

Dimensionality reduction when data are density functions

1

Pedro Delicado²

Dept. d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya,
Barcelona, Spain, e-mail: pedro.delicado@upc.edu

Abstract: Functional data analysis (FDA) deals with samples where a whole function is observed for each individual. A particular case is when the observed functions are densities, that is also an example of infinite dimensional compositional data. We focus on the dimensionality reduction problem for this particular type of data. Several methods are considered here: functional principal components analysis (FPCA), multidimensional scaling (MDS) and variations of both. Special attention is given to the standard graphics used to represent the output of these procedures. We show that some of them are no valid when working with density functions. Alternative graphical outputs are suggested. Both artificial and real data (population pyramids of all countries in the world) illustrate our proposals.

Keywords: Compositional data, functional data analysis, multidimensional scaling, population pyramids, principal components analysis.

1. Introduction

Functional Data Analysis (FDA) deals with the statistical description and modeling of samples of random functions. Functional versions for a wide range of statistical tools (ranging from exploratory and descriptive data analysis to linear models or multivariate techniques) have been developed. Others techniques are specific of FDA, because they exploit the functional nature of this kind of data. See Ramsay and Silverman (2005) for a general perspective on FDA and Ferraty and Vieu (2006) for a non-parametric approach. Ramsay and Silverman (2002) present applications of FDA to a wide range of problems and disciplines. A particular case of functional data appears when the observed functions are density functions, that are also an example of infinite dimensional compositional data (Egozcue *et al.*, 2006).

Assume we have observed n functions $f_1, \dots, f_n$. In general, they belong to an infinite-dimensional functional space. The dimensionality reduction problem consists in looking for a low dimensional configuration X (a $n \times q$ matrix, $q < n$, with rows x_i , $i = 1, \dots, n$) and an application ρ from $\mathbf{R}^q$ to the functional space such that $\rho(x_i)$ is close (in some sense) to the observed f_i , for $i = 1, \dots, n$. Usually, dimensionality reduction aims at visualizing data, which requires a plane representation, that implies $q = 2$.

In this work we compare several methods for dimensionality reduction for this particular type of data: functional principal components analysis (FPCA) with or without a previous data transformation and multidimensional scaling (MDS) for different inter-densities

¹The present paper is financially supported by the Spanish Ministerio de Educación y Ciencia and FEDER grant MTM2006-09920.

²Address of correspondence: C/Jordi Girona 1-3, Campus Nord, Edifici C5-214, 08034

distances, one of them taking into account the compositional nature of density functions. Special attention is given to the standard graphics used to represent the output of these procedures. We show that some of these graphics are no valid when working with density functions. Alternative graphical outputs are suggested to overcome this problem. The proposed tools are applied to both artificial and real data (population pyramids of all countries in the world).

2. Dimensionality reduction for density functions

Ferraty and Vieu (2006) define a *functional variable* as a random variable $\mathbf{f}$ taking values in an infinite functional space, usually $L_2(I) = \{f : I \rightarrow \mathbf{R}, \text{ such that } \int_I f(t)^2 dt < \infty\}$, where $I = [a, b]$, with a and b real numbers or $\pm\infty$, and $a < b$. An observation f of $\mathbf{f}$ is called a *functional data*. A *functional data set* $f_1, \dots, f_n$ is the observation of n independent functional variables $\mathbf{f}_1, \dots, \mathbf{f}_n$ identically distributed as $\mathbf{f}$.

We are particularly interested in the case where the observed functions are density functions in $[a, b]$: they are non negative and integrate up to 1 on $[a, b]$. So we assume that the functional space in our case is

$$\mathcal{F}(I) = \{f : I \rightarrow \mathbf{R}, \text{ such that } f(t) \geq 0 \text{ for all } t \in I, \text{ and } \int_I f(t) dt = 1\}.$$

Egozcue *et al.* (2006) note that density functions share a series of features with *compositional data* (p -dimensional non-negative data with constant sum; they live in the simplex S^{p-1} when the sum is equal to 1). In fact Egozcue *et al.* (2006) suggest that density functions are infinite dimensional compositional data. Therefore they propose to extend Aitchison's geometry (well developed for data living in the finite dimensional simplex; see Pawlowsky-Glahn and Egozcue, 2001) to the infinite dimensional simplex $\mathcal{F}(I)$. In particular, when $I = [a, b]$ is a finite interval they give a Hilbert space structure to the subset of $\mathcal{F}(I)$ formed by densities whose logarithm is square-integrable. In this space, the distance between two densities f and g is defines as

$$d_A(f, g) = \left[\frac{1}{2\eta} \int_a^b \int_a^b \left(\log \frac{f(x)}{f(y)} - \log \frac{g(x)}{g(y)} \right)^2 dx dy \right]^{1/2},$$

where $\eta = b - a$ and the subscript A refers to Aitchison's geometry.

2.1 Functional Principal Component Analysis

In the context of FDA on $L_2(I)$, a version of the Principal Component Analysis (PCA) has been developed: the *Functional Principal Component Analysis (FPCA)*. The objective of FPCA can be stated as follows. Given a functional random sample with mean function $\bar{f}(t) = (1/n) \sum_{i=1}^n f_i(t)$, for all $t \in I$, we look for functions $g_1, \dots, g_q$ (*principal functions* or *principal components*) in $L_2(I)$ and real numbers ψ_{ij} , $i = 1, \dots, n$, $j = 1, \dots, q$, such that

$$\sum_{i=1}^n \int_I \left((f_i(t) - \bar{f}(t)) - \sum_{j=1}^q \psi_{ij} g_j(t) \right)^2 dt$$

was minimum. Moreover, the functions $g_1, \dots, g_q$ are asked to be orthonormal. In other words, we are looking for a representation of functional data in a q -dimensional space

(that spanned by the functions $g_1(\cdot), \dots, g_q(\cdot)$):

$$f_i(t) \approx \bar{f}(t) + \sum_{j=1}^q \psi_{ij} g_j(t), \quad t \in I, \quad i = 1 \dots n. \quad (1)$$

The dimensionality reduction problem is approached in this context by defining the matrix X with elements (i, j) equal to ψ_{ij} , $i = 1, \dots, n$, $j = 1, \dots, q$, also known as *PC scores*. Principal component functions can be interpreted as the main variation modes of the observed functions around the global mean function. A graphical output according to this idea is to plot the mean function $\bar{f}$ and add and subtract from it the eigen-function $g_j(t)$ multiplied by an appropriate constant C . We refer to them as “mean +/- PC j” graphics. This gives us an idea of how the observed functions differ from the mean for observations that have significant positive or negative values in this principal component. The other standard graphical output for FPCA is the scatterplot of k -th PC against j -th PC scores (“PC k vs PC j”), that are similar to those used in usual PCA.

Let us return to the case of observing density functions, so our functional data belongs to $\mathcal{F}(I)$, that in general does not coincide with $L_2(I)$. Therefore FPCA previously introduced could present some problems now, as the examination of the right hand side of Equation (1) reveals: it is not sure that $\tilde{f}_i(t) = \bar{f}(t) + \sum_{j=1}^q \psi_{ij} g_j(t)$ is always a density function, even having observed density functions. It is easy to prove that $\bar{f}(t)$ is a density function and that $\int_I g_j(t) dt = 0$, but $\tilde{f}_i(t)$ can be negative for some $t \in I$ because both ψ_{ij} and $g_j(t)$ can take negative values. A standard solution to this problem is to look for a transformation $\Psi : \mathcal{F}(I) \mapsto L_2(I')$ and then to apply FPCA to the transformed functions. For instance, $\Psi(f)(\cdot) = \log(f(\cdot))$ is a sensible choice when the transformed densities are in $L_2(I)$.

Graphics of type “mean +/- PC j” could be misleading when working with density functions. The main reason is that it is not guaranteed that the plotted functions, $\bar{f} \pm C g_j(t)$, are density functions. Moreover, if the observed densities belong to a subset of $\mathcal{F}(I)$ (for instance, if all of them are normal densities) it could be possible that the mean function does not belong to the subset, as well as function $\bar{f} \pm C g_j(t)$. Therefore the graphic “mean +/- PC j” leads to the erroneous idea that the linear subspace $\{\bar{f} \pm C g_j(t) : C \in \mathbf{R}\}$ belongs to the subset. Additionally, if a transformation Ψ is used these graphics are in the space of transformed densities, where the interpretation of PCs could be much more difficult than in $\mathcal{F}(I)$. In this work we propose a way to circumvent these difficulties.

2.2 Multidimensional Scaling

Multidimensional Scaling (MDS) is a generalization of PCA when the information about data is given by an inter-individuals distance matrix, instead of by a standard data matrix. Assume that there are n individuals and that a distance (or dissimilarity) function between individuals is available. Let $d_{ij} \geq 0$ be the dissimilarity between individuals i and j . It is assumed that $d_{ij} = d_{ji}$ and that $d_{ii} = 0$ for all $i, j = 1 \dots, n$. Let Δ be the $n \times n$ matrix with element (i, j) equal to d_{ij} . Assume that for $q \leq n$ there exists a $n \times q$ data matrix X such that the Euclidean distance between the i -th and j -th rows of X is d_{ij} . We say that X is an *Euclidean configuration* of Δ . Such a configuration does not always exist. When it does, Δ is said to be *Euclidean*. In this case the X can be chosen having orthogonal columns, that are called *principal coordinates*.

In this paper we apply MDS to the specific case of density functions. Many distances can be computed between density functions. For instance, given that density functions are al-

ways in L_1 , we can base MDS on L_1 distances between observed densities: $\|f_i - f_j\|_1 = \int_a^b |f_i(x) - f_j(x)| dx$. Other distances can also be used: for instance, the Hellinger distance (that is the L_2 distance between squared root of densities), the L_2 distance between densities (assuming they are well defined; for instance assuming that densities are bounded on $[a, b]$) the L_2 distance between logs of densities (assuming they are well defined) or the symmetrized version of the Kullback-Leibler divergence:

$$d_{KL}(f_i, f_j) = \int_a^b \log \left(\frac{f_i(x)}{f_j(x)} \right) f_i(x) dx + \int_a^b \log \left(\frac{f_j(x)}{f_i(x)} \right) f_j(x) dx.$$

More in general, let $\Psi : \mathcal{F}(I) \mapsto \mathcal{F}_\Psi$ be a transformation from $\mathcal{F}(I)$ to another functional space $\mathcal{F}_\Psi$, and let $d^*(\cdot, \cdot)$ be a distance between elements of $\mathcal{F}_\Psi$. Let f and g be two density functions, then $d_\Psi(f, g) = d^*(\Psi(f), \Psi(g))$ is a distance between f and g . For instance, assuming that $\mathcal{F}_\Psi = L_2(I')$ and using the L_2 distance in $L_2(I')$ as d^* , the results of MDS from distance d_Ψ coincide with those obtained by doing FPCA on the transformed functions $\Psi(f_i)$, $i = 1, \dots, n$.

Taking into account Egozcue *et al.* (2006), where density functions are seen as infinite dimensional compositional data and the corresponding Aitchison's distance d_A is introduced, MDS can also be applied to the matrix Δ_A of distances $d_A(f_i, f_j)$ between density functions. Observe that

$$d_A(f, g) = \frac{1}{\sqrt{2\eta}} d_{L_2(I \times I)}(f^*, g^*)$$

where $f^* : I \times I \mapsto \mathbf{R}$ is defined as $f^*(x, y) = \log(f(x)/f(y))$, and g^* is defined analogously. We conclude that Δ_A is a Euclidean matrix.

As regards graphical outputs for MDS, it is usual to plot column k against column j of the Euclidean configuration X . This is analogous to graphics of type “PC k vs PC j ” in FPCA. Nevertheless there are not analogous graphics to those of type “mean +/- PC j ”. In this work we propose a way to do it.

References

- Egozcue J.J., Díaz-Barrero J.L. and Pawlowsky-Glahn V. (2006) Hilbert space of probability density functions based on aitchison geometry, *Acta Mathematica Sinica*, 22, 1175–1182.
- Ferraty F. and Vieu P. (2006) *Non parametric functional data analysis. Theory and practice*, Springer.
- Pawlowsky-Glahn V. and Egozcue J.J. (2001) Geometric approach to statistical analysis on the simplex, *Journal Stochastic Environmental Research and Risk Assessment*, 15, 5, 384–398.
- Ramsay J. and Silverman B.W. (2002) *Applied Functional Data Analysis: Methods and Case Studies*, Springer-Verlag, New York.
- Ramsay J. and Silverman B.W. (2005) *Functional Data Analysis*, Springer, New York, second edition.