Computational Statistics and Data Analysis 104 (2016) 24–34

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Functional archetype and archetypoid analysis☆

Irene Epifanio*

Dept. Matemàtiques and Institut de Matemàtiques i Aplicacions de Castelló, Campus del Riu Sec. Universitat Jaume I, 12071 Castelló, Spain

ARTICLE INFO

Article history:
Received 5 November 2015
Received in revised form 16 June 2016
Accepted 17 June 2016
Available online 24 June 2016

Keywords:
Archetype analysis
Functional data analysis
Unsupervised learning
Extreme point
Global human development

ABSTRACT

Archetype and archetypoid analysis can be extended to functional data. Each function is approximated by a convex combination of actual observations (functional archetypoids) or functional archetypes, which are a convex combination of observations in the data set. Well-known Canadian temperature data are used to illustrate the analysis developed. Computational methods are proposed for performing these analyses, based on the coefficients of a basis. Unlike a previous attempt to compute functional archetypes, which was only valid for an orthogonal basis, the proposed methodology can be used for any basis. It is computationally less demanding than the simple approach of discretizing the functions. Multivariate functional archetype and archetypoid analysis are also introduced and applied in an interesting problem about the study of human development around the world over the last 50 years. These tools can contribute to the understanding of a functional data set, as in the classical multivariate case.

1. Introduction

Archetype analysis (AA) is a statistical technique that seeks to approximate data by a convex combination of pure or extremal types called archetypes. Archetypes are built as a convex combination of the observations. AA was first introduced by Cutler and Breiman (1994). More recently, archetypoid analysis (ADA) was introduced by Vinué et al. (2015a). Unlike AA, the pure types in ADA are not a mixture (convex combination) of observations, but real observations. It has been shown that human understanding and interpretation of data is made easier when they are represented by their extreme constituents (Davis and Love, 2010) by the principle of opposites (Thurau et al., 2012). In other words, extremes are better than central points for human interpretation. Their applications have been growing in recent years, especially after the AA algorithm was implemented in the R package **archetypes** (Eugster and Leisch, 2009). ADA is available in the R package **Anthropometry** (Vinué et al., 2015a). The fields of application include, for instance, market research (Li et al., 2003; Porzio et al., 2008; Midgley and Venaik, 2013), biology (D'Esposito et al., 2012), genetics (Thøgersen et al., 2013), sports (Eugster, 2012), industrial engineering (Epifanio et al., 2013; Vinué et al., 2015a), the evaluation of scientists (Seiler and Wohlrabe, 2013), astrophysics (Chan et al., 2003; Richards et al., 2012), e-learning (Theodosiou et al., 2013), multi-document summarization (Canhasi and Kononenko, 2013, 2014) and different machine learning problems (Mørup and Hansen, 2012; Stone, 2002).

In the seminal paper by Cutler and Breiman (1994), one of the illustrative examples worked with functional observations, i.e., data consisting of a set of functions, although they converted them into a matrix by considering a set of values of each curve (after being smoothed) at certain points. Functional data analysis (FDA) comprises statistical procedures for

☆ The code and data for reproducing the examples are available at http://www3.uji.es/~epifanio/RESEARCH/faa.rar.

* Fax: +34 964728429.
E-mail address: epifanio@uji.es.

http://dx.doi.org/10.1016/j.csda.2016.06.007
0167-9473/

functional observations (a whole function is a datum). The objectives of FDA are essentially the same as those of any other branch of statistics. Although FDA is relatively new field, some classic references in this field include: Ramsay and Silverman (2005), who provide an excellent overview, Ferraty and Vieu (2006), with new methodologies for studying functional data nonparametrically, and Ramsay and Silverman (2002), who offer interesting applications in different fields, and Ramsay et al. (2009) regarding software in this field. More recent advances and interesting applications of FDA comprise a variety of fields such as aviation safety (Gregorutti et al., 2015), chromatography (Rakêt and Markussen, 2014), the quality of cookies and the relationship with the flour kneading process (Jacques and Preda, 2014), the relationship between the geometry of the internal carotid artery and the presence or absence of an aneurysm (Usset et al., 2016; Sangalli et al., 2009) and the analysis of hippocampal differences in Alzheimer's disease (Epifanio and Ventura-Campos, 2011, 2014).

A first attempt to extend AA to functional data was made by Costantini et al. (2012). Functions were expressed in a functional basis, and standard multivariate AA was applied to the coefficients in this basis. This method is only valid when the basis is orthonormal. The same thing is true when computing standard principal component analysis (PCA) of the basis coefficients in order to carry out functional principal component analysis (FPCA). In this paper, a methodology is developed for obtaining functional archetypes and archetypoids, whatever the basis used for approximating the functions.

Interest in describing and displaying the important features of a set of curves is not recent. Jones and Rice (1992) considered curves with extreme scores of principal components. This could be viewed as searching the archetypoid functions. However, unlike PCA, the goal of AA is to obtain extreme individuals, and curves with extreme PCA scores do not necessarily return archetypal observations. This is explained in Cutler and Breiman (1994) and shown in Epifanio et al. (2013) through a problem where archetypes could not be recovered with PCA even if all the components had been considered.

In this paper, AA and ADA are extended to univariate and multivariate functional data (more than one function is available per individual). Section 2 presents a review of archetype and archetypoid analysis in the classical multivariate case, and their respective extensions to FDA are introduced and illustrated through a well-studied data set in the field of FDA. The location of functional archetypes and archetypoids is also analyzed. Human development is a topic of considerable political and public interest, suffice it to say that all United Nations member states committed to help achieve the Millennium Development Goals established in 2000, by 2015 (United Nations, 2015). In Section 3, the proposal is applied to understanding statistics about human development for all countries, in order to obtain the big picture of global development. This can make it easier to interpret the large amount of data about sustainable development even for non-experts. The code in R (R Development Core Team, 2015) and data for reproducing the results are available at http://www3.uji.es/~epifanio/RESEARCH/faa.rar. Conclusions and future work are discussed in Section 4.

2. Definition of functional AA and ADA

2.1. AA and ADA for (standard) multivariate data

Let $\mathbf{X}$ be an $n \times m$ matrix that contains a usual multivariate data set with n observations and m variables. The objective of AA is to find a $k \times m$ matrix $\mathbf{Z}$, whose rows are the k archetypes in those data, in such a way that data can be approximated by mixtures of the archetypes. To obtain the archetypes, AA computes two matrices α and β which minimize the residual sum of squares (RSS) that arises from combining the equation where $\mathbf{x}_i$ is approximated by a mixture of $\mathbf{z}_j$'s (archetypes) ($\sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2$) and the equation where $\mathbf{z}_j$'s is expressed as a mixture of the data ($\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$):

$$RSS = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j \right\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l \right\|^2, \tag{1}$$

under the constraints

(1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and
(2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1, \ldots, k$.

Constraint (1) means that the approximations of $\mathbf{x}_i$ are a convex combination of archetypes, $\hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$. Each α_{ij} is the weight of the archetype j for the observation i; that is to say, the α coefficients indicate how much each archetype contributes to the approximation of each observation. Constraint (2) means that archetypes $\mathbf{z}_j$ are a mixture of the observations, $\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$.

Note that archetypes are not necessarily actual observations. This would happen if only one β_{jl} is equal to 1 in constraint (2) for each j. This implies that β_{jl} can only take on the values 0 or 1, since $\beta_{jl} \geq 0$ and the sum of constraint (2) is 1. In ADA, the continuous optimization problem of AA transforms into the following mixed-integer optimization problem:

$$RSS = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j \right\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l \right\|^2, \tag{2}$$

under the constraints