MODELLISTICA E CALCOLO SCIENTIFICO
MODELING AND SCIENTIFIC COMPUTING

MOX-Report No. 05/2017

Logratio approach to distributional modeling

Menafoglio, A.; Hron, K.; Filzmoser, P.

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

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

Logratio approach to distributional modeling

Alessandra Menafoglio[1], Karel Hron[2], Peter Filzmoser[3]

[1]MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy
[2]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Olomouc, Czech Republic.
[3]Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria
alessandra.menafoglio@polimi.it; hronk@seznam.cz; p.filzmoser@tuwien.ac.at

Abstract

Symbolic data analysis (SDA) as introduced in [3, 21] provides a unified approach to analyze distributional data, resulting from capturing intrinsic variability of groups of individuals as input observations. In parallel to the SDA approach, a concise methodology has been developed since the early 1980s to deal with *compositional data* — i.e., data carrying only relative information [1, 22] — through the logratios of their parts. Most methods in compositional data analysis aims to treat multivariate observations which can be identified with probability functions of discrete distributions. Nevertheless, a methodology to capture the specific features of continuous distributions (densities) has been recently introduced [8, 27]. The aim of this work is to describe a general setting that includes both the discrete and the continuous setting, and to provide specific details to both frameworks focusing on the implications on SDA. The theoretical developments are illustrated with real-world case studies.

Keywords: compositional data; Bayes spaces; centred log-ratio transformation; multivariate functional principal component analysis

1 Introduction

There are several types of variables in symbolic data analysis that naturally induce a probability distribution. For the discrete case, the prominent case is formed by categorical modal variables. A categorical modal variable Y with a finite domain $O = \{m_1, \ldots, m_D\}$ is a multi-state variable such that, for each element of Y, a category set is given and, for each category m_l, a weight (e.g., frequency or probability) is provided. If the weight is a frequency, it represents the proportion of individuals of the underlying microdata set characterized by this category. Consequently, from the probabilistic point of view, we would get a probability function over a set of categories. Although the unit sum of weights is taken as a usual representation, it is rather a convention than a real need. For example, the weights could also contain concentrations of chemical elements, or household expenditures in local currency. Alternatively, probability functions can also be obtained if one considers histogram-valued variables with either absolute or relative frequencies, and assumes that the classes of the histograms are fixed for each variable [18]. In this case, the "observations" (the histograms) are again of the same data type, each class of the histograms being a part (category).

In all the above cases, the main point is that weights (frequencies) contain quantitatively expressed relative contributions on a whole. The concrete representation of weights (probabilities, concentrations, ppm and so on) can be chosen arbitrarily without any loss of information. This idea could also be adapted to the continuous case, where the domain of symbolic variables is characterized by a subset of the real line, typically a bounded or unbounded interval. Then the probability function is replaced by a density, a non-negative Borel measurable function with unit integral constraint. An example can be age/income distribution in a certain region, i.e., the finite domain is replaced by an infinite one. And again, even if we used a representation of density that would lead to another integral value, the main feature – i.e., that density conveys relative contributions of Borel sets (subsets of the domain) to the overall probability (weight, frequency) – remains unaltered. In other words, both compositional data and density functions as distributional variables share the property of *scale invariance*. Additionally also their *relative scale* should be taken into account. For example, for the case of densities, the relative increase of a probability over a Borel set from 0.05 to 0.1 (2 multiple) differs from the increase 0.5 to 0.55 (1.1 multiple), although the absolute differences are the same in both cases. Accordingly, not only scale invariance, but also relative scale of distributional variables should be reflected by their statistical processing. The features of both discrete and continuous distributional variables are captured in the geometry of the Bayes space [27], that results in the Aitchison geometry on the simplex [10] when considering the special case of discrete distributions (expressed through compositional data).

Even though compositional data analysis belongs to multivariate statistics and the statistical processing of densities to functional data analysis (FDA, [23]), they both represent just univariate cases from the perspective of symbolic data analysis. Therefore, the main challenge in this setting is to extend the existing compositional methodology to handle more than one symbolic variable (discrete or continuous), simultaneously. The aim of this chapter is to take a step forward in this direction. Therefore, the next section is devoted to describe Bayes spaces and, as a special case, the Aitchison geometry for compositional data, that form the milestones to introduce the statistical analysis of compositions and density functions through the logratio approach. Concrete aspects of their modeling, with extension to multivariate symbolic variables, are discussed in Section 3. Two real-world data sets, representing discrete and continuous distributions, are employed in Section 4 to illustrate the methodological developments. Finally, Section 5 concludes.

2 The Bayes space embedding for compositional vectors

2.1 An introduction to Bayes spaces

The distribution of a random variable is characterized by a σ-finite positive measure μ on a measurable space $(\Omega, \mathcal{A})$. Although in practice exclusively probability measures P are considered for this purpose, the condition of normalization by $\mathsf{P}(\Omega) = 1$ is rather a convention than an actual need. In fact, any probability measure forms just a representation of a family of proportional measures $\mathcal{M} = \{\mu \,|\, \exists c > 0 : \forall \mathsf{A} \in \mathcal{A},\, \mu(\mathsf{A}) = c\,\mathsf{P}(\mathsf{A})\}$, which are equivalent from the viewpoint of the *relative* information they provide. Indeed, a rescaling of the measure leaves the ratios (or logratios) between its "parts" unchanged — i.e., between the measure of the measurable subsets of Ω —, which in turn is the only relevant information embedded into the measure itself. As such, two measures μ, ν are equivalent if they are proportional, denoted hereafter by $\mu =_{B(\lambda)} \nu$, where λ is a reference measure on $(\Omega, \mathcal{A})$. Given a measure μ, if there exists its Radon-Nikodym derivative with respect to λ (i.e., the density $d\mu/d\lambda$), it is identified with the measure μ itself. As long as $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is concerned, the reference measure λ is often set to the Lebesgue measure. However, any (probability) measure P could be considered as well. Given a reference measure P, the equivalence classes of σ-finite measures can be equipped with the geometrical structure of a Bayes space as in [26], whose origin is precisely the reference measure P. More specifically, a Bayes space is a space of ($B(\mathsf{P})$-equivalence classes of σ-finite) measures on $(\Omega, \mathcal{A})$ endowed with a vectorial structure induced by the perturbation and powering operations $(\oplus, \odot)$, which are defined as

$$
\begin{aligned}
(\mu \oplus \nu)(\mathsf{A}) &= {}_{B(\mathsf{P})} \int_{\mathsf{A}} \frac{d\mu}{d\mathsf{P}}(x) \frac{d\nu}{d\mathsf{P}}(x)\, d\mathsf{P}(x); && (1)\\
(\alpha \odot \mu)(\mathsf{A}) &= {}_{B(\mathsf{P})} \int_{\mathsf{A}} \left(\frac{d\mu}{d\mathsf{P}}(x)\right)^{\alpha} d\mathsf{P}(x), && (2)
\end{aligned}
$$

with μ, ν elements of the space and α a real number. Both perturbation and powering can also be expressed in terms of densities; for $f = d\mu/d\mathsf{P}$ and $g = d\nu/d\mathsf{P}$ we get

$$
(f \oplus g)(x) =_{B(\mathsf{P})} f(x)g(x), \quad (\alpha \odot f)(x) =_{B(\mathsf{P})} f(x)^{\alpha}. \tag{3}
$$

The results of both operations are densities again, possibly rescaled to unit integral constraint using the closure operation $\mathcal{C}(f) = \frac{f}{\int f\, d\mathsf{P}}$. Subtraction (or perturbation-subtraction) of densities is then defined as $f \ominus g = f \oplus (-1 \odot g) =_{B(\mathsf{P})} f/g$. This operation can be used, e.g., to change the reference measure to P_1 by employing the well-known chain rule, $(d\mu/d\mathsf{P}_1)(d\mathsf{P}_1/d\mathsf{P}) = d\mu/d\mathsf{P}$.

Given a reference measure P, we call $B^2(\mathsf{P})$ the Bayes space whose elements are ($B(\mathsf{P})$-equivalence classes of σ-finite) measures μ such that

$$
\int \left| \ln \frac{d\mu}{d\mathsf{P}} \right|^2 d\mathsf{P} < +\infty.
$$

Here, measures are identified with the corresponding Radon-Nikodym densities. In $B^2(\mathsf{P})$ an inner product can be defined as [8, 7]

$$\langle f, g\rangle_{B^2(\mathsf{P})} = \frac{1}{2\mathsf{P}(\Omega)} \int\int \ln\frac{f(x)}{f(y)} \ln\frac{g(x)}{g(y)}\, d\mathsf{P}(x)\, d\mathsf{P}(y), \tag{4}$$

for f, g densities in $B^2(\mathsf{P})$. The induced notions of norm and distance are then

$$||f||_{B^2(\mathsf{P})} = \frac{1}{2\mathsf{P}(\Omega)} \int\int \ln^2\frac{f(x)}{f(y)}\, d\mathsf{P}(x)\, d\mathsf{P}(y)$$

and

$$d_{B^2(\mathsf{P})}(f, g) = \frac{1}{2\mathsf{P}(\Omega)} \int\int \left(\ln\frac{f(x)}{f(y)} - \ln\frac{g(x)}{g(y)}\right)^2 d\mathsf{P}(x)\, d\mathsf{P}(y),$$

respectively. The space $B^2(\mathsf{P})$ equipped with the operations of perturbation and powering $(\oplus, \odot)$, and the inner product $\langle\cdot,\cdot\rangle$ is a separable Hilbert space [27].

The reference measure P may be chosen according to convenience. Although several options are discussed in [27], two cases are thoroughly considered in the literature: (a) the continuous uniform measure P_c (i.e., the Lebesgue measure) [27, 5, 19, 20], and (b) the discrete uniform measure P_d (i.e., the counting measure), which leads to the Aitchison geometry [1, 10]. The continuous uniform measure, defined on the interval $I = [a, b](\equiv \Omega)$ through its density as

$$\frac{d\mathsf{P}_c}{d\lambda}(x) = 1,$$

can be considered as a reference for functional distributional variables (i.e., continuous densities) with bounded domain. Nevertheless, P_c has often been considered as reference even for variables with (theoretically) unbounded domain, e.g., by neglecting subdomains with very rare occurrence. In all these cases, the inner product simplifies to

$$\langle f, g\rangle_{B^2(\mathsf{P}_c)} = \frac{1}{2\eta} \int_a^b \int_a^b \ln\frac{f(x)}{g(x)} \ln\frac{f(y)}{g(y)}\, dx\, dy,$$

with $\eta = b - a$, and by virtue of the Weierstrass theorem, continuous densities belong to $B^2(\mathsf{P}_c)$.

In case of multivariate compositional data, the discrete uniform measure is usually employed as a reference measure on $\Omega = \{m_1, \ldots, m_D\}$, i.e.,

$$d\mathsf{P}_d(x) = 1,\ x \in \Omega,$$

thus obtaining the Aitchison geometry. Here, compositions with D parts, $\mathbf{x} = (x_1, \ldots, x_D)'$, are identified with discrete probability functions over Ω (thus referring to a categorical modal variable). Having set the unit sum representation of compositions, the sample space of compositional data becomes the $(D-1)$-dimensional simplex

$$\mathcal{S}^D = \left\{\mathbf{x} = (x_1, \ldots, x_D)', x_i > 0, \sum_{i=1}^{D} x_i = 1\right\}.$$

In this setting, the closure operation reads $\mathcal{C}(\mathbf{x}) = \mathbf{x}/\sum_{i=1}^{D} x_i$; as before, both $\mathbf{x}$ and $\mathcal{C}(\mathbf{x})$ belong to the same equivalence class. For two compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and a real α, the operations of perturbation and powering read

$$\mathbf{x} \oplus \mathbf{y} =_{B^2(\mathsf{P}_d)} \mathcal{C}(x_1 y_1, \ldots, x_D y_D)', \quad \alpha \odot \mathbf{x} =_{B^2(\mathsf{P}_d)} \mathcal{C}(x_1^{\alpha}, \ldots, x_D^{\alpha}),$$

respectively, and the Aitchison inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

This geometry is the basis of compositional data analysis of multivariate compositional vectors, based on the logratio approach. In the next subsection, we illustrate a practical strategy to employ the Bayes space geometry, either continuous or discrete, for the statistical analysis of compositions.

2.2 Statistical analysis in Bayes spaces

A statistical analysis of continuous or discrete density functions needs to properly account for both the data dimensionality and the geometrical structure governing Bayes spaces. In fact, although continuous densities are functional data and discrete compositions are multivariate observations, they both are featured by the basic properties of compositions (as scale invariance and relative scale), that are captured neither by FDA nor by classical multivariate methods. For instance, most methods of FDA rely on the assumption that the data belong to the space $L^2(\mathsf{P})$ of squared-integrable functions with respect to a reference measure P (usually set to the Lebesgue measure). However, the geometrical structure of the space $L^2(\mathsf{P})$ is not appropriate for compositions (e.g., the point-wise sum of compositions does not result in a composition). Similarly, most multivariate statistical methods are built in the Euclidean setting, which is not appropriate to analyze discrete compositions. Nevertheless, as long as the data are embedded in a separable Hilbert space, one can map the observations in $L^2(\mathsf{P})$ or in the Euclidean space $\mathbb{R}^D$, and accordingly perform the statistical analysis via FDA or multivariate statistics, while accounting for the Bayes space geometry.

Let us first focus on the continuous case, having set the reference measure to a measure P (not necessarily a probability measure). As separable Hilbert spaces, an isometric isomorphism exists between $B^2(\mathsf{P})$ and $L^2(\mathsf{P})$. An instance of such an isometry is provided by the centered logratio (clr) transformation, defined for a density $f = d\mu/d\mathsf{P}$ as

$$\mathrm{clr}(f) = \ln f - \frac{1}{\mathsf{P}(\Omega)} \int \ln f \, d\mathsf{P}. \tag{5}$$

Consequently, for $\alpha \in \mathbb{R}$, $f, g \in B^2(\mathsf{P})$ the following relations hold,

$$\mathrm{clr}(f \oplus g) = \mathrm{clr}(f) + \mathrm{clr}(g), \ \mathrm{clr}(\alpha \odot f) = \alpha \cdot \mathrm{clr}(f),$$

$$\langle f, g \rangle_{B^2(\mathsf{P})} = \langle \mathrm{clr}(f), \mathrm{clr}(g) \rangle_{L^2(\mathsf{P})}.$$

We note that these relations enable one to handle clr-transformed densities in the L^2 setting. Due to its construction, clr transformations fulfill the integral constraint $\int \operatorname{clr}(f)\, d\mathsf{P} = 0$ that should be taken into account in any statistical analysis based on clr-transformed data. Moreover, in case of a uniform reference measure $\mathsf{P} \equiv \mathsf{P}_c$, (5) reads

$$\operatorname{clr}(f)(t) = \ln f(t) - \frac{1}{\eta} \int_a^b \ln f(\tau)\, d\tau. \tag{6}$$

Note that it would also be possible to get rid of the zero integral constraint resulting from the clr transformation, e.g., by expressing the densities via the Fourier coefficients of a basis in $B^2(\mathsf{P}_c)$ (such as Legendre polynomials [25]); though, most recent literature works propose clr-based methods [19, 15].

The situation is a bit different for compositional data (the case of P_d), where the clr transformation of a composition $\mathbf{x}$ (in fact, coordinates with respect to a generating system on the simplex) results in

$$\operatorname{clr}(\mathbf{x}) = (y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^D x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^D x_i}} \right)'. \tag{7}$$

Sequential binary partitioning (SBP) [9] provides a range of possibilities to build interpretable coordinates. Indeed, SBPs enables one to construct $D-1$ coordinates with respect to an orthonormal basis of the simplex, on the basis of balances between groups of compositional parts, expressed through their geometric means. The use of SBP usually requires some prior knowledge about the problem at hand. However, an "automated" versions of orthonormal coordinates can be considered as well [12]. For instance, for a composition $\mathbf{x}$ one can obtain the $(D-1)$-dimensional real vector $\mathbf{z} = (z_1, \ldots, z_{D-1})'$, as [14]

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^D x_j}}, \; i = 1, \ldots, D-1. \tag{8}$$

Note that only the first coordinate contains the part x_1 in terms of its logratio to the remaining parts at hand, thus it conveys information about the dominance of x_1 "in average". The remaining coordinates $(z_2, \ldots, z_{D-1})$ then represent the subcomposition including the parts $x_2, \ldots, x_D$. We notice that if the l-th part is of interest, one can consider a permutation of the parts in the input composition such that x_l, $l = 1, \ldots, D$, takes the first position, the others being placed arbitrarily (different orthonormal coordinate systems are just rotations of each other [10]). In this case, the first element of the corresponding coordinates, denoted by $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$ would have the above interpretation.

An explicit relation exists between the clr transformation and the coordinates $\mathbf{z}$, as $y_l = \sqrt{\frac{D-1}{D}} z_1^{(l)}$. This relation can be used to support the interpretation of clr variables.

Once compositional data are expressed either via clr or in orthonormal coordinates, all the standard methods of multivariate statistics that rely on the Euclidean geometry [6] can be employed. We discuss this in more detail in the next section from the perspective of SDA.

3 Implications for Symbolic Data Analysis

In both the discrete and the continuous setting, most cases of logratio modeling using the Bayes space methodology represent just univariate cases from the perspective of symbolic data analysis. Indeed, SDA methods are usually employed to cope with more distributional variables simultaneously. This raises an urgent need to provide set of coordinates for compositional data and densities, that would even enable one to perform joint analyses of discrete and continuous distributional data, through multivariate statistical analyses.

In symbolic data analysis, this is traditionally achieved by quantile representation of distributional data [16]. The principle is to express the observed variable values by some predefined quantiles of the underlying distribution. For example, for categorical multi-valued variables, quantiles may be determined from the ranking of the categories based on their frequencies, or other designed methods. In the simplest case, when quartiles are chosen, the representation for each variable is defined by the 5-uple (Min, Q_1, Q_2, Q_3, Max), that forms a kind of coordinates of the variable. Although one can honor scale invariance by following this approach, their relative scale is not taken into account.

The issue of coordinates in case of compositional data can be addressed either through the centered logratio (7), or by orthonormal coordinates (8). Although for some methods (e.g., principal component analysis and the associated compositional biplot [2]) the clr coordinates are preferable, while in other cases both options are allowed (e.g., cluster analysis, or regression analysis with compositional response [4]), whenever possible the orthonormal coordinates are employed. The reason for this relies in the fact that they guarantee a regular covariance matrix of the observations, which is a must for most robust multivariate methods [11]. In the continuous case, a set of coordinates can be obtained by using the Fourier coefficients of a basis in $B^2(\mathsf{P}_c)$ [8], or a B-spline representation of clr transformed densities [17]. Note that, in the latter case, one should take care of the fact that B-spline bases are not orthonormal. Notice that, except in very particular cases [27], density functions need infinitely many coefficients to be described. Thus, in general, an appropriate dimensionality reduction needs to be performed prior to their statistical analysis.

Amongst the compositional methods which are suitable to be extended to SDA problems, we focus here on two special cases that illustrate the potential of the logratio approach to analyze distributional data. In the next section, linear regression with a real response and several compositional covariates is presented, followed by multivariate principal component analysis for density functions.

3.1 Linear regression with discrete distributions as covariates

In [28] a regression model is presented, where both the response and the explanatory variables are compositional data. Although the model was not originally intended to provide a link with symbolic data analysis, it is particularly well-suited for our purposes. In the following, we employ a simplified version of this model, based on p compositions

$\mathbf{x}_1,\dots,\mathbf{x}_p$, containing $D_1,\dots,D_p$ parts ($D := D_1+\dots+D_p$), that explain a real response variable Y. Note that this setting represents a generalization of the so called *experiments with mixtures* [24], that has been adapted to the logratio methodology in [14].

Instead of analyzing the original compositional data, we express these in orthonormal coordinates, $\mathbf{z}_1,\dots,\mathbf{z}_p$, where $\mathbf{z}_j = (z_{j,1},\dots,z_{j,D_j-1})'$, $j=1,\dots,p$, and consider the regression model

$$\mathsf{E}(Y|(\mathbf{z}_1,\dots,\mathbf{z}_p)) = \beta_0+z_{1,1}\beta_{1,1}+\dots+z_{1,D_1-1}\beta_{1,D_1-1}+\dots+z_{p,D_p-1}\beta_{p,D_p-1}. \quad (9)$$

The linear model for the observations is

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta}+\boldsymbol{\varepsilon}, \quad (10)$$

where the $n\times(D-p+1)$ design matrix $\mathbf{Z}$ is defined as

$$\mathbf{Z} = \begin{pmatrix} 1 & \mathbf{z}'_{1,1} & \dots & \mathbf{z}'_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & \mathbf{z}'_{n,1} & \dots & \mathbf{z}'_{n,p} \end{pmatrix}.$$

The model thus contains $D-p+1$ regression parameters. Under the usual assumptions, the parameters can be estimated by a least squares (LS) method, i.e., by minimizing the sum of squared residuals RSS. This yields to the estimates $\widehat{\beta}_0,\widehat{\beta}_{1,1},\dots,\widehat{\beta}_{p,D_p-1}$. The result can be then used for prediction purposes, or for further statistical inference.

Under the Gaussian assumption, a series of tests can be performed. For instance, one may want to evaluate whether the j-th composition, $j=1,\dots,p$, has a significant influence on the explanatory variable Y. For this purpose, the following test statistic can be employed,

$$Q_j = \frac{1}{(D_j-1)S^2}\widehat{\boldsymbol{\beta}}'_j\mathbf{W}_j^{-1}\widehat{\boldsymbol{\beta}}_j,\ j=1,\dots,p, \quad (11)$$

where $S^2 = RSS/(D-p+1)$, $\widehat{\boldsymbol{\beta}}_j = (\widehat{\beta}_{j,1},\dots,\widehat{\beta}_{j,D_j-1})'$ and the $(D_j-1)\times(D_j-1)$ matrix $\mathbf{W}_j$ is formed by the block of $(\mathbf{Z}'\mathbf{Z})^{-1}$ that corresponds to $\boldsymbol{\beta}_j$ as part of $\boldsymbol{\beta} = (\beta_0,\boldsymbol{\beta}'_1,\dots,\boldsymbol{\beta}'_p)'$. Under the null hypothesis, the statistic Q_j follows a Fisher distribution with D_j-1 and $n-D+p-1$ degrees of freedom.

If for the j-th composition (distributional variable) the above hypothesis is rejected, one may want to investigate which of its part(s) does have significant influence on Y. A solution can be provided again in terms of orthonormal coordinates. Indeed, one may take advantage of the interpretation of (8), leading to coordinates $\mathbf{z}_j^{(l_j)} = (z_{j,1}^{(l_j)},\dots,z_{j,D_j-1}^{(l_j)})'$ and the corresponding parameters $\boldsymbol{\beta}_j^{(l_j)} = (\beta_{j,1}^{(l_j)},\dots,\beta_{j,D_j-1}^{(l_j)})'$. Here, only the first coordinate of $\mathbf{z}_j^{(l_j)}$ and the corresponding regression parameter are of primary interest. Concretely, if the significance of the regression parameter $\beta_{j,1}^{(l_j)}$ is confirmed by the rejection of the corresponding hypothesis on a significance level α, then the relative information concerning the l_j-th part of the composition $\mathbf{x}_j$ (resulting from summarizing logratios to the other parts of $\mathbf{x}_j$) has an influence on the response. The decision can be taken based on the test statistic

$$T_{jl_j} = \frac{\widehat{\beta}_{j,1}^{(l_j)}}{\sqrt{S^2\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(l_j,l_j)}}},\ j = 1, \ldots, p,\ l_j = 1, \ldots, D_j, \tag{12}$$

where $\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(l_j,l_j)}$ denotes the diagonal element of the matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$ which corresponds to the coefficient $\widehat{\beta}_{j,1}^{(l_j)}$. Under the null hypothesis, T_{jl_j} follows a Student's t distribution with $n - D + p - 1$ degrees of freedom. Note that for an exhaustive search for significance of coordinates in single explanatory parts, pD regression models would need to be built. Nevertheless, the estimate of the intercept parameter as well as the coefficient of determination for the regression model (9) are always the same [14], due to the coordinates orthonormality.

3.2 Multivariate functional principal component analysis when data are density functions

In this subsection we shall focus on multivariate continuous densities, named multivariate functional compositions (mFCs). These are defined as K-dimensional vectors whose components are elements of the Bayes space $B^2(\mathsf{P})$, for a continuous reference measure P. In this subsection, we will always consider as reference measure the uniform P_c, and denote $B^2(P_c)$ by B^2 for the sake of simplicity. For instance, Figure 1 represents a dataset of population pyramids in 57 districts of Upper Austria, that are mFCs of dimension $K = 2$: they are coupled density functions, describing the age density of males and females in these regions. We aim to introduce a methodology allowing to explore the variability of a dataset of mFCs, and consistently perform dimensionality reduction.

In multivariate and functional statistics (functional) principal component analysis (PCA) is widely employed to attain these types of goals. In (functional) PCA the focus is posed on the main modes of variability of the sample, whose interpretation is often insightful in terms of the observed phenomenon. In the recent literature, [15] introduces the simplicial functional principal component analysis (SFPCA) as an extension of functional PCA to the Bayes space setting. Here, we consider an approach similar to that introduced in [15] to derive an extension of simplicial principal component analysis to the multivariate, simplicial and functional setting, that relies on the Bayes space geometry introduced in Section 2.

We first note that mFCs are not multivariate density functions: only the marginal densities are available, up to a scale factor. Instead, a mFC can be considered as an element of the space $[B^2]^K = B^2 \times ... \times B^2$, which is a separable Hilbert space if equipped with the component-wise B^2 operations:

$$(\boldsymbol{f} \oplus \boldsymbol{g})_i = f_i \oplus g_i,\ (\alpha \odot \boldsymbol{f})_i = \alpha \odot f_i,\ \boldsymbol{f} = (f_i), \boldsymbol{g} = (g_i) \in [B^2]^K, \alpha \in \mathbb{R},$$

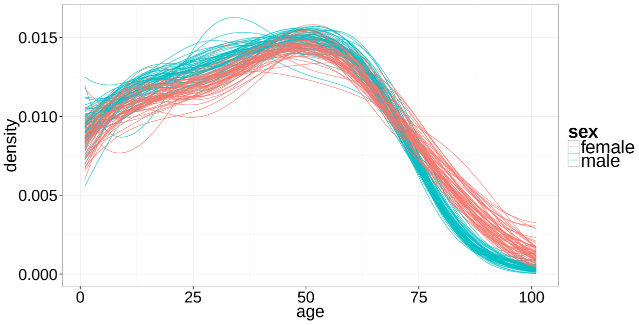

Figure 1: Population pyramids in the 57 districts of Upper Austria.

and the inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{[B^2]^K} = \sum_{i=1}^K \langle f_i, g_i \rangle_{B^2}$, for $\boldsymbol{f} = (f_i), \boldsymbol{g} = (g_i) \in [B^2]^K$.

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be a dataset of mFCs, e.g., that displayed in Figure 1. To simplify the notation and without loss of generality, hereafter we assume the dataset to be centered. Note that, one can always consider the centered version of a given dataset, that is $\widetilde{\boldsymbol{X}}_1, ..., \widetilde{\boldsymbol{X}}_N$, with $\widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i \ominus \overline{\boldsymbol{X}}$ and $\overline{\boldsymbol{X}} = \frac{1}{N} \odot \bigoplus_{i=1}^N \boldsymbol{X}_i$. Centering the observation can be interpreted as setting the reference measure to the sample mean $\overline{\boldsymbol{X}}$.

Multivariate SFPCA (mSFPCA) of $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ consists of finding the main modes of variability of the dataset. These are the orthogonal directions in $[B^2]^K$ that display the maximum variability of the dataset. They are identified by a collection of orthogonal elements $\{\boldsymbol{\zeta}_j\}_{j \geq 1}$, $\boldsymbol{\zeta}_j \in [B^2]^K$, of unitary norm, that are found by maximizing the following objective functional

$$\frac{1}{N} \sum_{i=1}^N \langle \boldsymbol{X}_i, \boldsymbol{\zeta} \rangle^2_{[B^2]^K} \text{ subject to } \|\boldsymbol{\zeta}\|_{[B^2]^K} = 1;\ \langle \boldsymbol{\zeta}, \boldsymbol{\zeta}_k \rangle_{A^2} = 0,\ k < j, \tag{13}$$

where $\langle \boldsymbol{X}_i, \boldsymbol{\zeta} \rangle^2_{[B^2]^K}$ represents the projection of $\boldsymbol{X}_i$ along the direction identified by $\boldsymbol{\zeta}$, and the orthogonality condition $\langle \boldsymbol{\zeta}, \boldsymbol{\zeta}_k \rangle_{[B^2]^K} = 0$, for $k < j$, is meaningful only for $j \geq 2$.

It can be shown that, for each $j = 1, 2, ...,$ maximization of (13) leads to a unique solution in $[B^2]^K$ ([13], Theorem 3.2). Indeed, the principal components are uniquely found as the eigenfunctions of the sample covariance operator $V : [B^2]^K \to [B^2]^K$, that acts on $\boldsymbol{x} \in [B^2]^K$ as

$$V\boldsymbol{x} = \frac{1}{N} \odot \bigoplus_{i=1}^N \langle \boldsymbol{X}_i, \boldsymbol{x} \rangle_{[B^2]^K} \odot \boldsymbol{X}_i.$$

The $N - 1$ non-zero eigenvalues of the operator V, $\lambda_1 < \lambda_2 < ... < \lambda_{N-1}$, represent the variability of the dataset along its main modes of variability $\boldsymbol{\zeta}_1, ..., \boldsymbol{\zeta}_{N-1}$.

For actual computation of the eigenpairs $(\lambda_j, \boldsymbol{\zeta}_j)$, $j = 1, ..., N-1$, we propose to employ the clr-transformation (5), in order to map the problem in $L^2(\mathsf{P})$ (L^2 for short) and proceed as in the multivariate functional case. Specifically, we propose the following procedure

(1) **Transform:** For $i \in 1, ..., N$, transform the i-th observed mFC as $\boldsymbol{clr}(\boldsymbol{X}_i)$, where the mapping $\boldsymbol{clr}$ acts as a component-wise clr transformation: $\boldsymbol{clr}(\boldsymbol{f}) = (clr(f_l)) \in [L^2]^K$, for $\boldsymbol{f} = (f_l) \in [B^2]^K$;

(2) **Solve in** $[L^2]^K$: Compute the multivariate FPCs $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_{N-1}$ in $[L^2]^K$ and the corresponding eigenvalues $\lambda_1, ..., \lambda_{N-1}$;

(3) **Back-transform**: Employ the inverse $\boldsymbol{clr}$-transformation to $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_{N-1}$, i.e., apply component-wise the inverse of the clr-transformation, and set $\boldsymbol{\zeta}_j = \boldsymbol{clr}^{-1}(\boldsymbol{\xi}_j)$.

It is possible to prove that (i) the eigenvalues found at step (1) are the same as those of the operator V, and (ii) this procedure leads to a correct characterization of the set of eigenpairs of V, since the clr-transformation is an isometric isomorphism between B^2 and L^2. The proof of these points can be obtained by generalizing the arguments presented in [15] (not shown).

To reduce the dimensionality of the dataset, we can then follow the same lines of the classical setting. For instance, we can employ the scree plot to determine the relevant mSFPCs in terms of the proportion of explained variability. The interpretation of the mSFPCs can be based on graphical displays, such as the plot of the eigenfunctions (possibly transformed via clr), or the perturbation of the mean via the eigenfunction perturbed by a coefficient. The former allows to single out contrasts between parts of the domains to which different weights are attributed; the latter enables one to visualize the portion of variability around the mean which is captured by the corresponding principal component.

4 Case studies

4.1 Effect of GDP components and causes of death on life expectancy

Eurostat provides various data sets at `http://ec.europa.eu/eurostat/data` that refer to economy, population, health, education, etc., of the EU countries. For the purpose of illustrating the procedure outlined in Section 3.1, we consider the life expectancy as response variable, and two compositions as explanatory variables. The first composition includes the most important components of the GDP (Gross Domestic Product), namely the *private final consumption expenditure* (private), the *government final consumption expenditure* (governmt), the *gross fixed capital formation* (capital), the *exports*, and the *imports*. All these data are taken from the year 2011, for the EU countries, as well as for Norway and Switzerland, and we use the data reported in absolute values (million Euros). The second composition contains the most relevant causes of death. Again, we use data from 2011, for the same countries as before, and take the absolute numbers as a basis. The following groups are considered (the abbreviations in brackets refer to the ICD codes, and to the abbreviations we are using later on): *Certain infectious and parasitic diseases* (A00-B99) (infect), *Malignant neoplasms* (C00-C97) (neoplasm), *Endocrine nutritional and metabolic diseases* (E00-E90) (nutrition), *Mental and behavioral disorders* (F00-F99) (mental), *Diseases of the nervous system and*

Table 1: Absolute numbers of causes of death for the male population of some selected countries.

	nervous	infect	neoplsm	nutrition	mental	circul	respirat	digest
BE	1971	1104	15381	1128	1477	14228	5782	2287
BG	547	334	9807	720	41	35453	2483	2205
CZ	996	647	15051	1204	413	24303	3282	2583
DK	807	385	8118	900	1324	6540	2727	1275
DE	10765	7565	119818	12396	11003	145647	32050	20562

Table 2: Coordinates for the causes of death (male) for variable *nervous* at the first position, for some selected countries.

	z_1	z_2	z_3	z_4	z_5	z_6	z_7
BE	0.51	1.22	-1.45	1.15	1.17	-1.11	-0.66
BG	0.96	1.64	-1.77	0.76	4.29	-2.22	-0.08
CZ	0.93	1.54	-1.63	0.83	2.30	-1.73	-0.17
DK	0.79	1.72	-1.31	0.86	0.66	-1.02	-0.54
DE	0.88	1.39	-1.38	0.85	1.23	-1.42	-0.31

the sense organs (G00-H95) (nervous), *Diseases of the circulatory system* (I00-I99) (circulatory), *Diseases of the respiratory system* (J00-J99) (respiratory), and *Diseases of the digestive system* (K00-K93) (digestive).

The life expectancy as well as the causes of death are available for the total population, and for males and females separately. Therefore, in the analyses below we investigate models for these three cases separately. The GDP composition is of course unchanged.

In order to get an impression about the raw data, Table 1 shows for the males and for some selected countries the absolute numbers of the considered causes of death. It is clear that these raw values would not be meaningful for a direct analysis, since the population sizes in the countries are very different.

A first impression about the data structure is provided in Figure 2. We compare the relative dominance of death by diseases of the nervous system and the sense organs (nervous) with the life expectancy, separately for males (left) and females (right). Thus, in the second composition, the variable *nervous* is put to the first position, and the first coordinate after applying Equation (8) represents all relative information about *nervous*.

Table 2 shows for the data presented in Table 1 the resulting coordinates. The first coordinate (z_1) is used on the horizontal axis on the left plot in Figure 2.

According to the figure, high values on this coordinate correspond to dominance of the disease *nervous*, which relates to low life expectancy, and vice versa. The most important diseases covered by *nervous* are Alzheimer and Parkinson.

The regression model (10) is now applied to the problem, and the idea is to identify economic and/or health information that relate to life expectancy. The regression models which are considered here (total, male, female) lead to multiple R^2 values of

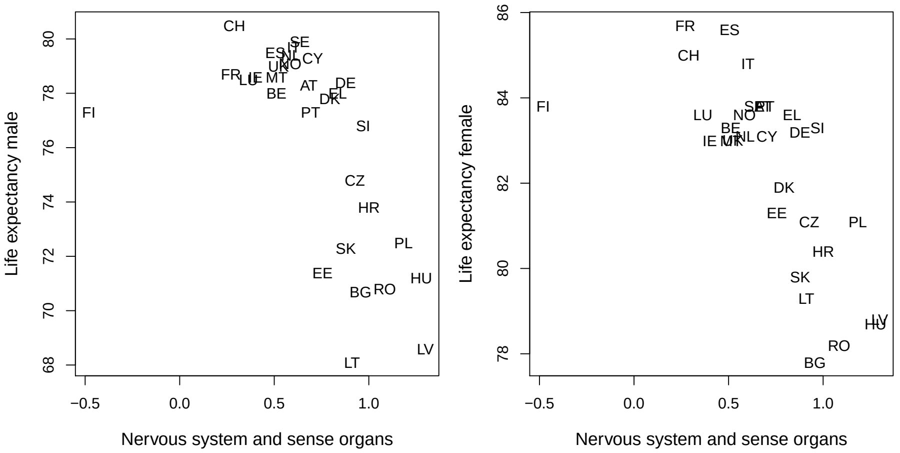

Figure 2: Relation between all relative information to the diseases of the nervous system and the sense organs with the life expectancy; left for males, right for females.

Table 3: Results of the test (11) for both compositions, for models based on the total population, the males, and the females, respectively. Shown are the resulting p-values of the test for the two compositions.

	Total	Males	Females
GDP compositions	<**0.001**	0.22	<**0.001**
Causes of death	<**0.001**	0.13	<**0.001**

more than 0.9. We apply the test statistic (11) to the different settings, and the resulting p-values are reported in Table 3. Both compositions have significant influence for the models based on the total and on the female population, whereas for the males we do not obtain significance.

The second test according to (12) tests for significance of the single parts in the compositions via their corresponding coordinates. The results are presented in Table 4 for the first composition, and in Table 5 for the second composition. We realize that none of the parts in the first composition is significant on its own. In order to get significance, we would need to go for other coordinates from (8) or even to consider a more complex coordinate system [9]. In contrast, several parts from the second composition are contributing significantly. For example, the part *nervous* that was under consideration in Figure 2 has significant contribution in all settings (total, male, female), and the regression coefficient is negative, as it was expected from the plot. So, dominance of this disease (and subsequent death) refers to countries with lower life expectancy. Dominance of *neoplasm* for females also relates to low life expectancy, while dominance of the other significant diseases *circulatory* and *digestive* are in relation to countries with higher life expectancy.

Note that the analysis above would lead to exactly the same results if the absolute values of the compositions (million Euro for GDP composition, numbers of death

Table 4: Results of the test (12) for the first composition, for models based on the total population, the males, and the females, respectively. Shown are the resulting p-values of the test, and the regression coefficients (coeff.) for the parts of the first composition.

	Total		Males		Females	
	p-value	coeff.	p-value	coeff.	p-value	coeff.
private	0.50	-1.32	0.98	0.08	0.58	-0.85
governmt	0.18	2.42	0.47	1.64	0.22	1.66
capital	0.56	-1.19	0.43	-2.18	0.47	-1.19
exports	0.68	1.57	0.92	0.52	0.49	2.00
imports	0.72	-1.49	0.99	-0.06	0.61	-1.61

Table 5: Results of the test (12) for the second composition, for models based on the total population, the males, and the females, respectively. Shown are the resulting p-values of the test, and the regression coefficients (coeff.) for the parts of the second composition.

	Total		Males		Females	
	p-value	coeff.	p-value	coeff.	p-value	coeff.
infect	0.60	0.40	0.10	1.60	0.66	0.24
neoplasm	0.17	-3.71	0.52	-2.09	**0.01**	-5.06
nutrition	0.21	-0.85	0.07	-1.67	0.49	-0.37
mental	0.99	0.003	0.78	0.18	0.29	0.42
nervous	**0.004**	-2.69	**0.005**	-3.70	<**0.001**	-2.76
circulatory	**0.03**	3.40	**0.04**	3.35	<**0.001**	3.77
respiratory	0.42	-1.056	0.17	-2.74	0.39	1.02
digestive	**0.005**	4.49	**0.004**	5.06	**0.01**	2.74

causes) would have been expressed in relative units, like proportions or percentages.

4.2 Dimensionality reduction of population pyramids via mSFPCA in Bayes spaces

We demonstrate the results of the methodology proposed in Subsection 3.2 on the dataset of population pyramids displayed in Figure 1, and presented in [15]. A similar dataset has been considered in [5]. To perform the computations, we resort to numerical integration to deal with clr-transforms and we solve numerically the eigenproblem in $[L^2]^K$ involved in step (2). Another strategy may be employed as well, e.g., by representing the data via a functional basis and expressing the solution through the corresponding coefficient [23, 15]. Figure 3 summarizes the obtained results. Figure 3a displays the variability explained by the first eight SFPCs. It shows a rapid decrease in these variances, which suggests a possible dimensionality reduction to two or three mSFPCs. However, the variability of the estimated scores along the third component (i.e., of $\langle \boldsymbol{X}_i, \boldsymbol{\zeta}_j \rangle^2_{[B^2]^K}$, with $j = 3$, $i = 1, ..., N$) appears affected by the presence of an outlier. Hence, we focus on the first two components for scope of interpretation and dimensionality reduction. To ease the interpretation, Figure 3c-d display the clr transformation of the elements of $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$, i.e., $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ obtained from step (2); colors are used to identify the gender. The transformed eigenfunctions can be interpreted as in FPCA, e.g., looking for meaningful contrasts between portions of the domain. Notice that the clr-transformed eigenfunctions are continuous, non-constant and fulfill the zero-integral constraints which is characteristic for clr-transformed FCs. As such, contrasts are expected in all the ξ_i, $i \geq 1$. Considering the first mSFPCs, we notice that in both elements, a contrast exists between the oldest population (age>80/75 years, for men and women, respectively) and the younger one. We note that this result is consistent with that of [15], that analyzes separately men and women subpopulations. Hence, high scores along the first mSFPC are expected for the municipalities with a higher incidence of the elder population than the mean, and vice versa. This is evident also when observing the plot displayed in Figure 3e. Here, the effect of the variability along the first principal component is visualized via the perturbation of the mean by the first mSFPC powered by $\pm 2 \cdot \sqrt{\lambda_1}$. Having fixed the sign of the eigenfunction $\boldsymbol{\xi}_1$ as in Figure 3c, data with high corresponding scores (dark grey line) tend to have heavier tails than the mean and vice versa.

The interpretation of the second mSFPC in Figure 3d is in terms of contrast between men and women subpopulations, with a positive contribution in men for right tails (age>93 years) higher than the mean, and a negative contribution in women's right tails (age>75 years) higher than the mean. Overall, Figure 3f shows that low scores along the second mSFPC associate with more pronounced peaks in the density functions and vice versa. Figures 3g and h display the contribution to the variability along the two mSFPCs: in each panel, the elements $\langle \boldsymbol{X}_i, \boldsymbol{\zeta}_k \rangle_{[B^2]^K} \odot \boldsymbol{\zeta}_k$, $i = 1, ..., N$, $k = 1, 2$, are represented. In agreement with the previous comments, these plots suggest that most variability is displayed within the right tails. In addition to this, further evidence of the previous interpretation is given by plotting the elements with maximum scores (black curves). Indeed, high scores along the first mSFPC in Figure 3c correspond to higher incidence of the old population than the mean in both men and women; instead, high scores along the second mSFPC correspond to higher incidence of the old population than the mean in men, and lower incidence in women. Similar interpretation — with opposite score signs — are obtained from the elements with minimum scores. In this sense, the second mSFPC provides a contrast between the behavior of men and women subpopulations for the elder ages. Finally, Figure 3j displays the approximation of the densities which are attained via the first two SFPCs, that together explain more than 80% of the overall variability.

5 Conclusions

This contribution has been devoted to the logratio approach to symbolic data analysis of distributional data. In our setting, the relative information embedded in compositions is being analyzed, based on the logratios between the values of the compositional parts (either discrete or continuous). Here, we described a unifying framework for both the continuous and the discrete case, based on the theory of Bayes spaces. We illustrated the discrete case through a regression setting, for a real response modeled in terms of a number of compositions. Here, we considered specific representations of the compositions in terms of coordinates, in order to use the classical tools for inference. To this end, we employed a particular type of the isometric logratio (ilr) coordinates. In the continuous setting, we analyzed multivariate distributional data in the form of densities by extending multivariate functional principal component analysis. Here, the key to bring theory to practice was to employ the centered logratio (clr) transformation to simplify computations of eigenfunctions.

The examples on regression and functional principal component analysis served as illustrations of the great potential of this theory that enables one to deal with both compositional data and densities in the common framework of the Bayes space methodology, adapted to the SDA case. This opens new views even to cope with mixed types of data (e.g., Euclidean, functional, compositional), that remains one of the greatest challenges for the future.

Acknowledgements

Karel Hron gratefully acknowledges the support of the grant No. GA15-06991S of the Grant Agency of the Czech Republic.

References

[1] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, London, 1986.

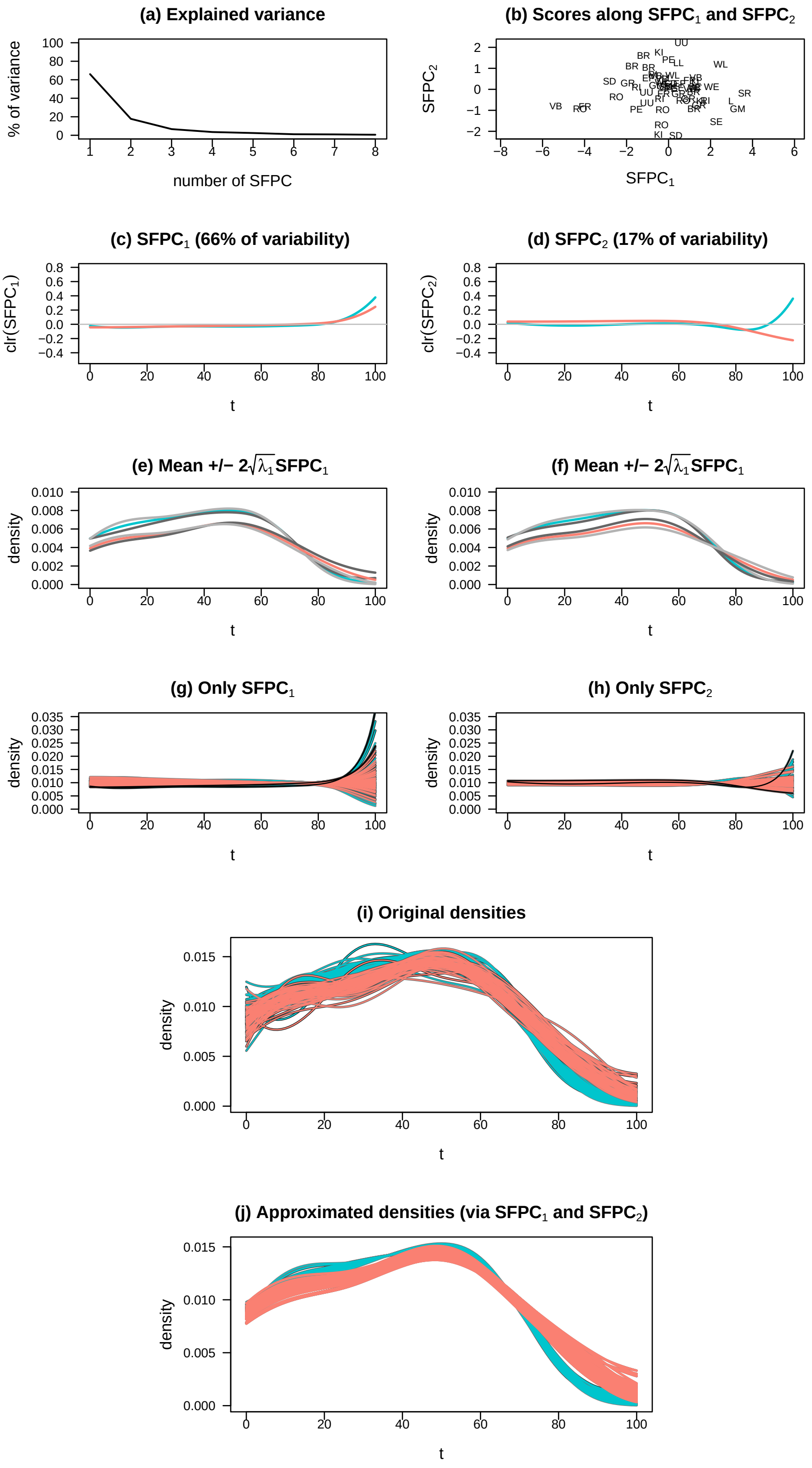

Figure 3: Results of mSFPCA on the population pyramids. In panels (c) to (f): solid dark grey lines indicate the perturbation of the mean by the mSFPC ζ_j powered by $+2 \cdot \sqrt{\lambda_j}$, $j = 1, 2$; solid light grey lines indicate the perturbation of the mean by the mSFPC ζ_j powered by $-2 \cdot \sqrt{\lambda_j}$, $j = 1, 2$.

[2] J. Aitchison and M. Greenacre. Biplots of compositional data. *Applied Statistics*, 51(4):375–392, 2002.

[3] L. Billard and E. Diday. *Symbolic data analysis*. Wiley, Chichester, 2006.

[4] F. Bruno, F. Greco, and M. Ventrucci. Spatio-temporal regression on compositional covariates: modeling vegetation in a gypsum outcrop. *Environmental and Ecological Statistics*, DOI: 10.1007/s10651-014-0305-4, 2015.

[5] P. Delicado. Dimensionality reduction when data are density functions. *Computational Statistics & Data Analysis*, 55(1):401 – 420, 2011.

[6] M.L. Eaton. *Multivariate Statistics. A Vector Space Approach*. John Wiley & Sons, New York, 1983.

[7] J. J. Egozcue and V. Pawlowsky-Glahn. Changing the reference measure in the simplex and its weighting effects. *Austrian Journal of Statistics*, under review, 2016.

[8] J.J. Egozcue, J.L. Díaz-Barrero, and V. Pawlowsky-Glahn. Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica, English Series*, 22(4):1175–1182, 2006.

[9] J.J. Egozcue and V. Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828, 2005.

[10] J.J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

[11] P. Filzmoser and K. Hron. Robust statistical analysis. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis: Theory and Applications*, pages 59–72. Wiley, Chichester, 2011.

[12] E. Fišerová and K. Hron. On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43:455–468, 2011.

[13] L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer, 2012.

[14] K. Hron, P. Filzmoser, and K. Thompson. Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, 39(5):1115–1128, 2012.

[15] K. Hron, A. Menafoglio, M. Templ, K. Hrůzová, and P. Filzmoser. Simplicial principal component analysis for density functions in Bayes spaces. *MOX-report 25/2014*, Politecnico di Milano, 2015.

[16] M. Ichino. Symbolic PCA for histogram-valued data. In *Proceedings IASC 2008*, 2008.

[17] J. Machalová, K. Hron, and J.S. Monti. Preprocessing of centred logratio transformed density functions using smoothing splines. arXiv:1501.07047, 2015.

[18] S. Makosso-Kallyth and E. Diday. Adaptation of interval PCA to symbolic histogram variables. *Advances in Data Analysis and Classification*, 6(2):147 – 159, 2012.

[19] A. Menafoglio, A. Guadagnini, and P. Secchi. A kriging approach based on aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment*, 28(7):1835–1851, 2014.

[20] A. Menafoglio, P. Secchi, and A. Guadagnini. A Class-Kriging predictor for Functional Compositions with application to particle-size curves in heterogeneous aquifers. MOX-report 58/2014, Politecnico di Milano, 2014.

[21] M. Noirhomme-Fraiture and P. Brito. Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, 4(2):157–170, 2011.

[22] V. Pawlowsky-Glahn, J.J. Egozcue, and R. Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. Wiley, Chichester, 2015.

[23] J. Ramsay and B.W. Silverman. *Functional Data Analysis, 2nd ed.* Springer, New York, 2005.

[24] H. Scheffé. Experiments with mixtures. *Journal of the Royal Statistical Society - B*, 20:344–360, 1958.

[25] R. Tolosana-Delgado, K.G. van den Boogaart, T. Mikes, and H. von Eynatten. Statistical treatment of grain-size curves and empirical distributions: densities as compositions? In *Proceedings of CoDaWork 2008*, 2008.

[26] K.G. van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn. Bayes linear spaces. *SORT*, 34(2):201–222, 2010.

[27] K.G. van den Boogaart, J.J. Egozcue, and V. Pawlowsky-Glahn. Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194, 2014.

[28] H. Wang, L. Shangguan, J. Wu, and R. Guan. Multiple linear regression modeling for compositional data. *Neurocomputing*, 122:490–500, 2013.

MOX Technical Reports, last issues

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

04/2017 Dede', L; Garcke, H.; Lam K.F.
A Hele-Shaw-Cahn-Hilliard model for incompressible two-phase flows with different densities

02/2017 Arena, M.; Calissano, A.; Vantini, S.
Monitoring Rare Categories in Sentiment and Opinion Analysis - Expo Milano 2015 on Twitter Platform.

03/2017 Fumagalli, I.; Parolini, N.; Verani, M.
On a free-surface problem with moving contact line: from variational principles to stable numerical approximations

01/2017 Riccobelli, D.; Ciarletta, P.
Rayleigh-Taylor instability in soft elastic layers

58/2016 Antonietti, P. F.; Bruggi, M. ; Scacchi, S.; Verani, M.
On the Virtual Element Method for Topology Optimization on polygonal meshes: a numerical study

57/2016 Bassi, C.; Abbà, A.; Bonaventura, L.; Valdettaro, L.
Large Eddy Simulation of gravity currents with a high order DG method

56/2016 Guerciotti, B.; Vergara, C.; Ippolito, S.; Quarteroni, A.; Antona, C.; Scrofani, R.
A computational fluid-structure interaction analysis of coronary Y-grafts

55/2016 Antonietti, P. F.; Facciola' C.; Russo A.; Verani M.;
Discontinuous Galerkin approximation of flows in fractured porous media on polytopic grids

54/2016 Vergara, C.; Le Van, D.; Quadrio, M.; Formaggia, L.; Domanin, M.
Large Eddy Simulations of blood dynamics in abdominal aortic aneurysms

52/2016 Paolucci, R.; Evangelista, L.; Mazzieri, I.; Schiappapietra, E.
The 3D Numerical Simulation of Near-Source Ground Motion during the Marsica Earthquake, Central Italy, 100 years later