La Matematica (2023) 2:635–658
https://doi.org/10.1007/s44007-023-00058-x

ORIGINAL RESEARCH ARTICLE

Archetypal Curves in the Shape and Size Space: Discovering the Salient Features of Curved Big Data by Representative Extremes

Irene Epifanio[1] · Vicent Gimeno[2] · Ximo Gual-Arnau[3] · M. Victoria Ibáñez-Gual[2]

Received: 29 December 2022 / Revised: 18 July 2023 / Accepted: 31 July 2023 /
Published online: 8 September 2023

Abstract

Curves are complex data. Tools for visualizing, exploring, and discovering the structure of a data set of curves are valuable. In this paper, we propose a scalable methodology to solve this challenge. On the one hand, we consider two distances in the shape and size space, one well-known distance and another recently proposed, which differentiate the contribution in shape and in size of the elements considered to compute the distance. On the other hand, we use archetypoid analysis (ADA) for the first time in elastic shape analysis. ADA is a recent technique in unsupervised statistical learning, whose objective is to find a set of archetypal observations (curves in this case), in such a way that we can describe the data set as convex combinations of these archetypal curves. This makes interpretation easy, even for non-experts. Archetypal curves or pure types are extreme cases, which also facilitates human understanding. The methodology is illustrated with a simulated data set and applied to a real problem. It is important to know the distribution of foot shapes to design suitable footwear that accommodates the population. For this purpose, we apply our proposed methodology to a real data set composed of foot contours from the adult Spanish population.

✉ Irene Epifanio
epifanio@uji.es

Vicent Gimeno
gimenov@uji.es

Ximo Gual-Arnau
gual@uji.es

M. Victoria Ibáñez-Gual
mibanez@uji.es

1 Department of Mathematics-IF, Universitat Jaume I, Campus del Riu Sec, 12071 Castelló, Spain

2 Department of Mathematics-IMAC, Universitat Jaume I, Campus del Riu Sec, 12071 Castelló, Spain

3 Department of Mathematics-INIT, Universitat Jaume I, Campus del Riu Sec, 12071 Castelló, Spain

Keywords Elastic shape analysis · Square-root velocity function (SRVF) · Archetype analysis · Unsupervised machine learning · Footwear · Anthropometry

Mathematics Subject Classification 62H99 · 62P30 · 53A04

1 Introduction

Shape analysis is an extensive area of research since there are many applications in which it is interesting to quantify the difference between shapes, register shapes, see how one shape deforms into another, calculate the mean shape, or classify and cluster different shapes. In the case of planar shapes, one way to approach their analysis is to consider landmarks on the boundary curve of the object of interest to characterize its shape [1]. However, choosing the appropriate landmarks for each shape and registering these points between different shapes is not always an easy task; therefore, the function that defines the boundary curve of the object can be used to characterize its shape. Then, every plane shape is characterized by a parameterized curve and the set of all plane shapes is a Riemannian manifold with the appropriate metric, see for instance [2]. Although a certain parameterization of the boundary curve can be fixed for each plane shape (for example, parameterization by arc length), in elastic shape analysis, the metric must be preserved under reparameterizations of the curve. In [3], the authors use the square-root velocity functions (SRVF) and propose an elastic metric for the study of shapes represented by curves, which is valid not only for plane curves but also for curves in Euclidean $n-$dimensional spaces. Depending on the applications, it may be interesting for the metric to be invariant with respect to translations, rotations, and scaling as well; in this case, the feature space is called the shape space. However, in many applications (for example, in the analysis of anatomical structures in medicine), it is also important to distinguish the size of curves, in addition to shape. In this second case, the feature space is called the shape and size space. Elastic metrics for the shape space and the shape and size space can be found in [4]. Furthermore, in [5], the authors propose a metric that distinguishes between the contribution of the difference in shape and the difference in size of two elements in the shape and size space.

On the other hand, elastic shape analysis of curves has been used in different data science problems, such as principal component analysis (PCA) [4, 6, 7], cluster analysis (CLA) [4, 8, 9], classification [4, 9, 10], and outlier detection [5, 6, 11, 12]. However, until now, elastic shape analysis has not been used in archetypal analysis.

Archetype analysis (AA) was defined by [13]. It is an exploratory unsupervised statistical learning technique [14] that lies between two well-known techniques, PCA and CLA [15]. The objective of AA is to express the data as a mixture (convex combination) of a set of archetypes. The archetypes are also a mixture of data points. Both these facts make the results of AA easy to understand, even for non-experts. Note that the interpretation of mixtures is simple, unlike the linear combinations of variables that form the factors of the PCA. Although the centers of CLA are also easy to understand, its modeling flexibility is diminished by the fact that data can only be assigned to one cluster, unlike AA. There is also another fact that stimulates the human comprehension of the results of AA versus other unsupervised techniques: archetypes are extreme or

pure profiles, and human beings understand them better than central points since we interpret opposite components better [16, 17]. Archetypes in Statistics have the same meaning as in everyday life [18].

Vinué [19] defined archetypoid analysis (ADA), which is a variant of AA. In ADA, the archetypal profiles are concrete observations. This is very useful for our case, as will be explained later since we are not dealing with multivariate data but with curves.

Visualization is an important task in exploratory analysis since it allows us to discover unrevealed aspects of the data. In our case, this is even more important since curves are complex data [11]. ADA not only allows us to display the main features of the data set by its extremes (archetypoids) but also allows us to explore data and extract information through the approximation of the data as a mixture of the archetypoids. ADA makes it possible to see and describe the whole data set through only a small set of representative observations that are easy to understand [18]. Note that considering extreme curves to display the main characteristics of a data set of curves was proposed by [20], who considered extreme principal component scores. However, the goal of PCA is not to find extreme cases as it is for ADA. Therefore, the cases with extreme PCA scores do not necessarily correspond to archetypal cases [13]. Even if all PCs were taken into account, archetypes could not be recovered [21].

AA and ADA are applied in many diverse fields, such as biology [22], computer vision [23–28], education [29, 30], engineering [31], genetics [32], machine learning problems [15], market research [17], neuroscience [33–35], psychology [36],and sports [37, 38]. Since the proposal by [21], AA has become a standard in the accommodation problem in industrial design [39], where extreme cases are searched to give designers an efficient way to develop and assess a product design. The designer considers a small set of boundary cases so that if the design fits well for those cases, it will also fit well for the not so extreme cases. However, the accommodation problem has not restricted to the multivariate case, but it has also been applied in other cases, such as shapes with landmarks [40, 41].

In this paper, we consider ADA when the metrics defined in [4] and [5] for the shape and size space are considered. This is the first time that ADA has been used in elastic shape analysis. ADA has been used before in shape analysis but with landmarks [40, 41], working in the tangent space. Our motivating problem is an accommodation problem with curves; therefore,we need to find archetypal curves. The main contributions of this work are as follows: to propose the first methodology to obtain archetypal curves in the shape and size space and to analyze their use with simulated data and to apply it to a real problem. Furthermore, the code is made available.

The outline of the paper is as follows. On the one hand, Sect. 2 reviews the SRVF representation of curves and the elastic metrics. On the other hand, several multivariate statistical methods are reviewed: a multidimensional scaling procedure and AA and ADA for the multivariate case. The proposed methodology for finding archetypal curves is introduced in Sect. 3. Sections 4 and 5 discuss the results when the new methodology is applied to a simulated data set and a real data set, respectively. Finally, some conclusions are given in Sect. 6.

2 Background

2.1 Elastic Metrics

For the study of shapes, or shapes and sizes, described by curves, the curve is usually represented in a specific metric space. Since a metric space is endowed with a distance function, when someone wishes to compare two shapes, or two shapes and sizes, of two domains bounded by their respective curves, it is necessary to compute the distance of the two points that represent such curves.

The classical metric spaces to represent curves are sub-spaces of the Hilbert space of functions $\mathbb{L}^2([0,1],\mathbb{R}^n)$, *i.e.*, the space of square integrable functions from $[0,1] \subset \mathbb{R}$ to $\mathbb{R}^n$. For example, in the square-root velocity function (SRVF) approach, every parameterized curve $\beta : [0,1] \to \mathbb{R}^n$ is represented by

$$\beta(t) \mapsto q(t) = \frac{\beta'(t)}{\sqrt{|\beta'(t)|}}.$$

It is easy to check that $q(t)$ associated with the curve $\beta : [0,1] \to \mathbb{R}^n$ belongs to $\mathbb{L}^2([0,1],\mathbb{R}^n)$ because

$$\|q\|^2_{\mathbb{L}^2} = \int_0^1 |q(t)|^2 dt = \text{length}(\beta) < \infty.$$

Moreover, it is interesting to remark here that the space of curves in $\mathbb{R}^n$ and the space of functions $\mathbb{L}^2([0,1],\mathbb{R}^n)$ can actually be identified because not only is each curve represented in the space $\mathbb{L}^2([0,1],\mathbb{R}^n)$ but every function $q(t) \in \mathbb{L}^2([0,1],\mathbb{R}^n)$ can be associated with the following curve

$$\beta(t) = \int_0^t q(s)|q(s)|ds.$$

By using this identification between the space of parameterized curves and the Hilbert space $\mathbb{L}^2([0,1],\mathbb{R}^n)$, we can use the distance $\mathrm{d}_{\mathbb{L}_2}$ in $\mathbb{L}^2([0,1],\mathbb{R}^n)$ as a distance between curves. Remember that given two points q_1 and q_2 in $\mathbb{L}^2([0,1],\mathbb{R}^n)$, since $\mathbb{L}^2([0,1],\mathbb{R}^n)$ is a Hilbert space, we can use its inner product to obtain the distance from q_1 to q_2 as

$$\begin{aligned}\mathrm{d}_{\mathbb{L}^2}(q_1,q_2) :=& \|q_2 - q_1\| = \sqrt{\langle q_2 - q_1, q_2 - q_1\rangle_{\mathbb{L}^2}} \\ =& \sqrt{\int_0^1 (q_2(t) - q_1(t))^2 dt}.\end{aligned}$$

When we think about a curve as a geometric object embedded or immersed in $\mathbb{R}^n$, we must conclude that the previous distance is not good enough to characterize the space of curves. This is mainly because this distance is not invariant under reparameterizations. Let us briefly recall the concept of reparameterization. Given a curve $\gamma : [0,1] \to \mathbb{R}^n$

and given a diffeormorphism $\phi : [0, 1] \to [0, 1]$ such that $\phi(0) = 0$ and $\phi(1) = 1$, we will say that the curve

$$\beta : [0, 1] \to \mathbb{R}^n, \quad t \mapsto \beta(t) := \gamma(\phi(t))$$

is the reparameterization by ϕ of the curve α. The set of reparameterizations with the composition law has a group structure, and we will denote it as Γ. A required condition for a distance to be useful in order to compare curves is, therefore, to be invariant under the action of the group of reparameterizations. These invariant metrics are called *elastic metrics*.

By using the identification of the space of curves with $\mathbb{L}^2([0, 1], \mathbb{R}^n)$, the action of the group Γ on the space of curves naturally induces an action on $\mathbb{L}^2([0, 1], \mathbb{R}^n)$ given by

$$\Gamma \times \mathbb{L}^2([0, 1], \mathbb{R}^n) \to \mathbb{L}^2([0, 1], \mathbb{R}^n), \quad (\phi, q) \mapsto \phi(q)(t) := q(\phi(t))\sqrt{\phi'(t)}.$$

Since two rotated curves represent the same bounded shape, we have to take care of the group of rotations as well. The group of rotations $SO(n)$ acts by matrix multiplication on the space of curves, and by again using the identification between the space of curves and $\mathbb{L}^2([0, 1], \mathbb{R}^n)$, there is a global action of the group $SO(n) \times \Gamma$ on $\mathbb{L}^2([0, 1], \mathbb{R}^n)$ given by

$$\begin{aligned} SO(n) \times \Gamma \times \mathbb{L}^2([0, 1], \mathbb{R}^n) &\to \mathbb{L}^2([0, 1], \mathbb{R}^n), \\ (O, \phi, q) &\mapsto (O, \phi)(q)(t) := O(q(\phi(t)))\sqrt{\phi'(t)}. \end{aligned}$$

Then, the space of interest, the shape and size space, will be represented as the orbit space

$$\mathcal{S} := \mathbb{L}^2([0, 1], \mathbb{R}^n)/SO(n) \times \Gamma.$$

In order to simplify the notation, let us denote this as $G := SO(n) \times \Gamma$. Then, given $q \in \mathbb{L}^2([0, 1], \mathbb{R}^n)$ the associated orbit $[q] \in \mathcal{S}$ will be obtained as

$$[q] := G.q = \{(O, \phi)(q) \, : \, (O, \phi) \in G\}.$$

Classically, see [4], the distance between orbits has been used to define an elastic metric in $\mathcal{S}$ as

$$\begin{aligned} \mathrm{d}_2([p], [q]) &:= \inf_{(O', \phi'), (O'', \phi'') \in G} \mathrm{d}_{\mathbb{L}^2}\left((O', \phi')p, (O'', \phi'')(q)\right) \\ &= \inf_{(O', \phi'), (O'', \phi'') \in G} \mathrm{d}_{\mathbb{L}^2}\left((O', \phi')\left(p, (O', \phi')^{-1}(O'', \phi'')(q)\right)\right) \\ &= \inf_{(O, \phi) \in G} \mathrm{d}_{\mathbb{L}^2}\left(p, (O, \phi)(q)\right). \end{aligned}$$

This is a distance in the shape and size space. In order to "take away" the length of the curves, normalized curves of length 1 can be used instead as

$$\mathrm{d}_4([p],[q]) := \inf_{(O,\phi)\in G} \cos^{-1}\left(\left\langle \frac{p}{\|p\|_{\mathbb{L}^2}}, (O,\phi)\left(\frac{q}{\|q\|_{\mathbb{L}^2}}\right)\right\rangle_{\mathbb{L}^2}\right).$$

In this paper, we focus on distances in the shape and size space as d_2. Recently, [5] proposed a new distance given by

$$\mathrm{d}_{4s}([p],[q]) := \sqrt{\mathrm{d}_4^2\left(\frac{p}{\|p\|_{\mathbb{L}^2}}, \frac{q}{\|q\|_{\mathbb{L}^2}}\right) + \ln^2\left(\frac{\|p\|_{\mathbb{L}^2}}{\|q\|_{\mathbb{L}^2}}\right)}.$$

This new distance is scale invariant in the sense that for any $\lambda \neq 0$

$$\mathrm{d}_{4s}([p],[q]) = \mathrm{d}_{4s}([\lambda p],[\lambda q]).$$

Moreover, in some cases, see Figures 6 and 7 of [5], for instance, when authors deal with a set of curves with a very large range of lengths, the new d_{4s} distance could improve the d_2 distance. Both distances are compared in clustering setting by [5].

2.2 Multidimensional Scaling

Let $\boldsymbol{D}$ be the $m \times m$ matrix containing the observed dissimilarity from the object p to the object q.

Let us recall that a distance matrix $\boldsymbol{D}$ is Euclidean if and only if $\boldsymbol{B}$ is positive semidefinite [42, Theorem 14.2.1], where $\boldsymbol{B} = (\boldsymbol{I} - m^{-1}\boldsymbol{e}\boldsymbol{e}')\boldsymbol{M}(\boldsymbol{I} - m^{-1}\boldsymbol{e}\boldsymbol{e}')$, $\boldsymbol{M}$ is a matrix with elements m_{pq} = -0.5* d_{pq}^2, $\boldsymbol{I}$ is the $m \times m$ identity matrix, and $\boldsymbol{e}$ is the $m \times 1$ vector with all its elements equal to unity.

If the distances are Euclidean distances, they can be represented exactly in at most $m-1$ dimensions [42, Theorem 14.4.1] by means of classical multidimensional scaling (cMDS) [43]. The objective of cMDS is to return a set of points such that the distances between them are approximately equal to the original distances since the dimension of the space that the data are to be projected in is usually less than $m - 1$. On the other hand, if the distances are not Euclidean, we can use cMDS as an approximation, which is optimal for a kind of discrepancy measure [42, Theorem 14.4.2]. However, it is possible to use h-plot, an alternative methodology proposed by [44, 45], which even works when the dissimilarity is not a distance. The aim of the h-plots is not to conserve the interpoint distances exactly but to preserve relationships between dissimilarity variables. This perspective is especially useful when the distance is not Euclidean, as in this case for d_2, d_4, d_{4s}, as the distances cannot be projected exactly in a Euclidean space. There are negative eigenvalues in the respective $\boldsymbol{B}$ matrices (see Sect. 4 and Sect. 5.1 for examples). Note that the original spaces are not Euclidean, so neither are the distances.

2.2.1 h-Plot for MDS

D is treated as a data matrix with h-plot, where each variable $d_{q.}$ measures the distance from q to other objects. (In case of asymmetrical relationships, variables measuring the distance from an object to q, $d_{.q}$, should be also considered). The variance–covariance matrix of **D**, **S**, is computed. **S** is always positive semidefinite and we solve the eigenvalue problem. Let λ_1 and λ_2 denote the two largest eigenvalues and q_1 and q_2 the corresponding unit eigenvectors. Then, the h-plot in two dimensions is $H_2 = (\sqrt{\lambda_1}q_1, \sqrt{\lambda_2}q_2)$. Analogously, it can be defined for higher dimensions.

The Euclidean distance between the rows h_p and h_q is approximately the sample standard deviation of the difference between variables $d_{p.}$ and $d_{q.}$. Therefore, if these variables are similar, their difference and, as a consequence, the standard deviation of their difference will be small and they will be represented near to each other and vice versa. If the scale of the distances is linearly modified, the obtained configuration does not change, only the scale of the axes is modified. The goodness of fit can be easily assessed by $(\lambda_1^2 + \lambda_2^2)/\sum_j \lambda_j^2$, where a high measure, close to 1, indicates a better fit. H-plots were compared with eleven methods by [44], with very good performance, even for asymmetrical relationships [45].

2.2.2 Congruence Coefficient

The best method to asses configurations is pictures [46, sec. 19.7]. However, we can use the congruence coefficient (CC), a correlation coefficient about the origin, to approximately assess the configurational similarity of two configurations C_1 and C_2. In configuration C_1 (C_2), the dissimilarity between ith and jth objects is $d_{ij}(C_1)$ ($d_{ij}(C_2)$).

CC is defined for symmetric dissimilarity matrices:

$$CC = \frac{\sum_{i<j} d_{ij}(C_1)d_{ij}(C_2)}{(\sum_{i<j} d_{ij}^2(C_1))^{1/2}(\sum_{i<j} d_{ij}^2(C_2))^{1/2}}$$

CC ranges from 0 to 1. If C_1 and C_2 are perfectly similar geometrically, CC = 1. We say that two configurations are similar when they can be brought to a complete match by rigid motions and dilations.

In the experimental sections, the configuration C_1 provided by distances d_2 and d_{4s} is compared with the configuration C_2 obtained after projecting by h-plot, using the Euclidean distance for computing the interpoint distances.

The idea of assessing the goodness of approximations by means of correlation between distances has been also used elsewhere. For example, [47] used the correlation between the Procrustes distances and the Euclidean distances in the tangent space in shape statistics with landmarks.

2.3 Archetypal Analysis

Let us review ADA and AA in the multivariate case. Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_m)$ be an $m \times r$ data matrix with m observations and r variables.

In AA, we search for k archetypes, which are mixtures of observations, i.e., a $k \times r$ matrix $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_k)$, whose mixture approximates each row $\mathbf{x}_i$

$$\mathbf{x}_i \sim \hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j. \tag{1}$$

The $m \times k$ matrix $\alpha = (\alpha_{ij})$ contains the approximator mixture coefficients, while the $k \times m$ matrix $\beta = (\beta_{jl})$ contains the constructor mixture coefficients, i.e.,they build the archetypes according to:

$$\mathbf{z}_j = \sum_{l=1}^{m} \beta_{jl} \mathbf{x}_l. \tag{2}$$

To find the matrices α and β and, therefore, $\mathbf{Z}$, we should minimize the following residual sum of squares (RSS), where ($\| \cdot \|$ denotes the Frobenius matrix norm for matrices and the Euclidean norm for vectors):

$$RSS = \|\mathbf{X} - \alpha\beta\mathbf{X}\|^2 = \sum_{i=1}^{m} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j \right\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{x}_l \right\|^2, \tag{3}$$

under the restrictions

(1) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ and $i = 1, \ldots, m$, $j = 1, \ldots, k$ and

(2) $\sum_{l=1}^{m} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ and $j = 1, \ldots, k$ and $l = 1, \ldots, m$.

In ADA, we search for k archetypoids, which are actual observations of the data set. Therefore, the minimization problem is similar to that of AA but restriction 2) is replaced by:

2) $\sum_{l=1}^{m} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, \ldots, k$ and $l = 1, \ldots, m$. In this way, $\beta_{jl} = 1$ for one and only one l, otherwise $\beta_{jl} = 0$.

In ADA, as in AA, each α_{ij} returns the weight of the archetypoid $\mathbf{z}_j$ for the observation $\mathbf{x}_i$; that is to say, the α approximator coefficients indicate how much each archetypoid contributes to the approximation of each observation.

Archetypes are located on the boundary of the convex hull if $k > 1$, while it is the mean if $k = 1$ [13]. Archetypoids are not necessarily on the boundary of the convex hull

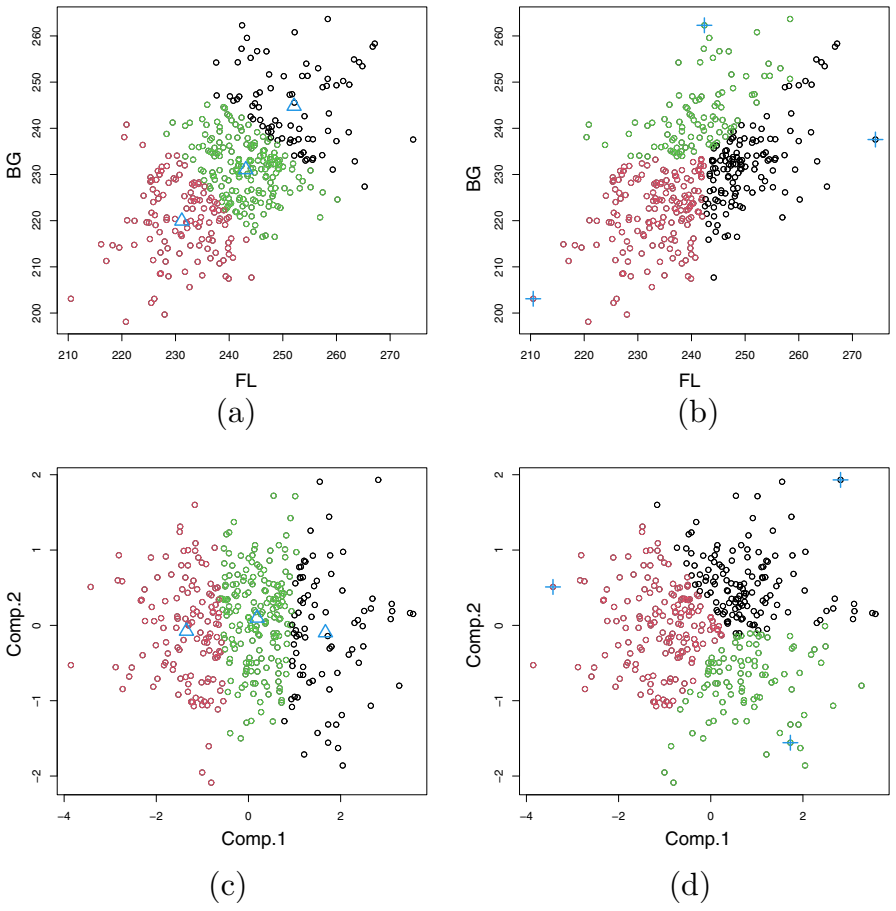

Fig. 1 Results for FL and BG. **a** k-means with cluster assignments to each centroid represented by blue triangles. **b** ADA with assignments to the maximum alpha. Archetypoids are represented by blue crosses. **c** PC projected k-means results. **d** PC projected ADA results

if $k > 1$, and it is the medoid if $k = 1$ [19]. The medoid is the observation for which the average dissimilarity between it and all the other observations is minimal. Therefore, it is the most centrally located observation. Note that medoids are the elements of the data set.

In this paper, we will focus on ADA, as explained in Sect. 3.

2.3.1 Toy Example

We use a two-dimensional data set to clarify the meaning of ADA and their differences with PCA and CLA. We consider two variables of 381 right feet of Spanish women: the Foot Length (FL) and Ball Girth (BG). Details about the data set and variables are provided in Sect. 5. PCA, k-means and ADA with $k = 3$ are applied to standardized data. Figure 1 shows the results.

Archetypoids are extreme feet. The first archetypoid has very low FL and BG values, the second archetypoid is characterized by a very high value for BG, but a medium value for FL, while archetypoid 3 is characterized by a very high FL value, but a medium–high value for BG. The rest of the feet are explained by mixtures of these archetypal feet. For instance, a foot with values of 254 and 254 for FL and BG, respectively, is described by 64% of archetypoid 2 plus 36% of archetypoid 3. k-means does not return this kind of information, it only indicates the cluster assignments.

Centroids of k-means are in the middle of the data cloud, with less extreme dimensions than archetypoids. Therefore, archetypoids are more easily interpretable. Centroids have more uniform shapes, and they have the same FL and BG ratio as the mean foot. This does not happens with archetypoids. We can visualize this in the PC projections. The first PC is a size component, while the second PC is a shape component, since the loadings are 0.7 and 0.7 for the 1st PC and 0.7 and -0.7 for the 2nd PC. Centroids are found in the zero horizontal line, while archetypoids are found near the border of the PC score space, although they are not the cases with the most extreme PC scores.

3 Methodology

In our problem, we do not have variables in $\mathbb{R}^r$, but we have the distances between the curves. When variables are unavailable, we can follow the strategy explained by [19] for finding archetypoids. The idea is to project the distances into a certain space $\mathbb{R}^r$ and find the archetypoids in that space. Note that by using archetypoids, as they are actual observations, we can determine the concrete curves in the original space. In this way, we can also visualize the archetypal curves. This would not be possible with archetypes since we cannot obtain a mixture of curves. However, we have the α coefficients, expressing the contribution of each original archetypoid to each original curve. The idea of projecting to an approximating linear space, when the original space is not vectorial, and working on that space, is widely used in Statistics [48].

The scheme of the procedure is as follows.

1. Compute $\mathbf{D}$, which is the $m \times m$ matrix where d_{pq} denotes the distance between the curves $[p]$ and $[q]$.

2. Use a multidimensional scaling method (MDS) to find a representation in $\mathbb{R}^r$ that conserves the pairwise distances, i.e., the information contained in $\mathbf{D}$, in some way. According to the method, a goodness of fit measure can be used to select r.

3. Calculate the archetypoids of the $m \times r$ matrix $\mathbf{X}$, the matrix obtained by the MDS method. This matrix contains the coordinates of the points estimated to represent the distances.

Regarding to $\mathbf{D}$, d_2 or d_{4s} are used in our case. As regards the MDS method, we consider h-plot in this work. It was used previously with good results in ADA when variables are not available [18, 19, 49, 50]. We want to emphasize that the scheme is flexible. Other choices can be selected for estimating $\mathbf{D}$ and for MDS.

3.1 Computational Details

As regards the implementation of the methods, the code is available in Section Code Availability. For computing d_2 and d_{4s} distances, we use the implementation by [4] and [5], respectively. [5] also used the code by [4] for computing d_4. For obtaining the infimum over orbits, two optimization methods are applied, Procrustes analysis and the dynamic programming algorithm. Full details are provided in [4, Appendix 2].

H-plot implementation is made as in [44], with *princomp* function for PCA in R.

To solve the mixed-integer optimization problem of ADA, [19] proposed an algorithm based on two phases: an initialization phase, called the BUILD phase, where a set of possible archetypoids are selected, and the SWAP phase, where the initial set is improved by exchanging the selected observations for unselected ones and checking if these replacements decrease the RSS. We use the R [51] implementation created by [40].

As regards the determination of the number of archetypoids, we use the elbow criterion, which has been used in previous papers, such as [13, 19, 52]. This criterion consists of displaying the RSS versus the number of archetypoids and determining the point where an elbow is found.

3.1.1 Scalability

Regarding scalability, there are two issues to consider. On the one hand, the number of sample points per curve is not a problem since the algorithm by [4] resamples the curves at the beginning to have 100 points, so the number of sample points is always constant. The number of sample points in curves is only used in the estimation of the distances.

On the other hand, let us analyze the scalability of our procedure when the number of curves is big. Let us analyze each part of the procedure. Firstly, the computation of the distances is made by each pair of curves. Therefore, it can be easily parallelized. Secondly, the h-plot method depends on the solution of an eigenvalue problem of a positive semidefinite matrix, which is a well-studied problem for large matrices [53]. Nowadays, there are even scalable methods for computing eigenvectors of non-symmetric matrices [54]. Thirdly, ADA method was made scalable by [55].

4 Application to a Simulated Data Set

We have simulated an artificial data set with 90 3D cylindric helixes, $\beta_i(t)$ $i = 1, \cdots, 90$, $t \in [0, 1]$, with

$$x_i = a_i \cos(8\pi t); \quad y_i = a_i \sin(8\pi t); \quad z_i = b_i t; \quad i = 1, \cdots, 90$$

where the parameters a_i and b_i of the helix are randomly obtained from two different probability distributions; the radius $a_i \sim Normal(50, 20)$ and $b_i \sim Uniform(30, 70)$ (so all these helixes will have different shapes and different lengths).

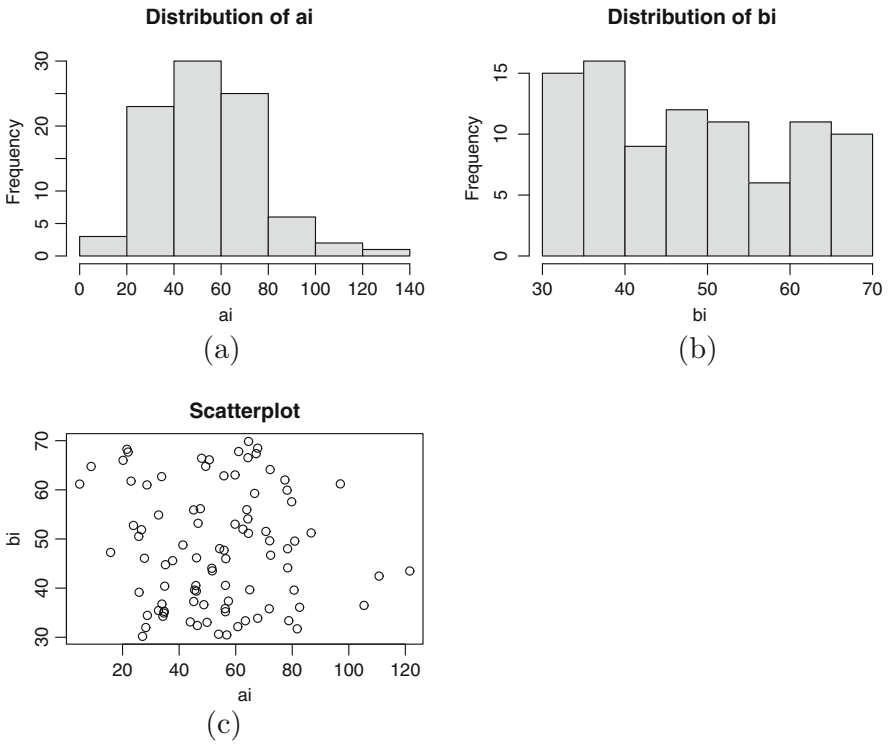

Fig. 2 Parameters of the simulated helixes. **a** Values simulated for $a_i, i = 1, \cdots, 90$ **b** Values simulated for $b_i, i = 1, \cdots, 90$. **c** Scatter-plot of the values simulated for $(a_i, b_i), i = 1, \cdots, 90$

Fig. 2 shows the values obtained in the simulations for the parameters of the 90 helixes. Fig. 3a and b show two graphical representations of the simulated helixes.

The distance matrices between the 90 curves, D_{4s} and D_2, have been computed. As these distance matrices are not Euclidean (there are negative eigenvalues in $\boldsymbol{B}$ for both distance matrices, see Fig. 4), h-plots can be used as MDS. The goodness-of-fit measures of h-plots explained by [44] for d_{4s} are 87.94%, 99.92%, and 99.99% for $r = 1, 2, 3$, respectively, and 80.31%, 99.92%, and 99.97% for d_2. We use $r = 3$ in both cases. The resulting h-plots for two dimensions ($r = 2$) can be seen in Fig. 5. Their CC are 95% and 97%, respectively.

According to the elbow criteria, the screeplots shown in fig. 6 advises us to consider three ($k = 3$) archetypoids in each case.

The archetypoids obtained with the two distances are somewhat different (Figs. 7 and 8). The values of a_i in the helixes of the data set range from 4.82 to 121.57, and the values of b_i range between 30.18 and 79. Figure 8 shows the parameters of the 90 helixes together with the parameters of the archetypoids obtained with the two distances. The first archetypoid obtained from d_{4s} has a very low value for a_i and a very high value for b_i, while the first archetypoid from d_2 has also a low value for a_i and quite a high value for b_i, slightly lower than that obtained with d_{4s}. The second and third archetypoids from d_{4s} have low values for b_i, and intermediate and high

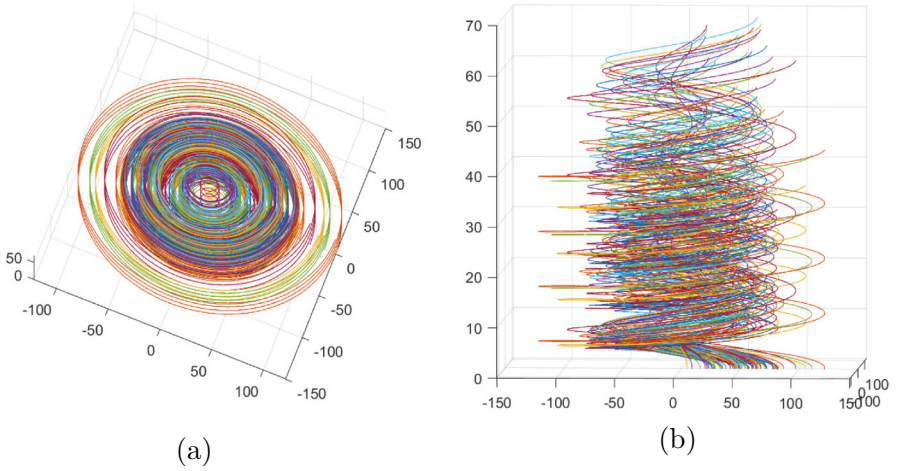

Fig. 3 Simulated curves. **a** and **b** show the 90 simulated helixes seen from different perspectives

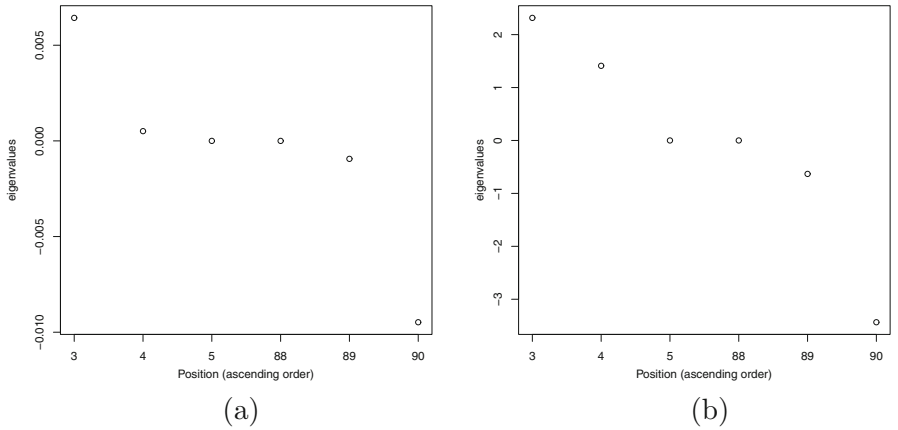

Fig. 4 3rd to 5th and 88th to 90th eigenvalues in $\boldsymbol{B}$. **a** Using the distance d_{4s}. **b** Using the distance d_2

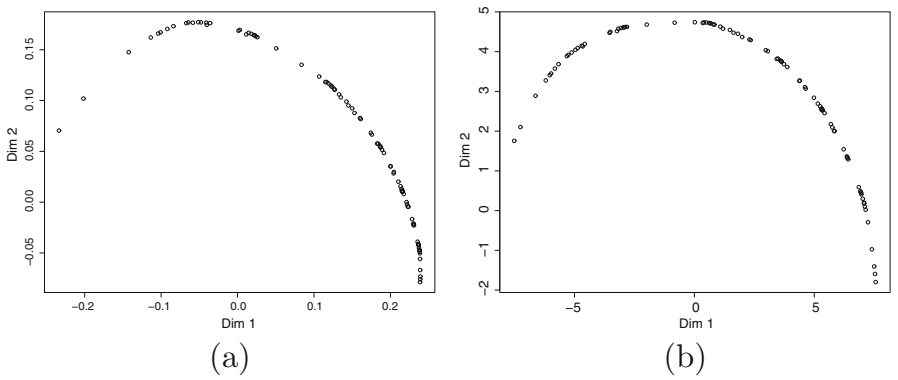

Fig. 5 H-plots of simulated curves. **a** Using the distance d_{4s}. **b** Using the distance d_2

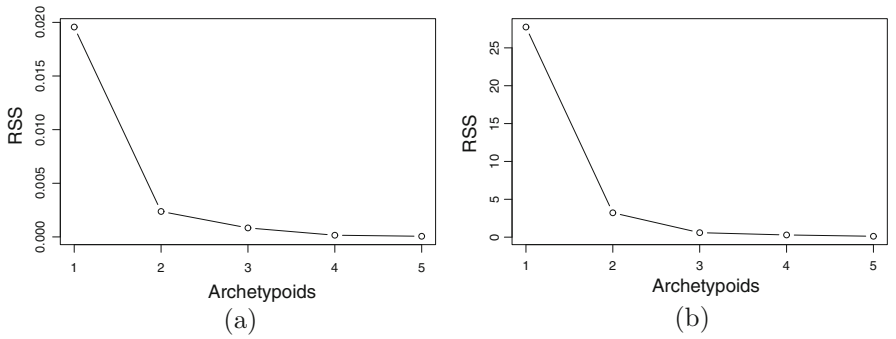

Fig. 6 Screeplot for simulated data. **a** Using the distance d_{4s}. **b** Using the distance d_2

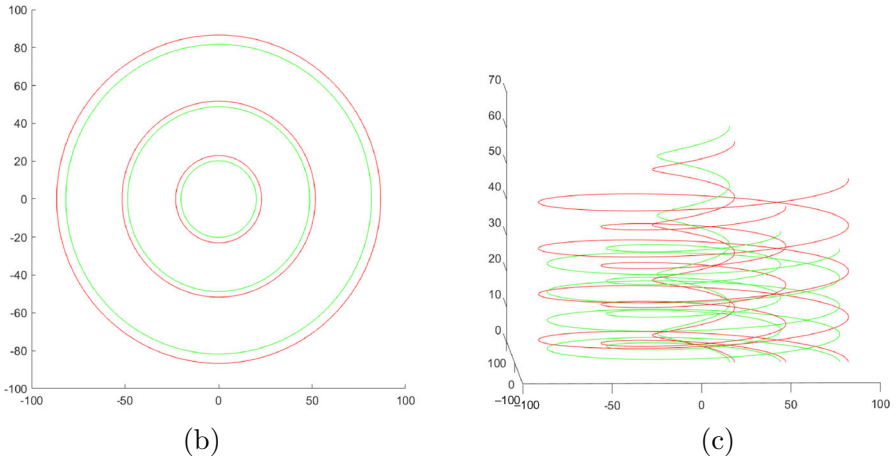

Fig. 7 Archetypoids obtained seen from two different perspectives. In green, the archetypoids obtained with d_{4s}, and in red, those obtained with d_2

values for a_i, respectively. However, the second and third archetypoids from d_2 also have intermediate and high values for a_i, respectively, but medium values for b_i. In summary, the parameters of the archetypoids found by d_{4s} are more extreme than the parameters of the archetypoids found by d_2.

5 Application to a Real Data Set

Suitable footwear design needs to take into account the distribution of foot shape [41]. If this is not taken into account, it will not only lead to lower sales, but can also cause pain and deformity, especially in women. This is the reason why there are many studies on foot shape, such as [56–61], etc.

Comprehension of the typology and distribution of body part shapes is not only critical in the apparel industry but also in ergonomic industrial design [62, 63], as well as in other scientific disciplines that include criminalistics [64], face classification with

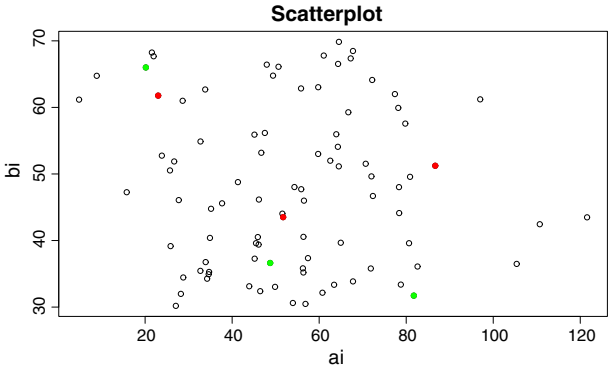

Fig. 8 Parameters of the archetypoids (of the helixes) obtained from the two distances. In green, the parameters (a_i, b_i) of the archetypoids obtained with d_{4s}, and in red the parameters (a_i, b_i) of the archetypoids obtained with d_2

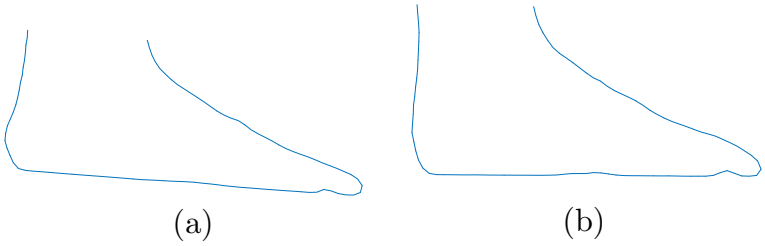

Fig. 9 Medoid curves of feet for men (**a**) and women (**b**) with d_{4s}

all its fields of application (forensic anthropology, crime prevention, human-machine interaction systems like e-commerce, e-learning, games, dating, and social networks) [65, 66], medicine [67–69], phylogeny [70], sport [71–73], etc. However, it is not just restricted to anthropometry; taxonomy is also important in morphometry in general, such as in plant or animal taxonomy [74, 75] and also in genetics [76].

ADA with landmarks was used in [41] for determining foot type in the adult Spanish population. Here we have carried out a similar study, but instead of using landmarks, we use curves.

Here, we use the data from [5]. The description of the acquisition of these data is detailed in [5]. Our curves consist of the longitudinal contour of right feet passing through the Ball Position. The sample size is 770, divided into 389 and 381 right feet of Spanish adult men and women, respectively.

The medoid shapes for men and women with d_{4s} are displayed in Figure 9.

5.1 Results and Discussion

In the interests of brevity and as an illustrative example, we only examine the results for d_{4s} although they could be carried out with d_2, analogously. The matrices $\boldsymbol{D}$ obtained with d_{4s} for men and women are not Euclidean, since the respective $\boldsymbol{B}$ are not positive

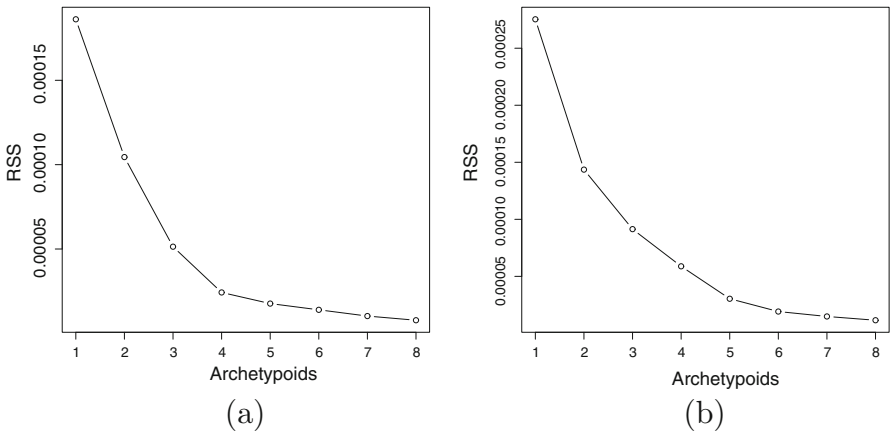

Fig. 10 Screeplot for women (**a**) and men (**b**)

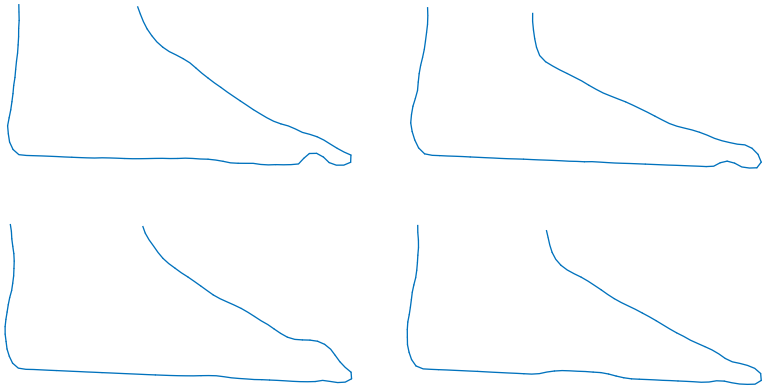

Fig. 11 Archetypoidal feet for women

semidefinite; 41% of the eigenvalues are negative. Therefore, we show the results using h-plot as MDS.

The goodness-of-fit measure for h-plotting (see [44] for details) is 99% for $r = 4$ for both men and women (it is 84.96%, 95.44%, 97.77%, and 99.43% for $r = 1, 2, 3,$ and 4, respectively, for men, while it is 81.89%, 91.66%, 96.92%, and 98.83% for $r = 1, 2, 3,$ and 4, respectively, for women). Therefore, we use $r = 4$. The CC are 97% and 96% for men and women, respectively.

Figure 10 shows the screeplot for women and men. The elbow is found at $k = 4$ and $k = 5$, for women and men, respectively. The archetypal feet are displayed in Figure 11 and Figure 12 for women and men, respectively.

In order to describe the archetypoidal curves obtained, Tables 1 and 2 display the percentiles of the four variables that most influence shoe fit according to footwear design experts. The variables are Foot Length, FL (distance between the rear and foremost point of the foot axis); Ball Girth, BG (perimeter of the ball section); Ball Width, BW (maximal distance between the extreme points of the ball section projected

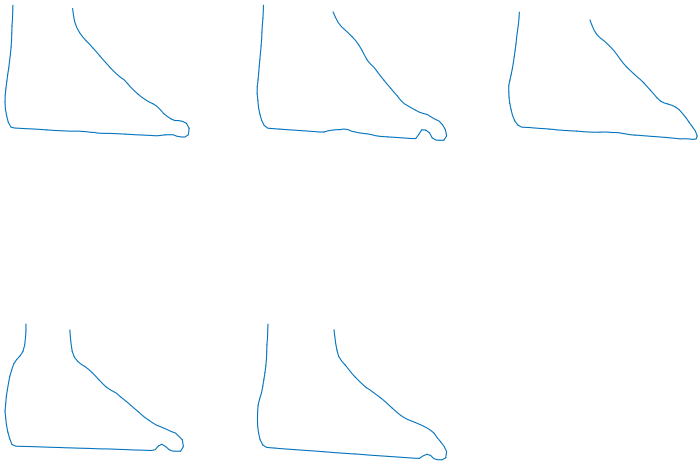

Fig. 12 Archetypoidal feet for men

Table 1 Percentiles of the main variables and divided by FL for archetypoidal feet of women

Archetypoid	FL	BG	BW	IH	BG/FL	BW/FL	IH/FL
1st	82	59	54	87	32	30	71
2nd	66	14	37	3	7	24	1
3rd	19	17	20	56	43	44	72
4th	13	65	55	62	94	90	85

Table 2 Percentiles of the main variables and divided by FL for archetypoidal feet of men

Archetypoid	FL	BG	BW	IH	BG/FL	BW/FL	IH/FL
1st	94	69	61	53	9	9	16
2nd	50	36	21	96	37	18	92
3rd	25	95	87	93	99	98	95
4th	63	17	12	7	7	4	5
5th	95	95	95	72	53	62	35

onto the ground plane); and Instep Height, IH (maximal height of the instep section, located at 50% of the foot length). We also show the percentiles of the variables after removing the scale, i.e.,by dividing each of the variables by FL: BG/FL, BW/FL, and IH/FL.

In the case of women, the first archetypoidal foot has high percentiles for FL and IH and medium percentiles for BG and BW; the second archetypoidal foot has low percentiles for BG and IH, and medium for FL and BW; the third archetypoid has low percentiles for FL, BG,and BW, and medium for IH; and the fourth archetypoid has a low percentile for FL and medium for BG, BW,and IH. This last archetypoidal

Table 3 Distribution of feet for women and men

	1st Arch.	2nd Arch.	3rd Arch.	4th Arch.	5th Arch.
Women	92	109	79	101	
Men	121	77	62	32	97

Arch. stands for archetypoid

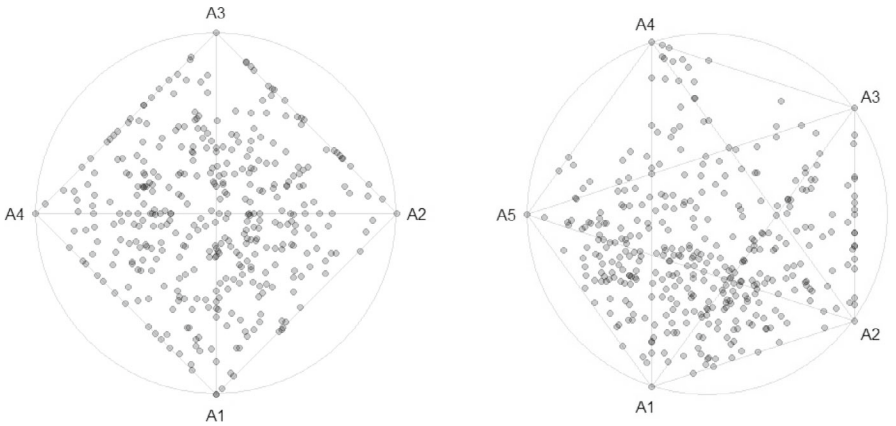

Fig. 13 Simplexplot for women (**a**) and men (**b**)

foot has an extreme shape, since it has very high percentiles (around 90) for variables BG/FL, BW/FL, and IH/FL.

In the case of men, the first archetypoidal foot has a high percentile for FL and medium percentiles for the rest of the variables (BG, BW, and IH), although the percentiles are low for the variables divided by FL; the second archetypoidal foot has a high percentile for IH; the third archetypoid has high percentiles for BG, BW, and IH; the fourth archetypoid has low percentiles for BG, BW, and IH; and the fifth archetypoidal foot has high percentiles for all four variables FL, BG, BW, and IH.

Let us see how the feet are distributed according to the archetypoidal curves. Table 3 shows the distribution of archetypoidal profiles for women and men. For each foot, we consider its α coefficients, and we assign each foot to the archetypoidal profiles for which the α coefficient is maximum. For example, the 4th archetypoidal profile for men is not very prevalent. A simplex visualization of the α coefficients is shown in Fig. 13, using the *simplexplot* function of the R package *archetypes* [77]. In this way, we have been able to visualize the set of curves of the feet and express them as a mixture of the archetypoids.

We have applied multivariate ADA with standardized features FL, BG, BW,and IH, for men and women. We have considered the same number of archetypoids as with curves to check if the same results could have been achieved using the features directly instead of the curves. Table 4 shows the percentiles of the multivariate archetypoids for women and men, respectively.

Table 4 Percentiles of the main variables and divided by FL for archetypoidal feet of women and men with multivariate features

Archetypoid	FL	BG	BW	IH	BG/FL	BW/FL	IH/FL
A1W	99	40	59	82	1	4	32
A2W	5	1	0	0	8	3	2
A3W	96	100	100	64	93	97	27
A4W	17	83	82	100	96	95	100
A1M	10	86	93	24	99	100	59
A2M	0	0	0	4	41	5	66
A3M	53	80	80	100	84	82	100
A4M	78	26	50	2	5	21	1
A5M	100	99	99	98	28	37	53

The first four archetypoids correspond to the sample of women. They are denoted by A1W, ..., A4W. The last five archetypoids correspond to the sample of men. They are denoted by A1M, ... A5M

The profiles returned by multivariate features and curves are somewhat different. In the case of women, the fourth archetypoid profiles are the most similar with no large differences in features FL, BG/FL, BW/FL, and IH/FL. There are some coincidences in the first archetypoid profiles, with some not too large differences in features FL, BW, and IH. This also happens with the second archetypoid profiles, with some not too large differences in features BG, IH, BG/FL, and IH/FL. The third archetypoid profiles have the largest differences. These differences are found in all the features except IH.

As regards men, the third, fourth, and fifth archetypoid profiles are the most similar with no large differences in features BG, BW, IH, BG/FL, BW/FL, and IH/FL; FL, BG, IH, BG/FL, and IH/FL; and FL, BG, BW, and IH/FL, respectively. The largest differences are provided between the profiles of the first and second archetypoids. They only coincide in feature BG for the first archetypoid and features BG/FL and BW/FL for the second archetypoid. Therefore, the archetypal profiles returned using the richer information of curves cannot be retrieved using multivariate data. The same occurs in [41], where results with 3D landmarks and multivariate data were also compared.

6 Conclusion

We have used archetypal analysis for the first time in elastic shape analysis. We have applied ADA to the projections with MDS of two distances (d_2 and d_{4s}). ADA has allowed us to see which the archetypal curves were and relate the rest of the curves to said archetypoids by the α coefficients. As curves are complex data, exploration and visualization of the data set are simplified by ADA. We have seen the application in a real problem concerning footwear. Furthermore, our proposal is scalable.

If we wanted to find the archetypal curves and the data set has different groups, the same idea as in [20] could be considered: using a clustering algorithm to find the groups and then applying ADA to find the archetypal curves of each group.

In future work, ADA could be replaced in our methodology by robust ADA [78] when dealing with outliers. Furthermore, our methodology could be extended to irregular or sparsely sampled curves [79]. A new line of research could involve using the distances and ADA in a different data science problem, such as the detection of outlier curves by extending the idea proposed by [80]. Furthermore, the fields of application are numerous, from medicine [81] to industry [82] or computer animation [83].

Acknowledgements The authors would like to thank Sebastian Kurtek for providing us with the code and the "Biomechanics Institute of Valencia" for providing us with the foot database.

Author Contributions Conceptualization, I.E.; data curation, I.E. and M.V.I.; formal analysis, I.E., V.G. and X.G-A.; funding acquisition, I.E. and V.G.; investigation, I.E., M.V.I., V.G. and X.G-A.; methodology, I.E., M.V.I., V.G. and X.G-A.; project administration, I.E.; resources, I.E., M.V.I, V.G. and X.G-A.; software, I.E. and M.V.I.; supervision, I.E.; validation, I.E., M.V.I., V.G. and X.G-A.; visualization, I.E., M.V.I., V.G. and X.G-A.; writing-original draft, I.E., M.V.I., V.G. and X.G-A.; writing-review and editing,I.E., M.V.I., V.G. and X.G-A. All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Work partially supported by the Research grant PID2022-141699NB-I00, PID2020-118763GA-I00,and PID2020-115930GA-I00 funded by Ministerio de Ciencia e Innovación, AICO/2021/252 from Generalitat Valenciana and UJI-B2020-22 from Universitat Jaume I, Spain. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Availability of Data The code and data for reproducing the results are publicly available at http://www3.uji.es/~epifanio/RESEARCH/ada_curves.zip.

Code Availability The code and data for reproducing the results are publicly available at http://www3.uji.es/~epifanio/RESEARCH/ada_curves.zip.

Declarations

Conflict of interest The authors have declared that no competing interests exist.

References

1. Kendall, D.G.: Shape manifolds, procrustean metrics, and complex projective spaces. Lond. Math. Soc. **16**, 81–121 (1984)
2. Younes, L., Michor, P.W., Shah, J., Mumford, D.: A metric on shape space with explicit geodesics. Rend. Lincei Mat. Appl. **19**, 25–57 (2008)
3. Srivastava, A., Klassen, E., Joshi, S.H., Jermyn, I.H.: Shape analysis of elastic curves in Euclidean spaces. IEEE Trans. Pattern Anal. Mach. Intell. **33**(7), 1415–1428 (2011)
4. Kurtek, S., Srivastava, A., Klassen, E., Ding, Z.: Statistical modeling of curves using shapes and related features. J. Am. Stat. Assoc. **107**(499), 1152–1165 (2012)

5. Epifanio, I., Gimeno, V., Gual-Arnau, X., Ibáñez-Gual, M.V.: A new geometric metric in the shape and size space of curves in $\mathbb{R}^n$. Mathematics **8**(10) (2020)
6. Kurtek, S., Su, J., Grimm, C., Vaughan, M., Sowell, R., Srivastava, A.: Statistical analysis of manual segmentations of structures in medical images. Comput. Vis. Image Underst. **117**(9), 1036–1050 (2013)
7. Cho, M.H., Asiaee, A., Kurtek, S.: Elastic statistical shape analysis of biological structures with case studies: a tutorial. Bull. Math. Biol. **81**(7), 2052–2073 (2019)
8. Xie, Q., Kurtek, S., Srivastava, A.: Analysis of AneuRisk65 data: elastic shape registration of curves. Electron. J. Stat. **8**(2), 1920–1929 (2014)
9. Epifanio, I., Gual-Arnau, X., Herold-Garcia, S.: Morphological analysis of cells by means of an elastic metric in the shape space. Image Anal. Stereol. **39**(1) (2020)
10. Laga, H., Kurtek, S., Srivastava, A., Golzarian, M., Miklavcic, S.J.: A Riemannian elastic metric for shape-based plant leaf classification. In: 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA), pp. 1–7 (2012)
11. Xie, W., Chkrebtii, O., Kurtek, S.: Visualization and outlier detection for multivariate elastic curve data. IEEE Trans. Vis. Comput. Gr. **26**(11), 3353–3364 (2020)
12. Harris, T., Tucker, J.D., Li, B., Shand, L.: Elastic depths for detecting shape anomalies in functional data. Technometrics **63**(4), 466–476 (2021)
13. Cutler, A., Breiman, L.: Archetypal analysis. Technometrics **36**(4), 338–347 (1994)
14. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference and Prediction. 2nd edn., Springer, New York (2009)
15. Mørup, M., Hansen, L.K.: Archetypal analysis for machine learning and data mining. Neurocomputing **80**, 54–63 (2012)
16. Thurau, C., Kersting, K., Wahabzada, M., Bauckhage, C.: Descriptive matrix factorization for sustainability: adopting the principle of opposites. Data Min. Knowl. Discov. **24**(2), 325–354 (2012)
17. Porzio, G.C., Ragozini, G., Vistocco, D.: On the use of archetypes as benchmarks. Appl. Stoch. Models Bus. Ind. **24**, 419–437 (2008)
18. Epifanio, I., Ibáñez, M.V., Simó, A.: Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. Am. Stat. **74**(2), 169–183 (2020)
19. Vinué, G., Epifanio, I., Alemany, S.: Archetypoids: a new approach to define representative archetypal data. Comput. Stat. Data Anal. **87**, 102–115 (2015)
20. Jones, M.C., Rice, J.A.: Displaying the important features of large collections of similar curves. Am. Stat. **46**(2), 140–145 (1992)
21. Epifanio, I., Vinué, G., Alemany, S.: Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem. Comput. Ind. Eng. **64**(3), 757–765 (2013)
22. D'Esposito, M.R., Palumbo, F., Ragozini, G.: Interval Archetypes: A New Tool for Interval Data Analysis. Stat. Anal. Data Min. **5**(4), 322–335 (2012)
23. Chen, Y., Mairal, J., Harchaoui, Z.: Fast and robust archetypal analysis for representation learning. In: CVPR 2014—IEEE Conference on Computer Vision & Pattern Recognition, pp. 1478–1485 (2014)
24. Bauckhage, C., Kersting, K., Hoppe, F., Thurau, C.: Archetypal analysis as an autoencoder. In: Workshop New Challenges in Neural Computation (2015)
25. Sun, W., Yang, G., Wu, K., Li, W., Zhang, D.: Pure endmember extraction using robust kernel archetypoid analysis for hyperspectral imagery. ISPRS J. Photogramm. Remote Sens. **131**, 147–159 (2017). https://doi.org/10.1016/j.isprsjprs.2017.08.001
26. Sun, W., Zhang, D., Xu, Y., Tian, L., Yang, G., Li, W.: A probabilistic weighted archetypal analysis method with Earth mover's distance for endmember extraction from hyperspectral imagery. Remote Sens. **9**(8), 841 (2017)
27. Mair, S., Boubekki, A., Brefeld, U.: Frame-based data factorizations. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 2305–2313. PMLR, International Convention Centre, Sydney, Australia (2017)
28. Cabero, I., Epifanio, I.: Archetypal analysis: an alternative to clustering for unsupervised texture segmentation. Image Anal. Stereol. **38**(2), 151–160 (2019)
29. Cabero, I., Epifanio, I.: Finding archetypal patterns for binary questionnaires. SORT **44**(1), 39–66 (2020)
30. Cabero, I., Epifanio, I., Gual-Arnau, X.: Analysis of archetypes to determine time use and workload profiles of Spanish university professors. Educ. Sci. **13**(3) (2023)
31. Millán-Roures, L., Epifanio, I., Martínez, V.: Detection of anomalies in water networks by functional data analysis. Math. Probl. Eng. 2018 (Article ID 5129735), 13 (2018)

32. Thøgersen, J.C., Mørup, M., Damkiær, S., Molin, S., Jelsbak, L.: Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. BMC Bioinform. **14**, 279 (2013)
33. Tsanousa, A., Laskaris, N., Angelis, L.: A novel single-trial methodology for studying brain response variability based on archetypal analysis. Expert Syst. Appl. **42**(22), 8454–8462 (2015)
34. Hinrich, J.L., Bardenfleth, S.E., Roge, R.E., Churchill, N.W., Madsen, K.H., Mørup, M.: Archetypal analysis for modeling multisubject fMRI data. IEEE J. Sel. Top. Signal Process. **10**(7), 1160–1171 (2016)
35. Olsen, A.S., Høegh, R.M.T., Hinrich, J.L., Madsen, K.H., Mørup, M.: Combining electro- and magnetoencephalography data using directional archetypal analysis. Front. Neurosci. **16** (2022). https://doi.org/10.3389/fnins.2022.911034
36. Fernández, D., Epifanio, I., McMillan, L.F.: Archetypal analysis for ordinal data. Inf. Sci. **579**, 281–292 (2021)
37. Eugster, M.J.A.: Performance profiles based on archetypal athletes. Int. J. Perform. Anal. Sport **12**(1), 166–187 (2012)
38. Vinué, G., Epifanio, I.: Forecasting basketball players' performance using sparse functional data. Stat. Anal. Data Min.: ASA Data Sci. J. **12**(6), 534–547 (2019)
39. Rodríguez Vega, G., Zaldívar Colado, U., Zaldívar Colado, X.P., Rodríguez Vega, D.A., de la Vega Bustillos, E.J.: Comparison of univariate and multivariate anthropometric accommodation of the northwest mexico population. Ergonomics **64**(8), 1018–1034 (2021)
40. Epifanio, I., Ibáñez, M.V., Simó, A.: Archetypal shapes based on landmarks and extension to handle missing data. Adv. Data Anal. Classif. **12**(3), 705–735 (2018)
41. Alcacer, A., Epifanio, I., Ibáñez, M.V., Simó, A., Ballester, A.: A data-driven classification of 3d foot types by archetypal shapes based on landmarks. PLoS ONE **15**, 1–19 (2020)
42. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Academic Press, London (1979)
43. Cox, T.F., Cox, M.A.: Multidimensional Scaling. CRC Press, Boca Raton (2000)
44. Epifanio, I.: h-plots for displaying nonmetric dissimilarity matrices. Stat. Anal. Data Min. **6**(2), 136–143 (2013)
45. Epifanio, I.: Mapping the asymmetrical citation relationships between journals by h-plots. J. Assoc. Inf. Sci. Technol. **65**(6), 1293–1298 (2014)
46. Borg, I., Groenen, P.: Modern Multidimensional Scaling Theory and Applications. Springer, New York (1997)
47. Rohlf, F.J.: Shape statistics: procrustes superimpositions and tangent spaces. J. Classif. **16**(2), 197–223 (1999)
48. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis: With Applications in R. John Wiley & Sons, Chichester (2016)
49. Vinué, G., Epifanio, I.: Archetypoid analysis for sports analytics. Data Min. Knowl. Discov. **31**(6), 1643–1677 (2017)
50. Alcacer, A., Epifanio, I., Valero, J., Ballester, A.: Combining classification and user-based collaborative filtering for matching footwear size. Mathematics **9**(7) (2021)
51. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2023). R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org
52. Eugster, M.J.A., Leisch, F.: From spider-man to hero—archetypal analysis in R. J. Stat. Softw. **30**(8), 1–23 (2009)
53. Saad, Y.: Numerical Methods for Large Eigenvalue Problems, Revised SIAM, Philadelphia (2011)
54. Schwarz, A., Karlsson, L.: Scalable eigenvector computation for the non-symmetric eigenvalue problem. Parallel Comput. **85**, 131–140 (2019)
55. Vinue, G., Epifanio, I.: Robust archetypoids for anomaly detection in big functional data. Adv. Data Anal. Classif. **15**(2), 437–462 (2021)
56. Krauss, I., Grau, S., Mauch, M., Maiwald, C., Horstmann, T.: Sex-related differences in foot shape. Ergonomics **51**(11), 1693–1709 (2008)
57. Delgado-Abellán, L., Aguado, X., Jiménez-Ormeño, E., Mecerreyes, L., Alegre, L.M.: Foot morphology in Spanish school children according to sex and age. Ergonomics **57**(5), 787–797 (2014)
58. Hong, Y., Wang, L., Xu, D.Q., Li, J.X.: Gender differences in foot shape: a study of Chinese young adults. Sports Biomechan. **10**(02), 85–97 (2011)

59. Krauss, I., Langbein, C., Horstmann, T., Grau, S.: Sex-related differences in foot shape of adult Caucasians—a follow-up study focusing on long and short feet. Ergonomics **54**(3), 294–300 (2011)
60. Tomassoni, D., Traini, E., Amenta, F.: Gender and age related differences in foot morphology. Maturitas **79**(4), 421–427 (2014)
61. Saghazadeh, M., Kitano, N., Okura, T.: Gender differences of foot characteristics in older Japanese adults using a 3D foot scanner. J. Foot Ankle Res. **8**(1), 29 (2015)
62. Jee, S.-C., Yun, M.H.: An anthropometric survey of Korean hand and hand shape types. Int. J. Ind. Ergonom. **53**, 10–18 (2016)
63. Lin, Y.-L., Lee, K.-L.: Investigation of anthropometry basis grouping technique for subject classification. Ergonomics **42**(10), 1311–1316 (1999)
64. Ritz-Timme, S., Gabriel, P., Obertovà, Z., Boguslawski, M., Mayer, F., Drabik, A., Poppa, P., De Angelis, D., Ciaffi, R., Zanotti, B., Gibelli, D., Cattaneo, C.: A new atlas for the evaluation of facial features: advantages, limits, and applicability. Int. J. Legal Med. **125**(2), 301–306 (2011)
65. Fuentes-Hurtado, F., Diego-Mas, J.A., Naranjo, V., Alcañiz, M.: Automatic classification of human facial features based on their appearance. PLoS ONE **14**(1), 1–20 (2019)
66. Sarakon, P., Charoenpong, T., Charoensiriwath, S.: Face shape classification from 3D human data by using SVM. In: The 7th 2014 Biomedical Engineering International Conference, pp. 1–5 (2014)
67. Koleva, M., Nacheva, A., Boev, M.: Somatotype and disease prevalence in adults. Rev. Environ. Health **17**(1), 65–84 (2002)
68. Buffa, R., Lodde, M., Floris, G., Zaru, C., Putzu, P.F., Marini, E.: Somatotype in Alzheimer's disease. Gerontology **53**(4), 200–204 (2007)
69. Singh, S.: Somatotype and disease: a review. Anthropologist **3**, 251–261 (2007)
70. Braga, J., Zimmer, V., Dumoncel, J., Samir, C., de Beer, F., Zanolli, C., Pinto, D., Rohlf, F.J., Grine, F.E.: Efficacy of diffeomorphic surface matching and 3d geometric morphometrics for taxonomic discrimination of early pleistocene hominin mandibular molars. J. Hum. Evol. **130**, 21–35 (2019)
71. Malousaris, G.G., Bergeles, N.K., Barzouka, K.G., Bayios, I.A., Nassis, G.P., Koskolou, M.D.: Somatotype, size and body composition of competitive female volleyball players. J. Sci. Med. Sport **11**(3), 337–344 (2008)
72. Sterkowicz-Przybycień, K., Sterkowicz, S., Biskup, L., Żarów, R., Kryst, Ł, Ozimek, M.: Somatotype, body composition, and physical fitness in artistic gymnasts depending on age and preferred event. PLoS ONE **14**(2), 1–21 (2019)
73. Ryan-Stewart, H., Faulkner, J., Jobson, S.: The influence of somatotype on anaerobic performance. PLoS ONE **13**(5), 1–11 (2018)
74. Viscosi, V., Cardini, A.: Leaf morphology, taxonomy and geometric morphometrics: a simplified protocol for beginners. PLoS ONE **6**(10), 1–20 (2011)
75. MacLeod, N.: The direct analysis of digital images (eigenimage) with a comment on the use of discriminant analysis in morphometrics. In: BIOLOGICAL SHAPE ANALYSIS: Proceedings of the 3rd International Symposium, pp. 156–182 (2015). World Scientific
76. Korem, Y., Szekely, P., Hart, Y., Sheftel, H., Hausser, J., Mayo, A., Rothenberg, M.E., Kalisky, T., Alon, U.: Geometry of the gene expression space of individual cells. PLoS Comput. Biol. **11**(7), 1–27 (2015)
77. Seth, S., Eugster, M.J.A.: Probabilistic archetypal analysis. Mach. Learn. **102**(1), 85–113 (2016)
78. Moliner, J., Epifanio, I.: Robust multivariate and functional archetypal analysis with application to financial time series analysis. Phys. A: Stat. Mech. Appl. **519**, 195–208 (2019)
79. Steyer, L., Stöcker, A., Greven, S.: Elastic analysis of irregularly or sparsely sampled curves. Biometrics **n/a**(n/a) (2022). https://doi.org/10.1111/biom.13706
80. Cabero, I., Epifanio, I., Piérola, A., Ballester, A.: Archetype analysis: a new subspace outlier detection approach. Knowl.-Based Syst. **217**, 106830 (2021)
81. Wu, Y., Kundu, S., Stevens, J.S., Fani, N., Srivastava, A.: Elastic shape analysis of brain structures for predictive modeling of PTSD. Front. Neurosci. **16** (2022). https://doi.org/10.3389/fnins.2022.954055
82. Ferro, F.F., Rampazzo, M., Beghi, A.: Elastic shape analysis for anomaly detection in fabric images. IFAC-PapersOnLine **54**(7), 67–72 (2021). 19th IFAC Symposium on System Identification SYSID 2021
83. Bauer, M., Eslitzbichler, M., Grasmair, M.: Landmark-guided elastic shape analysis of human character motions. Inverse Probl. Imaging **11**(4), 601–621 (2017)