MOX–Report No. 25/2014

Simplicial principal component analysis for density functions in Bayes spaces

Hron, K.; Menafoglio, A.; Templ, M.; Hruzova K.; Filzmoser, P.

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

Simplicial principal component analysis for density functions in Bayes spaces

Karel Hron[1], Alessandra Menafoglio[2], Matthias Templ[3,4], Klara Hrůzová[5] and Peter Filzmoser[3]

[1]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Olomouc, Czech Republic.
[2]MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy
[3]Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria
[4]Department of Methodology, Statistics Austria, Vienna, Austria hronk@seznam.cz; alessandra.menafoglio@polimi.it; templ@statistik.tuwien.ac.at; klara.hruzova@gmail.com; p.filzmoser@tuwien.ac.at

Abstract

Probability density functions are frequently used to characterize the distributional properties of large-scale database systems. As functional compositions, densities carry primarily relative information. As such, standard methods of functional data analysis (FDA) are not appropriate for their statistical processing. The specific features of density functions are accounted for in Bayes spaces, which result from the generalization to the infinite dimensional setting of the Aitchison geometry for compositional data. The aim of the paper is to build up a concise methodology for functional principal component analysis of densities. We propose the simplicial functional principal component analysis (SFPCA), which is based on the geometry of the Bayes space $\mathcal{B}^2$ of functional compositions. We perform SFPCA by exploiting the centred log-ratio transform, an isometric isomorphism between $\mathcal{B}^2$ and L^2 which enables one to resort to standard FDA tools. Advances of the proposed approach are demonstrated using a real-world example of population pyramids in Upper Austria.

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

1 Introduction

An increasing number of studies are nowadays based on large amounts of data. As a direct consequence, the importance of Functional Data Analysis (FDA, e.g., Ramsay

and Silverman, 2002, and references therein) has recently strongly increased. In recent years, a large body of literature has been developed in this field (e.g., Ramsay and Silverman, 2005; Horváth and Kokoszka, 2012, and references therein), however, still little attention has been paid to the problem of dealing with functional data that are probability density functions (Delicado, 2007, 2011; Nerini and Ghattas, 2007; Zhang and Müller, 2011; Menafoglio et al., 2014). Even though it might seem that density functions are just a special case of functional data –with a constant-integral-constraint equal to one– standard FDA methods appear to be inappropriate for their treatment, as they do not account for the particular constrained nature of the data. This problem is well known in the finite dimensional setting, where specific techniques have been worked out to deal with compositional data (e.g., Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001; Egozcue and Pawlowsky-Glahn, 2006; Egozcue, 2009; Pawlowsky-Glahn and Buccianti, 2011, and references therein), i.e., multivariate data carrying only relative information, usually represented in proportions or percentages. Those techniques are mainly based on a geometric perspective grounded on the Aitchison geometry in the simplex, which properly accounts for the compositional nature of the data. In this context, probability density functions have been recently interpreted as functional compositional data, i.e., functional data carrying only relative information. To handle this kind of data, the Aitchison geometry has been recently extended to the so called Bayes spaces: a Hilbert space structure for σ-finite measures, including probability measures, has been worked out in (van den Boogaart et al., 2014), based on the pioneering work of Egozcue et al. (2006) and the subsequent development of (van den Boogaart et al., 2010; Egozcue et al., 2013). The name Bayes spaces comes from the primary purpose of the approach, which is to assign a simple algebraic interpretation for the basic notions of mathematical statistics (e.g., the Bayes theorem as a paradigm of information acquisition van den Boogaart et al., 2010). The idea of Bayes spaces was first exploited in Delicado (2011) for the statistical analysis of density functions in the context of dimensionality reduction through functional principal component analysis and multidimensional scaling. Very recently, the Hilbert space structure of probability density functions with a compact support has been used by Menafoglio et al. (2014) to work out a kriging methodology for probability density functions. Even though the last developments in this field (van den Boogaart et al., 2014) enable one to deal with general σ-finite measures which are not necessarily compactly supported, this general theory seems to be still hard to be used in practice, mainly due to its highly technical construction involving reference measures different from the Lebesgue measure. However, from the application viewpoint, the hypothesis of finite support does not appear to be too restrictive (Delicado, 2011; Menafoglio et al., 2014), and thus we shall focus on compactly supported probability density functions.

The aim of this work is to make a step forward in the direction of functional principal component analysis in Bayes spaces, moving from the work of Delicado (2011). In particular, we shall geometrically work out the problem of functional principal component analysis (FPCA) in the Bayes space of probability density functions. Furthermore, we will propose the use of the centred log-ratio transform (clr, van den Boogaart et al., 2014; Menafoglio et al., 2014) for its practical implementation. We remark that, unlike

the non-linear transformations which are commonly used for probability density functions (e.g., the logarithmic transformation Delicado, 2011), the centred log-ratio transformation is an isometric isomorphism between the Bayes space of probability density functions and the space L^2 of square-integrable real measurable functions. From an application viewpoint this is extremely important, as it allows to solve the problem of FPCA with the usual L^2 geometry, while accounting for the non-linear geometry of Bayes spaces.

The remaining part of the paper is organized as follows. Section 2 introduces the Bayes space of probability densities functions as well as the clr transformation which will be used for their processing. FPCA is recalled in Section 3 for L^2 data. The extension of FPCA to Bayes spaces is proposed in Section 4. In Section 5, we apply the developed methodology to a real case study dealing with the age distributions in Upper Austria, with the aim of characterizing the main modes of variability of these densities. Section 6 eventually concludes the work.

2 Density functions as elements of the Bayes space

The theory of Bayes spaces (Egozcue and Pawlowsky-Glahn, 2006; van den Boogaart et al., 2010, 2014; Egozcue et al., 2013) has been introduced as a generalization to density functions of the Aitchison geometry. This is commonly used for compositional data, i.e., multivariate observations carrying only relative information (e.g., Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011, and references therein), which are usually collected in the form of constrained data summing up to a constant, usually set to 1 or 100, in case of proportions or percentages, respectively. Any probability density function $f(x)$ can be considered as a compositional vector with infinitely many parts (Egozcue and Pawlowsky-Glahn, 2006): as such, it inherits the key features of compositions (Egozcue, 2009).

Consider a σ-finite measure with support I defined on a measurable space $(\Omega, \mathcal{A})$. Like for compositional data, the constant sum constraint $\int_I f(x)\,\mathrm{d}x = 1 = \mathsf{P}(\Omega)$ leads to a representation within a class of functions that provide the same kind of information – namely, the equivalence class of functions which are proportional to the density function. In fact, any other representative $\widetilde{f}$ within this class, characterized by a constraint $\int_I \widetilde{f}(x)\,\mathrm{d}x = c$ for $c \in \mathbb{R}$, would carry the same information regarding the relative contribution of any Borel subsets of the real line to the measure of the support. This property is known as *scale invariance*. A second important feature of functional compositions is the *relative scale* property: the relative increase of a probability over a Borel set from 0.05 to 0.1 (twice a lot) differs from the increase 0.5 to 0.55 (1.1 multiple), although the absolute differences are the same in both cases.

Both the scale invariance and the relative scale properties are completely ignored when considering probability density functions just like unconstrained functional data. In particular, the usual notion of sum and product by a constant appears inappropriate when applied to compositions, as the space of functional compositions endowed with those operations is not a vector space (e.g., the point-wise sum of two compositions

is not a composition). Instead, the Hilbert space structure of Bayes spaces (van den Boogaart et al., 2014), which is based on an appropriate geometry, enables one to capture and properly account for these properties. In the following, we restrict our attention to density functions with compact support, as in (Delicado, 2011). Both theoretical and practical reasons motivate this choice. Indeed, when the support is the whole real line, the Lebesgue measure cannot be used as reference probability measure, leading to highly technical issues. Moreover, in most real datasets, finite values for the inferior and superior extremes of the support can be determined without a substantial loss of generality.

We call $\mathcal{B}^2(I)$ the Bayes space of (equivalence classes of) nonnegative functional compositions on a compact subset I of $\mathbb{R}$ with square-integrable logarithm (Egozcue et al., 2006; van den Boogaart et al., 2014). In the following, I will denote an interval $[a, b]$, but any compact subset of $\mathbb{R}$ could be dealt with analogously. Given two absolutely integrable density functions $f, g \in \mathcal{B}^2(I)$ and a real number $\alpha \in \mathbb{R}$ we indicate with $f \oplus g$ and $\alpha \odot f$ the perturbation and powering operation, respectively, defined as (Egozcue et al., 2006; van den Boogaart et al., 2014):

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)\,\mathrm{d}s}, \quad (\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha\,\mathrm{d}s}, \quad t \in I.$$

The resulting functions are readily seen to be probability density functions. Moreover, Egozcue et al. (2006) prove that $B^2(I)$ endowed with the operations $(\oplus, \ominus)$ is a vector space.

To endow $B^2(I)$ with a Hilbert space structure, Egozcue et al. (2006) define the inner product

$$\langle f, g\rangle_B = \frac{1}{2\eta}\int_I\int_I \ln\frac{f(t)}{f(s)}\ln\frac{g(t)}{g(s)}\,\mathrm{d}t\,\mathrm{d}s, \quad f, g \in \mathcal{B}^2(I) \tag{1}$$

which induces the following norm,

$$||f||_B = \left[\frac{1}{2\eta}\int_I\int_I \ln^2\frac{f(t)}{f(s)}\,\mathrm{d}t\,\mathrm{d}s\right]^{1/2},$$

where η stands for the length of the compact set I, namely $\eta = b - a$. $\mathcal{B}^2(I)$, endowed with the inner product (1), is proved to be a separable Hilbert space in (Egozcue et al., 2006). As such, it is isomorphic to the Hilbert space $L^2(I)$ of (equivalence classes of) square-integrable real functions on I. An isometric isomorphism between $\mathcal{B}^2(I)$ and $L^2(I)$ is defined by the *centred log-ratio* (clr) transformation (van den Boogaart et al., 2014; Menafoglio et al., 2014), defined for $f \in \mathcal{B}^2(I)$ as

$$\mathrm{clr}(f)(t) = f_c(t) = \ln f(t) - \frac{1}{\eta}\int_I \ln f(s)\,\mathrm{d}s. \tag{2}$$

We remark that such an isometry allows to compute operations and inner products among the elements in $\mathcal{B}^2(I)$ in terms of their counterpart in $L^2(I)$ among the clr-

transforms, i.e.

$$\begin{aligned} \mathrm{clr}(f \oplus g)(t) &= f_c(t) + g_c(t), \\ \mathrm{clr}(\alpha \odot f)(t) &= \alpha \cdot f_c(t), \\ \langle f, g \rangle_B &= \langle f_c, g_c \rangle_2 = \int_I f_c(t) g_c(t)\, \mathrm{d}t. \end{aligned}$$

However, the additional condition $\int_I f_c(t)\, \mathrm{d}t = 0$, needs to be taken into account for computation and analysis on clr transformed density functions, as we shall show in Section 4.

3 Principal component analysis for functional data

Principal component analysis (PCA) is a widely used multivariate statistical technique aiming to capture the main modes of variability of the data by means of a small number of linear combinations of the original variables. In the functional context, the same aim is reached by Functional Principal Component Analysis (FPCA). Here, we briefly recall FPCA, referring the reader, e.g., to Shang (2014) for a survey on this topic.

Let us consider a functional random sample $X_1, ..., X_N$ in $L^2(I)$, and indicate with $\langle x, y \rangle_2 = \int_I x(t)y(t)dt$ the inner product between two elements x, y in $L^2(I)$ and with $\|x\|_2 = (\int_I |x(t)|^2 dt)^{1/2}$ the induced norm. For ease of notation and without loss of generality, we assume the samples to be centred. FPCA looks firstly for the main mode of variability, i.e., for the element ξ_1 in $L^2(I)$ –called first functional principal component (FPC)– maximizing over $\xi \in L^2(I)$

$$\frac{1}{N} \sum_{i=1}^{N} \langle X_i, \xi \rangle_2^2 \text{ subject to } \|\xi\|_2 = 1. \tag{3}$$

The remaining FPCs, $\{\xi_j\}_{j \geq 2}$, capture the remaining modes of variability subject to be mutually orthogonal, and are thus obtained by solving problem (3) with the additional orthogonality constraint $\langle \xi_k, \xi \rangle_2 = 0, k < j$.

Analogously to the multivariate case, the FPCs $\{\xi_j\}$ coincide with the eigenvectors of the sample covariance operator $V : L^2(I) \to L^2(I)$ (e.g., Horváth and Kokoszka, 2012), acting on $x \in L^2(I)$ as

$$Vx = \frac{1}{N} \sum_{i=1}^{N} \langle X_i, x \rangle_2 X_i,$$

or, equivalently,

$$Vx = \int_I v(\cdot, t) x(t) dt,$$

the kernel $v : L^2(I) \times L^2(I) \to \mathbb{R}$ being the sample covariance function

$$v(s,t) = \frac{1}{N} \sum_{i=1}^{N} x_i(s) x_i(t), \quad s, t \in I.$$

Therefore, the j-th FPC ξ_j and the associated scores $\Psi_{ij} = \langle X_i, \xi_j \rangle_2$, $i = 1, ..., N$, are obtained by solving the eigenvalue equation

$$V\xi_j = \rho_j \xi_j, \tag{4}$$

where ρ_j denotes the j-th eigenvalue, with $\rho_1 \geq \rho_2 \geq ...$. As in multivariate PCA, for each j, the eigenvalue ρ_j is associated with the proportion of total variability explained by the FPC ξ_j.

Several computational methods can be employed to solve equation (4) (e.g., Ramsay and Silverman, 2005; Jones and Rice, 1992; Kneip and Utikal, 1992, and references therein). Ramsay and Silverman (2005) suggest to express each datum X_i, $i = 1, ..., N$, as a linear combination of K known basis functions $\phi_1, ..., \phi_K$ and to solve the eigenproblem (4) through an appropriate matrix coefficient. Indeed, suppose that each datum X_i, $i = 1, ..., N$, admits the basis expansion

$$X_i(\cdot) = \sum_{k=1}^{K} c_{ik} \phi_k(\cdot), \tag{5}$$

where $c_{ik} = \langle X_i, \phi_k \rangle_2$, $k = 1, ..., K$, or, in matrix notation, $\mathbf{X}(\cdot) = \mathbf{C}\boldsymbol{\phi}(\cdot)$, with $\mathbf{C} = (c_{ik}) \in \mathbb{R}^{N,K}$, $\mathbf{X}(\cdot) = (X_i(\cdot))$, and $\boldsymbol{\phi}(\cdot) = (\phi_i(\cdot))$. Then the variance-covariance function takes the form

$$v(s,t) = N^{-1} \boldsymbol{\phi}(s)' \mathbf{C}' \mathbf{C} \boldsymbol{\phi}(t), \quad s, t \in I.$$

Suppose further that the eigenfunction ξ_j, $j \geq 1$, admits the expansion

$$\xi_j(\cdot) = \sum_{k=1}^{K} b_{jk} \phi_k(\cdot),$$

$b_{jk} = \langle \xi_j, \phi_k \rangle_2$, $k = 1, ..., K$, or in matrix notation $\xi_j(\cdot) = \boldsymbol{\phi}(\cdot)' \mathbf{b}_j$. This yields

$$V\xi_j(\cdot) = \boldsymbol{\phi}(\cdot)' N^{-1} \mathbf{C}' \mathbf{C} \mathbf{W} \mathbf{b}_j,$$

where $\mathbf{W}_{kl} = \langle \phi_k, \phi_l \rangle_2$. Therefore the eigenvalue equation (4) reduces to

$$N^{-1} \mathbf{C}' \mathbf{C} \mathbf{W} \mathbf{b}_j = \rho_i \mathbf{b}_j, \tag{6}$$

and $\mathbf{b}_j$ is obtained as solution of the linear system (6). Note that in case of basis orthonormality $\mathbf{W} = \mathbf{I}$ the FPCA problem reduces to standard multivariate PCA of the coefficient matrix $\mathbf{C}$. Otherwise, Ramsay and Silverman (2005) show that problem (6) is equivalent to the eigenproblem

$$\frac{1}{N} \mathbf{W}^{1/2} \mathbf{C}' \mathbf{C} \mathbf{W}^{1/2} \mathbf{u} = \rho_i \mathbf{u}$$

with $\mathbf{u} = \mathbf{W}^{1/2} \mathbf{b}$, i.e., FPCA reduces to a multivariate PCA of the transformed coefficient matrix $\mathbf{C}\mathbf{W}^{1/2}$ followed by the transformation $\mathbf{b} = \mathbf{W}^{-1/2} \mathbf{u}$.

4 Simplicial Functional Principal Component Analysis

In this section, a simplicial version of FPCA will be derived by following the same scheme that led to the formulation of FPCs in Section 3. Let $X_1, ..., X_N$ be a centred sample in $\mathcal{B}^2(I)$. We consider the problem of finding the simplicial functional principal components (SFPCs) in $\mathcal{B}^2(I)$, i.e., the elements $\{\zeta_j\}_{j\geq 1}$, $\zeta_j \in \mathcal{B}^2(I)$, maximizing the following objective function over $\zeta \in \mathcal{B}^2(I)$:

$$\frac{1}{N}\sum_{i=1}^{N}\langle X_i, \zeta\rangle_B \text{ subject to } \|\zeta\|_B = 1;\ \langle \zeta, \zeta_k\rangle_B = 0,\ k < j, \tag{7}$$

where the orthogonality condition $\langle \zeta, \zeta_k\rangle_B = 0$, for $k < j$, holds only for $j \geq 2$.

$\mathcal{B}^2(I)$ being a separable Hilbert space, problem (7) is well posed (Horváth and Kokoszka (2012), Theorem 3.2). Indeed, analogously to the $L^2(I)$ case previously discussed, the j-th SFPC solves the eigenvalue equation

$$V\zeta_j = \lambda_j \odot \zeta_j,$$

(λ_j, ζ_j) being the j-th eigenpairs of the sample covariance operator $V : \mathcal{B}^2(I) \to \mathcal{B}^2(I)$, acting on $x \in \mathcal{B}^2(I)$ as

$$Vx = \frac{1}{N} \odot \bigoplus_{i=1}^{N}\langle X_i, x\rangle_B \odot X_i.$$

In order to proceed with (7) in practice, we apply the isometric isomorphism between $\mathcal{B}^2(I)$ and $L^2(I)$ defined by the clr-transform (2) that allows rewriting the problem (7) as a maximization of the term

$$\frac{1}{N}\sum_{i=1}^{N}\langle \text{clr}(X_i), \text{clr}(\zeta)\rangle_2 \text{ subject to } \|\text{clr}(\zeta)\|_2 = 1;\ \langle \text{clr}(\zeta), \text{clr}(\zeta_k)\rangle_2 = 0,\ k < j$$

over $\zeta \in \mathcal{B}^2(I)$. Therefore, for $j \geq 1$ the maximization problem (7) can be equivalently restated as

$$\frac{1}{N}\sum_{i=1}^{N}\langle \text{clr}(X_i), \xi\rangle_2 \text{ subject to } \|\xi\|_2 = 1;\ \langle \xi, \xi_k\rangle_2 = 0,\ k < j;\ \int_I \xi = 0, \tag{8}$$

where the orthogonality constraint is meaningful only for $j \geq 2$ and the zero-integral constraint accounts for the corresponding clr-transform property.

We now show that (8) is solved by the eigenvectors $\{\xi_j\}_{j\geq 1}$ of the sample covariance operator $V_{\text{clr}} : L^2(I) \to L^2(I)$ of the transformed sample $\text{clr}(X_1), ..., \text{clr}(X_N)$, acting on $x \in L^2(I)$ as

$$V_{\text{clr}}x = \frac{1}{N}\sum_{i=1}^{N}\langle \text{clr}(X_i), x\rangle_2 \,\text{clr}(X_i).$$

We first notice that, as in the previous case, the eigenvectors $\{\xi_j\}_{j\geq 1}$ would have solved problem (8), if it had been stated without the zero-integral condition $\int_I \xi = 0$. Therefore, it suffices to show that ξ_j fulfills $\int_I \xi_j = 0$ for all $j \geq 1$. To this end, we note that the zero-integral property of the clr-transformed sample implies that V_{clr} admits a zero eigenvalue with associated eigenvector $\xi_0 \equiv 1/\sqrt{b-a}$:

$$V_{\text{clr}}\, \xi_0 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{b-a}} \left[\int_I \text{clr}(X_i) \right] \text{clr}(X_i) \equiv 0.$$

Since the eigenvectors $\{\xi_j\}$ corresponding to the remaining eigenvalues $\{\rho_j\}$ are to be orthogonal to the eigenvector ξ_0, the ξ_j's need to satisfy the zero-integral condition $\int_I \xi_j = 0$, as $\langle \xi_j, \xi_0 \rangle_2 = 1/\sqrt{b-a} \int_I \xi_j$. Therefore, problem (7) can be restated in terms of clr-transforms as (8) and the SFPCs can be obtained by transforming the eigenvectors $\{\xi_j\}_{j\geq 1}$ associated to the non-null eigenvalues $\{\rho_j\}_{j\geq 1}$ of V_{clr} through the inverse of the function clr, namely $\zeta_j = \text{clr}^{-1}(\xi_j)$, with $\xi_j \neq \xi_0$.

To compute the eigenvectors ξ_j we resort to a method based on a B-spline basis expansion. Following Machalová et al. (2014), we consider for $\text{clr}(X_1), ..., \text{clr}(X_N)$ and ξ_j, $j \geq 1$, a B-spline basis fulfilling the zero-integral constraint through a zero-sum condition on the coefficients,

$$\begin{aligned} \text{clr}(X_i)(\cdot) &= \sum_{k=1}^{K} c_{ik} \phi_k(\cdot), \quad \sum_{k=1}^{K} c_{ik} = 0, \\ \xi_j(\cdot) &= \sum_{k=1}^{K} b_{jk} \phi_k(\cdot), \quad \sum_{k=1}^{K} b_{jk} = 0. \end{aligned} \tag{9}$$

Hence, with the same arguments used in Section 3, $\mathbf{b}_j = (b_{jk})$ is obtained as solution of the linear system

$$N^{-1} \mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b}_j = \rho_i \mathbf{b}_j,$$

where now the rows and columns of the matrix $\mathbf{C}'\mathbf{C}$ add up to zero due to the zero-sum constraints in (9). Also in this case, the zero sum constraint of $\mathbf{b}_j$ is inherently kept in the PCA algorithm.

The interpretation of SFPCs follows the main lines used in the $L^2(I)$ case, as the SFPCs represent the main modes of variability of the observations around the global mean function, but now within the space $\mathcal{B}^2(I)$ endowed with the Aitchison geometry. Hence, useful tools to visualize and interpret the results of SFPCs are the scores plan graph and the representation of the mean function perturbed by the j-th SFPC ξ_j powered by an appropriate constant, which in turn corresponds to the graphs of the mean $+/-$ the FPCs multiplied by a constant, advocated by Ramsay and Silverman (2005) in the $L^2(I)$ case.

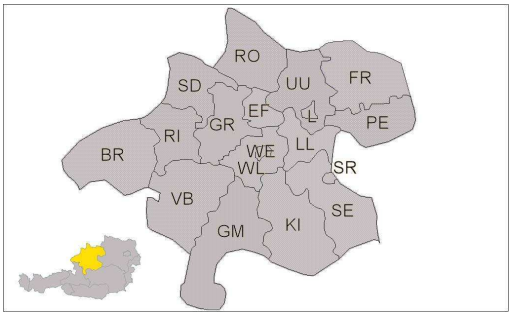


Figure 1: Upper Austria and its districts.

5 Analysis of Population Age Distributions in the Upper Austria Regions

In this section we perform the SFPCA of a real dataset dealing with the population age distributions in Upper Austria. This is the fourth-largest Austrian state in terms of land area and third-largest by population out of the nine states constituting Austria. Upper Austria is formed by 15 political districts, displayed in Figure 1. The dataset we consider collects the age distributions of men and women living in $N = 114$ municipalities of Upper Austria. We note that this kind of data is often referred to as *population pyramid* in the literature. A similar dataset –but referring to different countries in the world for the year 2000– has been considered in Delicado (2011) in the context of dimensionality reduction with particular emphasis on graphical displays. The aim of the current study is to characterize the available population age densities performing a dimensionality reduction according to the geometry of the Bayes space $\mathcal{B}^2(I)$ as opposed to the usual $L^2(I)$ geometry. For the purpose of the present analysis, the possible spatial dependence among the observations will not be considered. Instead, the geographical information will be taken into account for the interpretation of the scores.

The raw data have been smoothed by using the procedure detailed in (Machalová et al., 2014) and recalled in Section 4. In particular, the discrete clr-transforms of raw densities (Egozcue and Pawlowsky-Glahn, 2006) have been projected on a B-spline basis with compact support $I = [0, 100]$ and five equally spaced knots in $(0, 25, 50, 75, 100)$ years, with constraints to fulfill the zero-integral condition. The smoothed densities are displayed in Figure 2, coloured according to the gender information.

The smoothed data have been embedded into the space $\mathcal{B}^2(I)$ of functional compositions, and the methodology devised in Section 4 has been applied, resorting to the clr-transform (2) to make computations. In particular, here we consider the results of the SFPCA applied to the whole dataset –i.e., operated on the covariance operator estimated by simultaneously considering the groups of males and females. In this regard, we note that the estimated covariance structures of the two sub-populations appear pretty similar, even though the male group experiments a higher variability on the right tail of the distribution, as evidenced by the comparison of the sample covariance operators rela-

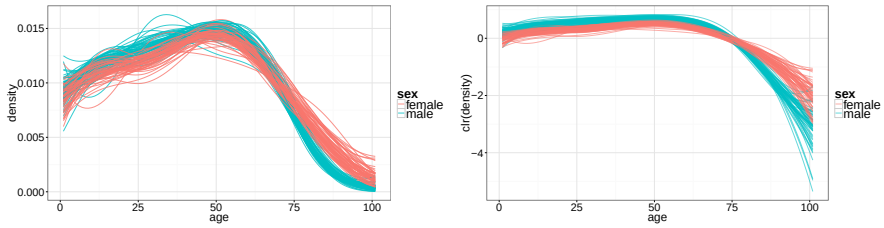
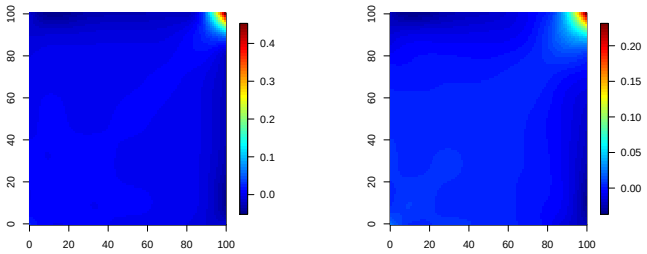


Figure 2: Population age densities in Upper Austria and their clr-transforms.

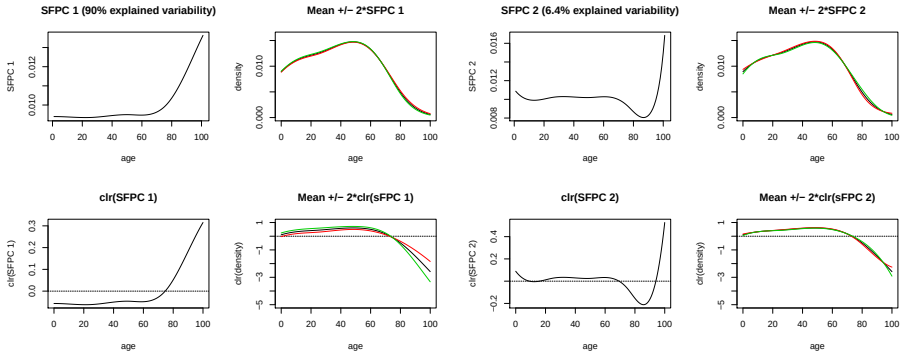


(a) Males (b) Females

Figure 3: Sample covariance operator estimated in $\mathcal{B}^2(I)$ from the smoothed data.

tive to the males and females groups (Figure 3a and 3b, respectively). In particular, the results of the separate SFPCA on the two subgroups appear consistent with the ones obtained on the whole dataset.

Figure 4 reports the first two SFPCs corresponding to the whole population, explaining 90% and 6.4% of the variability, respectively. To ease the SFPC interpretation, Figure 4 displays also the clr-transformed SFPC, together with the plot of the sample mean $+/-$ the clr-transform of the SFPCs multiplied by two. Figure 4a evidences that the first clr-SFPC contrasts the right tail of the distribution (age$>$75) against its left part (age$\leq$ 75). This is a clear consequence of the relative scale property of densities – captured by the clr transformation – that highlights the variability of small relative contributions. In particular, high scores in the first SFPC associates with a high incidence of the old population on the overall number of inhabitants; conversely, low scores associates with a high relative contribution of the youth to the overall population. The second SFPC, displayed in Figure 4b, still characterizes the variability of the right part of the distribution. Indeed, the main contribution to the second clr-SFPC is provided by the contrast between the 75-90 years-old population (associated with low scores) and the remaining part of the population (associated with high scores), with particular emphasis to the left and right boundaries of the densities support. We note particularly that high scores along the second SFPC associates with heavy tails and vice versa. However, we note that some uncertainty could affect the second estimated SFPC, due to the absence of data for ages lower than 2 years or higher than 92 years as these values formed



(a) First SFPC (b) Second SFPC

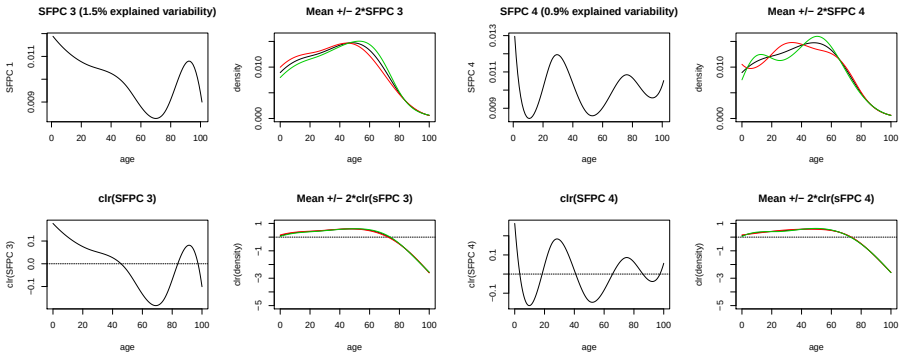
Figure 4: SFPCs and their clr-transform: SFPCs and the mean function perturbed by twice the SFPCs.

interval representatives in the aggregated data set.

Figure 5 reports the estimated third and fourth SFPCs. Although these components explain only 1.5% and 0.9% of the overall variability, the visualization of the mean function perturbed by $+/-$ twice the SFPCs seems to suggest that a very high variability is left to these components. This counterintuitive result is precisely due to specific features of the Bayes space $\mathcal{B}^2(I)$ as opposed to the standard geometry of the $L^2(I)$ space. By way of example, Figures 6-7 report the result of the FPCA performed according to the $L^2(I)$ geometry. We first notice that the $L^2(I)$ sample covariance operator (Figure 6) attributes more variability to the left part of the support than its $\mathcal{B}^2(I)$ counterpart. This directly reflects on the principal components: even though the first FPC (Figure 7a) is interpreted similarly to the first SFPC, it attributes much higher variability to the left part of the support. We further remark that the metric used to measure the variability readily reflect on the dimensionality reduction. Indeed, the scree-plot relative to the FPCA (gold line in Figure 6c) suggests the reduction to three or four SFPCs, as opposed to its SFPCA counterpart (black line in Figure 6c) which suggests a more synthetic representation based on one or two SFPCs instead.

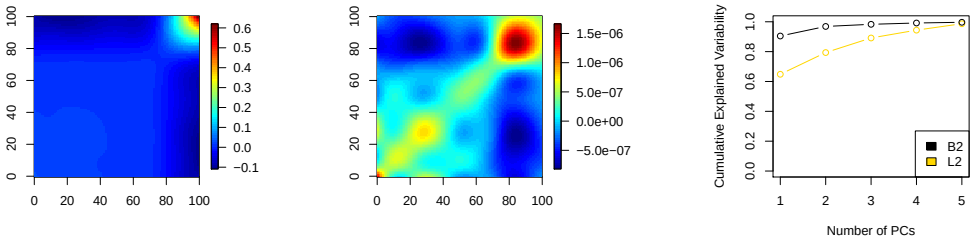
We focus on the latter case for the analysis of the scores. Figure 8 represents the plane of the scores relative to the first two SFPCs (left panel) and the second two SFPCs (right panel), coloured according to the gender information. This evidences that the first SFPC discriminates between the male and female subpopulations, the latter being associated with higher scores (i.e., higher life expectancy). This is readily interpretable in demographical and sociological terms, as male and female subpopulations are associated with different lifestyles, with a significant influence on the life expectancy.

In addition, the scores along the first two SFPCs seems to include also a geographical information. Indeed, the district of *Gmunden* (GM), see also Figure 1, shows high scores for both women and men along the first component (see Figure 9, in the light of Figure 8). This is also true for the South of *Steyr rural area* (SE). This evidences the fact that these regions are featured by a high incidence of the old population on the



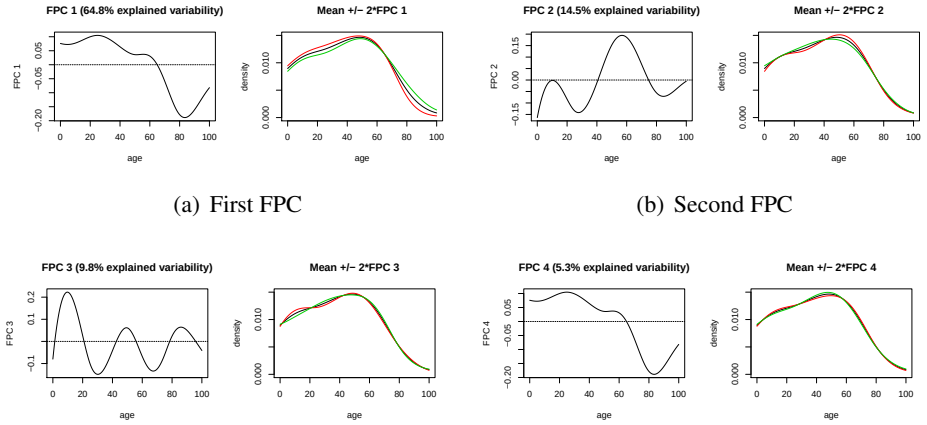
(a) Third SFPC (b) Fourth SFPC

Figure 5: SFPCs and their clr-transform: SFPCs and the mean function perturbed by twice the SFPCs.



(a) $\mathcal{B}^2$ sample covariance operator (b) L^2 sample covariance operator (c) Scree plot

Figure 6: Sample covariance operators estimated in $\mathcal{B}^2(I)$ and $L^2(I)$ from the smoothed data and scree plot for the SFPCA and FPCA.



(a) First FPC

(b) Second FPC

(c) Third FPC

(d) Fourth FPC

Figure 7: FPCs in the $L^2(I)$ geometry: FPCs and the mean function perturbed by twice the FPCs.

overall number of inhabitants. We also note that the North-West districts of *Braunau am Inn*, *Ried im Innkreis*, and *Grieskirchen* (BR, RI, GR) appear to be characterized by a high scores variability. In fact, these belong to a rural area mainly associated with low scores (i.e., younger population), characterized by some small towns of size between 10000 and 20000 inhabitants, which are represented by larger scores (i.e., older population). Overall, city and town areas prove to be associated with a very localized high incidence of the old population, which is particularly evident in the surrounding of *Linz* (L), the capital of Upper Austria which is nearby the center of the map in direction North-West.

Regarding the second SFPC (bottom panels in Figure 9), different regional structures appear in the North with respect to the rest of the map. Indeed, the district of *Rohrbach* (RO) is associated with pretty low scores for men (i.e., high incidence of 75-90 years old population), and very low scores for women. Similarly, very low scores appear for women in the *Schärding* regions (SD). These regions are mostly rural, with very few industries. Instead, high scores are recorded for women in the region around *Linz* (L), which also experiments high SFPC 1 scores. This is interpreted as an incidence of the old female population which is high, but still less pronounced than in the southern district of *Gmunden*, with a more significant incidence of the very young population ($<$7 years). Interestingly, this pattern is not so evident in the male population. High scores along the second SFPC are also evident in the industrialized regions of *Wels* (WE and WL) and *Vöcklabruck* (VB), which form a belt from *Linz* to the South-West and are characterized with a very young population (both for males and females).

In conclusion, the first SFPC scores appear to increase moving from the North to the South, except for the city and town areas. The second SFPC scores are mainly associated with the industrial districts, above all in the female subpopulation, evidencing quite a high incidence of the very young population in the *Linz* area as opposed to the

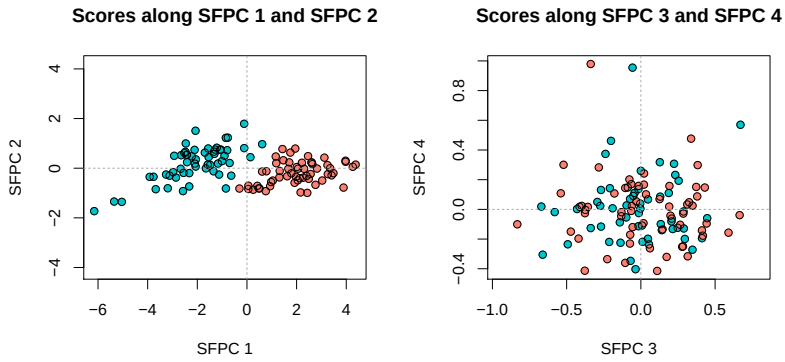


Figure 8: Scores along the first four SFPCs, coloured according to the gender information.

more rural districts of *Rohrbach* and *Schärding* (RO and SD).

6 Conclusions

The choice of an appropriate space to perform the analysis is crucial prior to any statistical processing using FDA methods. This is particularly evident in the presence of constrained data, such as functional compositions. We focused on the problem of the dimensionality reduction on probability density functions. In this case, although the numerical problems resulting from the unit-integral constraint of densities could be overcome by applying an appropriate preprocessing (e.g., log-transformation as proposed in Ramsay and Silverman (2005)), their inherent properties – as scale invariance and relative scale – are only captured by using the Bayes spaces methodology. The centred log-ratio transformation isometrically maps the Bayes space $\mathcal{B}^2(I)$ into the space $L^2(I)$, and it provides a way to easily apply the standard FDA methods in the presence of functional compositions. In this sense, it is possible to meaningfully employ the standard tools for the interpretation of FPCA (e.g., the plot of mean $+/-$ eigenvectors), if interpreted in the light of the Aitchison geometry which takes into account the relative information captured by density functions. We note that our proposal stands in continuity with the work of Delicado (2011), who pointed out that his most promising results were obtained through multidimensional scaling in the Bayes space $\mathcal{B}^2(I)$. However, out methodological developments provide a clear direction for the extension to density functions of several methods in use in FDA, besides opening a variety of further challenges for the future. One of these is the possibility of considering alternative computational tools, such as the log-ratio transformations with respect to an orthonormal basis in the Bayes space, with the aim of avoiding the zero-integral constraint resulting from the clr transformation. Even more promising could be the possibility of extending the support I to the general case of the whole real line (or any Borel subset) as proposed in van den Boogaart et al. (2014). The choice of a non-uniform reference measure still

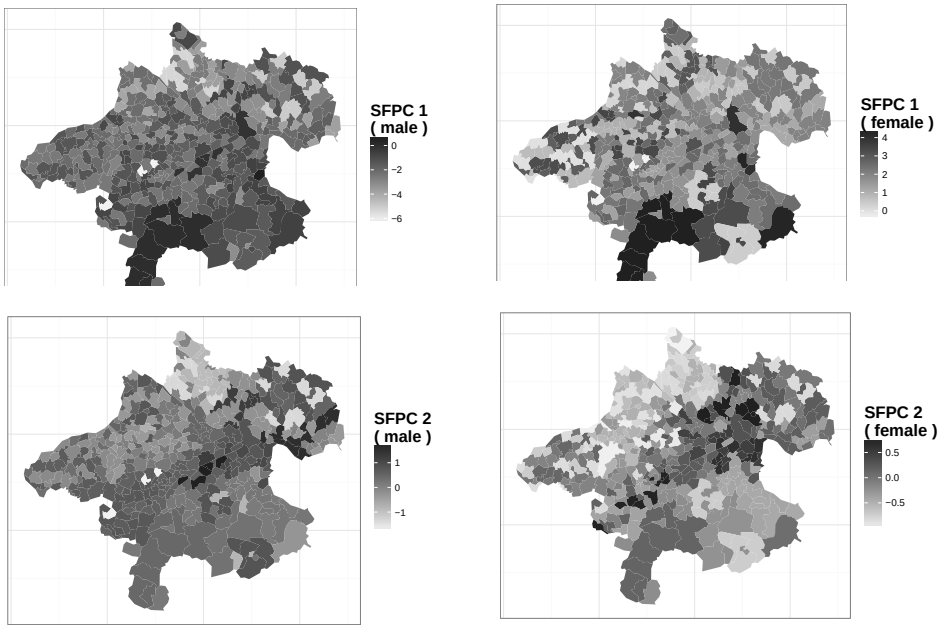


Figure 9: Geographical regional representation of SFPC 1 and SFPC 2 scores.

needs to be thoroughly discussed in terms of applicative consequences, and certainly deserves to be further investigated.

Acknowledgments

The authors gratefully acknowledge the support of the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic) and the grant IGA PrF 2014 028 Mathematical Models of the Internal Grant Agency of the Palacky University in Olomouc. Dr. Jitka Machalová (Palacký University in Olomouc) is thanked for computation of smoothed densities in Section 5.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd. (Reprinted 2003 with additional material by The Blackburn Press), London (UK). 416 p.

Delicado, P., 2007. Functional k-sample problem when data are density functions. Computational Statistics 22, 391–410.

Delicado, P., 2011. Dimensionality reduction when data are density functions. Computational Statistics and Data Analysis 55, 401–420.

Egozcue, J.J., 2009. Reply to "On the Harker Variation Diagrams; ..." by J.A. Cortés. Mathematical Geosciences 41 (7), 829–834.

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

Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. In Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264, 145–160.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35, 279–300.

Egozcue, J.J., Pawlowsky-Glahn, V., Tolosana-Delgado, R., Ortego, M.I., van den Boogaart, K.G., 2013. Bayes spaces: use of improper distributions and exponential families. Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas, DOI 10.1007/s13398-012-0082-6.

Filzmoser, P., Hron, K., Reimann, C., 2009. Principal component analysis for compositional data with outliers. Environmetrics 20 (6), 621–632.

Filzmoser, P., Hron, K., 2013. Robustness for compositional data. In Becker, C., Fried, R., Kuhnt, S. (eds) Robustness and complex data structures. Springer, Heidelberg, 117–131.

Horváth, L., Kokoszka, P. (2012). Inference for Functional Data with Applications. Springer Series in Statistics. Springer.

Hron, K., Templ, M., Filzmoser, P., 2010. Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis 54 (12), 3095–3107.

Johnson, R.A., Wichern, D.W., 2002. Applied Multivariate Statistical Analysis. Prentice Hall, London, fifth edition.

Jones, M.C., Rice, J.A., 1992. Displaying the important features of large collections of similar curves. The American Statistician 46 (2), 140–145.

Kneip, A., Utikal, K., 2001. Inference for density families using functional principal component analysis. Journal of the American Statistical Association 96, 519–542.

Machalová, J., Hron, K., Monti, G.S., 2014. Smoothing splines for centred logratio transformed density functions. Manuscript.

Maronna R., Martin R.D., Yohai V.J., 2006. Robust Statistics: Theory and Methods. John Wiley, New York (USA). 436 p.

Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A., 2011. Dealing with zeros, Ch. 4. In Pawlowsky-Glahn and Buccianti (2011), pp. 47–62.

Menafoglio, A., Guadagnini, A., and Secchi P., 2014. A Kriging Approach based on Aitchison Geometry for the Characterization of Particle-Size Curves in Heterogeneous Aquifers. Stoch. Env. Res. Risk. A.. DOI: 10.1007/s00477-014-0849-8.

Nerini, D., Ghattas, B., 2007. Classifying densities using functional regression trees: applications in oceanology. Computational Statistics and Data Analysis 51, 4984–4993.

Pawlowsky-Glahn, V., Buccianti, A. (Eds.), 2011. Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd., Chichester (UK). 378 p.

Pawlowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment 15 (5), 384–398.

R development core team, 2008, R: A language and environment for statistical computing: Vienna, `http://www.r-project.org`.

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

Ramsay, J., Silverman, B.W., 2005. Functional Data Analysis, 2nd ed.. Springer, New York.

Rousseeuw, P., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Shang, H.L., 2014. A survey of functional principal component analysis. Advances in Statistical Analysis, 98, 121–142.

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

Van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V., 2014. Bayes Hilbert spaces. Australian & New Zealand Journal of Statistics, 56(2), 171-194.

Zhang, Z., Müller, H.G., 2011. Functional density synchronization. Computational Statistics and Data Analysis, 55, 2234–2249.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

25/2014 Hron, K.; Menafoglio, A.; Templ, M.; Hruzova K.; Filzmoser, P.
Simplicial principal component analysis for density functions in Bayes spaces

24/2014 Ieva, F., Jackson, C.H., Sharples, L.D.
Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology

23/2014 Ieva, F., Paganoni, A.M., Tarabelloni, N.
Covariance Based Unsupervised Classification in Functional Data Analysis

22/2014 Arioli, G.
Insegnare Matematica con Mathematica

21/2014 Artina, M.; Fornasier, M.; Micheletti, S.; Perotto, S.
The benefits of anisotropic mesh adaptation for brittle fractures under plane-strain conditions

20/2014 Artina, M.; Fornasier, M.; Micheletti, S.; Perotto, S.
Anisotropic mesh adaptation for crack detection in brittle materials

19/2014 L.Bonaventura; R. Ferretti
Semi-Lagrangian methods for parabolic problems in divergence form

18/2014 Tumolo, G.; Bonaventura, L.
An accurate and efficient numerical framework for adaptive numerical weather prediction

17/2014 Discacciati, M.; Gervasio, P.; Quarteroni, A.
Interface Control Domain Decomposition (ICDD) Method for Stokes-Darcy coupling

15/2014 Esfandiar, B.; Porta, G.; Perotto, S.; Guadagnini, A;
Anisotropic mesh and time step adaptivity for solute transport modeling in porous media