error - MAD
FPCR	$d = 6$	0.1709	0.0348	0.1730	0.0399
FLDA	$d = 2$	0.1872	0.0440	0.1875	0.0408
SMBP	$d = 6$	0.1795	0.0469	0.1800	0.0593
FKNN	$k = 4$	0.0958	0.0355	0.1000	0.0297

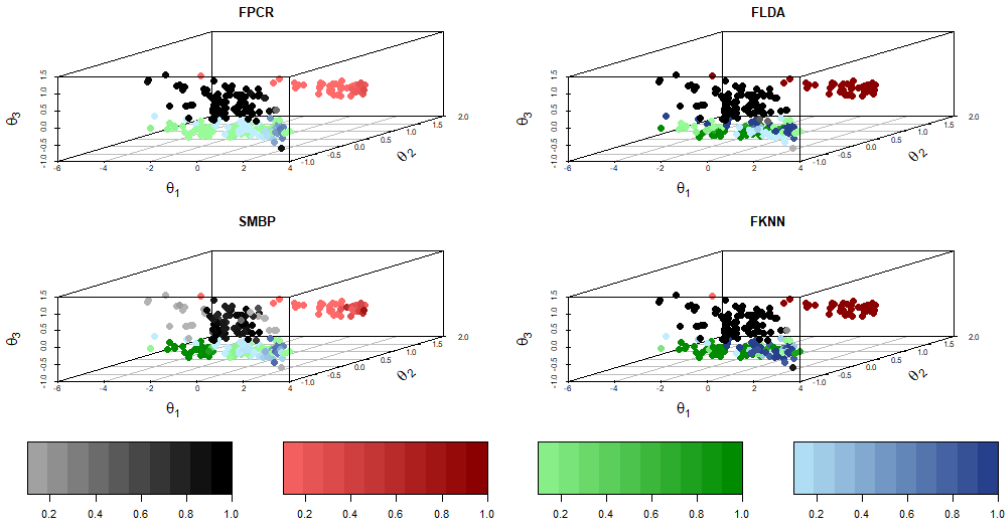

Figure 10. *Particle size distribution data: visualisation of classification results of each observation during CV individually through their 3-dimensional scores representation. The darker the color, the higher proportion of scenarios where the observation was classified correctly.*

6. Conclusions

The need of considering a proper sample space for representation of functional distributional data (expressed usually in terms of probability density functions) has been addressed throughout the years. This paper aims to provide a first concise overview of classification methods in the context of the Bayes space approach. In practice, it is common to transform PDFs into the L^2 space using the clr transformation and then proceed with data analysis - however, it is important to emphasize the embedding of the methods in the (original) Bayes space framework.

When looking at the results of classification performed on both simulated and real data sets, the nonparametric FKNN approach seems to outperform all of its competitors. On one hand, this should not be too surprising as, with FKNN, the "complete" information contained in the data is used (no simplification through parametrization is done). On the other hand, this way one does not have any model to work with, no parameters to assign interpretation to. With the remaining methods, it is possible to obtain additional information and interpretability from the parameters which can be useful especially for real-world applications. As demonstrated in the paper, this is indeed an important advantage of methods like FLR and FPCR. Considering the relative scale of PDFs by the clr transformation opens new perspectives concerning the discrimination power of portions of the domain unlike it could be deduced from the original PDFs. Note also that the choice of a proper classification method could be different, and even counterintuitive, if based on assessing the data structure of the original PDFs in terms of the L^2 space.

Another aspect worth mentioning is connected to the fact that the performance of the methods can be severely influenced by choosing different parameters and/or by different proportionality occurring in the data. The safest bet is to test a few sets of parameters and choose, in a data-driven fashion (e.g., via cross-validation) the combination which performs the best. Also choosing a more appropriate reference measure than the default uniform one (van den Boogaart et al., 2014; Talská et al., 2020) can contribute to a better classification performance.

Although in this paper classification of univariate densities was presented, the Bayes space methodology offers the possibility of an extension to multivariate densities, which is nowadays equipped with an orthogonal decomposition of PDFs into independent and interactive parts (Genest, Hron and Nešlehová (2023); Hron, Machalová and Menafoglio, 2022). The development of proper methods for the data analysis and classification in this very cutting-edge setting will be of our primary interest in the near future.

Acknowledgements

Ivana Pavlů and Karel Hron gratefully acknowledge the support of the Grant No. GA22-15684L of the Grant Agency of the Czech Republic and the Grants IGA_PrF_2022_008 and IGA_PrF_2023_009. Enea G. Bongiorno is member of the Gruppo Nazionale per l´Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto

Nazionale di Alta Matematica (INdAM). The financial support of Università del Piemonte Orientale is acknowledged by Enea G. Bongiorno.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Bongiorno, E. G. and Goia, A. (2016). Classification methods for Hilbert data based on surrogate density. *Computational Statistics and Data Analysis*, 9, 204–222.

Burba, F., Ferraty, F. and Vieu, P. (2009). k-nearest neighbors method in functional nonparametric regression. *Journal of Nonparametric Statistics*, 21, 453–469.

De Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39, 1–38.

Eaton, M. L. (1983). *Multivariate Statistics: A Vector Space Approach*. Wiley Series in Probability and Statistics. New York: Wiley.

Egozcue, J. J., Díaz-Barrero, J. L. and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica*, 22, 1175–1182.

Egozcue, J. J., Pawlowsky-Glahn, V., Tolosana-Delgado, R., Ortego, M. I. and van den Boogart, K. G. (2013). Bayes Spaces: use of improper distributions and exponential families. *RACSAM*, 107, 475–486.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer.

Filzmoser, P., Hron, K. and Templ, M. (2012). Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics*, 27, 585–604.

Genest, C., Hron, K. and Nešlehová, J. (2023). Orthogonal decomposition of multivariate densities in Bayes spaces and relation with their copula-based representation. *Journal of Multivariate Analysis*, 198, 105228.

Hartigan, J. A. (1975). *Clustering Algoritms*. New York: Wiley.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. New York: Springer.

Hron, K., Machalová, J. and Menafoglio, A. (2022). Bivariate densities in Bayes spaces: orthogonal decomposition and spline representation. *Statistical Papers*, 64, 1629–1667.

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K. and Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis*, 94, 330–350.

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8, 231–255.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 533–550.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Upper Saddle River: Pearson.

Kwong, K. S. and Iglewicz, B. (1996). On singular multivariate normal distribution and its applications. *Computational Statistics and Data Analysis*, 22(3), 271–285.

Machalová, J., Hron, K. and Monti, G. S. (2016). Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*, 43, 1419–1435.

Machalová, J., Talská, R., Hron, K. and Gába, A. (2021). Compositional splines for representation of density functions. *Computational Statistics*, 36, 1031–1064.

Nourollah Mousavi, S. and Sørensen, H. (2017). Functional logistic regression: a comparison of three methods. *Journal of Statistical Computation and Simulation*, 88(2), 250–268.

Pavlů, I., Filzmoser, P., Menafoglio, A. and Hron, K. (2022). Classification of continuous distributional data using the logratio approach. In P. Brito and S. Dias (Eds.), *Analysis of Distributional Data*, pp. 183–202. Portland: Chapman and Hall/CRC.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications*. Chichester: Wiley.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15, 384–398.

Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Chichester: Wiley.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Reiss, P. T. and Ogden, R. T. (2007). Functional Principal Component Regression and Functional Partial Least Squares. *Journal of the American Statistical Association*, 102, 984–996.

Šimíček, D., Bábek, O., Hron, K., Pavlů, I. and Kapusta, J. (2021). Separating provenance and palaeoclimatic signals from geochemistry of loess-paleosol sequences using advance statistical tools: Central european loess belt. *Sedimentary Geology*, 419.

Talská, R., Hron, K. and Matys Grygar, T. (2021). Compositional scalar-on-function regression with application to sediment particle size distributions. *Mathematical Geosciences*, 53, 1667–1695.

Talská, R., Menafoglio, A., Hron, K., Egozcue, J. J. and Palarea-Albaladejo, J. (2020). Weighting the domain of probability densities in functional data analysis. *Stat*, 9. e283. 10.1002/sta4.283.

Templ, M., Hron, K. and Filzmoser, P. (2011). robCompositions: An R-package for robust statistical analysis of compositional data. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 341–355. Chichester: Wiley.

van den Boogaart, K. G., Egozcue, J. J. and Pawlowsky-Glahn, V. (2010). Bayes linear spaces. *SORT- Statistics and Operations Research Transactions*, 34(2), 201–222.

van den Boogaart, K. G., Egozcue, J. J. and Pawlowsky-Glahn, V. (2014). Hilbert Bayes spaces. *Australian & New Zealand Journal of Statistics*, 54(2), 171–194.

Varmuza, K. and Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton: CRC Press.