Advances in Data Analysis and Classification manuscript No.
(will be inserted by the editor)

Archetypal shapes based on landmarks and extension to handle missing data

Irene Epifanio · María Victoria Ibáñez · Amelia Simó

Received: date / Accepted: date

Abstract Archetype and archetypoid analysis are extended to shapes. The objective is to find representative shapes. Archetypal shapes are pure (extreme) shapes. We focus on the case where the shape of an object is represented by a configuration matrix of landmarks. As shape space is not a vectorial space, we work in the tangent space, the linearized space about the mean shape. Then, each observation is approximated by a convex combination of actual observations (archetypoids) or archetypes, which are a convex combination of observations in the data set. These tools can contribute to the understanding of shapes, as in the usual multivariate case, since they lie somewhere between clustering and matrix factorization methods. A new simplex visualization tool is also proposed to provide a picture of the archetypal analysis results. We also propose new algorithms for performing archetypal analysis with missing data and its extension to incomplete shapes. A well-known data set is used to illustrate the methodologies developed. The proposed methodology is applied to an apparel design problem in children.

Keywords Statistical shape analysis · Archetype analysis · Archetypoid analysis · Anthropometric data · Children's wear · Missing data

Mathematics Subject Classification (2000) 62H11 · 62H25 · 62H30

1 Introduction

One of the main steps in image processing is the representation and description of the objects in the images. Specifically, here we concentrate on the analysis of their shape. A significant amount of research and activity has been carried out in recent

This paper has been partially supported by the Spanish Ministerio de Economía y Competitividad Project $DPI2013-47279-C2-1-R$.

I. Epifanio, M.V. Ibáñez and A. Simó
Department of Mathematics- IMAC, Universitat Jaume I
Tel.: +34-964728390
Fax: +34-964728429
E-mail: epifanio@uji.es

decades in the general area of shape analysis. By shape analysis, we mean a set of tools for describing, comparing, matching, deforming, and modeling shapes. Three major approaches can be identified in shape analysis based on how the object is treated in mathematical terms (Stoyan and Stoyan (1995)): Objects can be treated as subsets of $\mathbb{R}^m$, they can be described as sequences of points that are given by certain geometrical or anatomical properties (landmarks), or they can be defined using functions that represent their contours.

The majority of research has been restricted to landmark-based analysis, where objects are represented using k labeled points in the Euclidean space $\mathbb{R}^m$. These landmarks are required to appear in each data object and to correspond to each other in a physical sense. Seminal papers on this topic are Bookstein (1978), Kendall (1984), and Goodall (1991). The main references are Dryden and Mardia (1998) and Kendall et al (2009). In this paper we concentrate on this approach.

When the landmark-based approach is used, the corresponding shape space is a finite-dimensional Riemannian manifold, and statistical methodologies on manifolds must be used. There are several difficulties with generalizing probability distributions and statistical procedures to measurements in a non-vectorial space like a Riemannian manifold, but fortunately, there has been a significant amount of research and activity in this area over recent years. An excellent review is given by Pennec (2006).

Statistical learning can be supervised or unsupervised ((Hastie et al, 2009, Ch. 14) provide an excellent overview of unsupervised learning techniques) depending on whether or not outcome variables are present or not. This last is our case. In multivariate statistics, data decomposition techniques for finding the latent components are very popular. A data matrix is viewed as a linear combination of several factors. Different unsupervised methods emerge depending on the constraints on the factors and their combination (Mørup and Hansen (2012); Thurau et al (2012); Vinué et al (2015a)). For example, in clustering methods such as k-means (or k-medoids) data are explained by means of several centroids, which are the average of groups of data (or data points in the case of k-medoids). This makes factors easily interpretable. However, binary assignment of data to the clusters reduces their modeling flexibility, as compared with methods such as principal component analysis (PCA). PCA can explain data variability very well, but the factors obtained are not always easy to interpret, as they are a linear combination of the variables. Obviously, their objectives are different. Archetype analysis (AA) lies somewhere between these two techniques, in the sense that its modeling flexibility is higher than clustering methods and its factors are very easy to interpret. The same happens with fuzzy versions of k-means and k-medoids, although their objectives are different from that of AA. A summary table showing the relationship between several unsupervised methods in the multivariate context, which also applies to shapes, is given by Vinué et al (2015a).

AA was proposed by Cutler and Breiman (1994). The objective of AA is to represent data as a convex combination (mixture) of pure or extremal patterns called archetypes. Archetypes are a convex combination of data points. A variant of AA is archetypoid analysis (ADA). Unlike AA, the pure types in ADA are not a mixture of cases, but real (observed) cases. ADA represents the data as mixtures of extreme cases, and not as mixtures of mixtures, as AA does. ADA was proposed by Vinué et al (2015a).

The archetypal patterns obtained in both AA and ADA are extreme observations (archetypes belong to the convex hull of data (Cutler and Breiman (1994)). Representing data by their extreme constituents (Davis and Love (2010)) facilitates human interpretation and understanding of data due to the principle of opposites (Thurau et al (2012)). Central points are not as good as extreme points for human interpretability. Furthermore, the expression of data as a convex combination of archetypal patterns makes a human reading of data easier, unlike a linear combination without any constraints.

Besides ADA, AA has given rise to the development of other new methodologies and it has been extended to other kinds of data. Some examples are weighted and robust AA (Eugster and Leisch (2011)), interval archetypes (D'Esposito et al (2012)), functional AA and ADA (Epifanio (2016)), data-driven prototype identification (Ragozini et al (2017)), probabilistic AA (Seth and Eugster (2016b)), AA for nominal observations (Seth and Eugster (2016a)) and archetypal networks (Ragozini and D'Esposito (2015)).

AA and ADA have been applied in many different fields, such as astrophysics (Chan et al (2003)), biology (D'Esposito et al (2012)), developmental psychology (Ragozini et al (2017)), e-learning (Theodosiou et al (2013)), genetics (Thøgersen et al (2013)), global development (Epifanio (2016)), industrial engineering (Epifanio et al (2013); Vinué et al (2015a)), machine learning problems (Mørup and Hansen (2012)), market research (Li et al (2003); Porzio et al (2008); Midgley and Venaik (2013)), multi-document summarization (Canhasi and Kononenko (2013, 2014)), neuroscience (Tsanousa et al (2015); Hinrich et al (2016)) and sports (Eugster (2012); Vinué and Epifanio (2017)). AA and ADA algorithms are implemented in the R package **archetypes** (Eugster and Leisch (2009)) and **Anthropometry** (Vinué et al (2015b); Vinué (2017)), respectively.

Displaying figures that correspond to extreme scores of principal components is quite a common exploratory tool in statistical shape analysis (Dryden and Mardia (1998); Claude (2008)). This could be viewed as looking for the archetypal shapes. Nevertheless, unlike PCA, the purpose of AA or ADA is to find extreme observations, and cases with extreme PCA scores do not necessarily return archetypal cases. This is explained by Cutler and Breiman (1994) and shown by Epifanio et al (2013) through a problem where archetypes could not be recovered with PCA even if all the components had been considered.

Although it is common to study the average and variability of shapes in morphometrics, many applications have to cope with the analysis of extreme shapes. Several biological applications where the interest is on extreme shapes rather than mean shapes are introduced by Dryden and Zempléni (2006): the analysis of healthy and diseased muscle fiber cells and a time series of 4000 conformations of a DNA molecule over 4 nanoseconds. In several diseases, the morphology of certain organic structures is affected and they have extreme shapes when compared with those found in controls. For example, spines in scoliosis patients, or corneal endothelium cells in pathological cases (Ayala et al (2006); Zapater et al (2002)). Another potential biological application is taxonomy, since the interest is precisely on species (pure types or archetypes) and their hybrids (mixture of archetypes). Therefore, instead of the usual application of PCA in this field (Viscosi and Cardini (2011)), AA for shapes could be very useful.

However, there are sometimes error-prone shape data (Du et al (2015)), i.e. shape data that are not free from measurement errors, or some landmarks may even be missing for different observations (Brown et al (2012)); for example, if we are working with fossilized specimens, which are frequently subject to fragmentation, distortion and erosion (Arbour and Brown (2014)). Excluding the cases (or missed landmarks) with missing data is not a good option (Arbour and Brown (2014)), especially if the sample size is small and the collection of additional, more complete, material is not possible. Another option would be imputation, provided that the proportion of missing data is small. Nevertheless, if the proportion is large, the errors caused by imputation are increasingly important (Eirola et al (2013)). For that reason, we also propose a new procedure for computing AA with missing data in the multivariate case and we extend it to shapes with missing landmarks. A previous attempt to compute multivariate AA with missing data was introduced by Mørup and Hansen (2012), but by modifying the original objective function of AA.

In industrial design, extremes are also very important. Human modeling is widely used in many industries such as the aviation, automotive, defense and manufacturing sectors. The use of representative human models (cases) provides designers with an efficient way of applying the body size characteristics of the target population to ergonomic design and evaluation. Boundary cases are those located toward the edges of the distribution, and are very relevant in ergonomic product design assuming that the accommodation of boundaries ensures the accommodation of interior points. Note that if the "hard-to-fit" extreme subjects (the boundary cases) are located before the design process begins, the design could be improved from the very beginning. This would reduce the time and cost of the design process.

The problem that motivated us is concerned with the analysis of children's shapes for garment fit problems, although it could be applied to the design of other products for an appropriate target population. Lack of fit is a significant problem for both customers and apparel companies. In fact, the return rates due to size misfit are very high in online garment shops (Eneh (2015)). The idea is to identify individuals who represent the fitting problems of the target population by means of archetypal shapes. Then the designer may adapt the base pattern to ensure that each new pattern is adapted to the measurements of the extremes of a size. A 3D anthropometric study of the child population in Spain was carried out by the Biomechanics Institute of Valencia. The aim of this study was to obtain anthropometric data from the child population for the clothing industry. A total of 502 Spanish children aged 6 to 12 years old were randomly selected. They were scanned using the Vitus Smart 3D body scanner from Human Solutions, a non-intrusive laser system formed by four columns housing the optic system, which moves from head to feet in ten seconds, performing a sweep of the body.

The purpose of this work is to extend AA and ADA to statistical shape analysis, which will help make a shape data set easier to understand, displaying and describing the relevant features of the data. AA and ADA in the multivariate case is reviewed in Section 2, together with some aspects of shape analysis, and the novel extension of AA and ADA to shapes is introduced. In Section 3, AA and ADA with landmarks is illustrated with a very well-known data set, and the results are compared with those obtained with PCA, sparse PCA and k-means clustering. A new simplex visualization tool is also proposed to provide a picture of the archetypal analysis results. In Section 4, AA with missing data is proposed, together with its extension to shapes with missing landmarks. The procedure is again illustrated and compared with different alternatives in the multivariate case and shape case in the supplementary material (Online Resource 1). ADA with missing data is also discussed. In Section 5, children's shapes for garment design are analyzed. Finally, conclusions and further developments are discussed in Section 6. The code in R (R Development Core Team (2017)) and data for reproducing the examples are available at http://www3.uji.es/∼epifanio/RESEARCH/laa.rar.

2 Definition of AA and ADA with landmarks

2.1 AA and ADA in the multivariate case

Let $X = (x_1, ..., x_n)$ be an $n \times r$ data matrix with n cases and r variables. For AA, the $p \times r$ matrix $Z = (z_1, ..., z_p)$ will contain the p archetypes in those data, in such a way that row x_i is approximated by a mixture of the rows z_j's (archetypes):

$$x_i \sim \hat{x}_i = \sum_{j=1}^{p} \alpha_{ij} z_j, \tag{1}$$

with the mixture coefficients contained in the $n \times p$ matrix $\alpha = (\alpha_{ij})$. On the other hand, the archetypes z_j are obtained with mixture coefficients compiled in the $p \times n$ matrix $\beta = (\beta_{jl})$ according to:

$$z_j = \sum_{l=1}^{n} \beta_{jl} x_l. \tag{2}$$

To determine the matrices Z, α and β for a given data matrix X, the following residual sum of squares (RSS) is minimized ($\|\cdot\|$ denotes the Frobenius matrix norm for matrices, and the Euclidean norm for vectors):

$$RSS = \|X - \alpha\beta X\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{p} \alpha_{ij} z_j\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{p} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l\|^2, \tag{3}$$

under the constraints

1) $\sum_{j=1}^{p} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ and $i = 1, \dots, n$, $j = 1, \dots, p$ and

2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ and $j = 1, \dots, p$ and $l = 1, \dots, n$.

Constraint 1) indicates that the approximation of x_i is a convex combination of archetypes, $\hat{x}_i = \sum_{j=1}^{p} \alpha_{ij} z_j$. Each α_{ij} is the weight of the archetype j for the case i; that is to say, the α coefficients indicate how much each archetype contributes to the approximation of each case. Constraint 2) indicates that archetypes z_j are case mixtures, $z_j = \sum_{l=1}^{n} \beta_{jl} x_l$.

Archetypes z_j are not necessarily observed data points. z_j would only be an observed data point if one β_{jl} was equal to 1 in constraint 2) for each j. In ADA, therefore, β_{jl} can only take on the binary values 0 or 1, because $\beta_{jl} \geq 0$ and the sum of constraint 2) is 1. As a consequence, instead of the continuous optimization problem of AA, the following mixed-integer optimization problem has to be solved in ADA:

$$RSS = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{p} \alpha_{ij} z_j\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{p} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l\|^2, \tag{4}$$

under the constraints

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

Note that $\beta_{jl} = 1$ for one and only one l, otherwise $\beta_{jl} = 0$.

Due to their definitions, archetypes and archetypoids are extremal representatives of the data. If $p > 1$, archetypes are located on the boundary of the convex hull of the data (see Cutler and Breiman (1994)), while this does not necessarily hold for archetypoids (see Vinué et al (2015a)). If $p = 1$, the mean of the data is the archetype, and the archetypoid is the medoid of the data (Kaufman and Rousseeuw (1990)).

An alternating minimizing algorithm was proposed by Cutler and Breiman (1994) to solve the problem (3). It alternates between estimating the best α for given archetypes Z and the best archetypes Z for a given α. The convex least squares problems were solved with a penalized version of the non-negative least squares algorithm by Lawson and Hanson (1974). Eugster and Leisch (2009) implemented that algorithm in the R library **archetypes**, where data are standardized by default. We have based our algorithm on this one, but the data are not standardized and the Frobenius norm is considered, as indicated in equation (3), instead of the spectral norm used by Eugster and Leisch (2009).

An algorithm based on the idea of the Partitioning Around Medoids (PAM) clustering algorithm (Kaufman and Rousseeuw (1990)) was proposed by Vinué et al (2015a) to solve the problem (4). A BUILD phase and a SWAP phase are considered in that algorithm. From an initial set of archetypoids calculated in the BUILD step, the SWAP phase improves that set by exchanging selected cases for unselected observations and by checking if these replacements reduce the RSS. That algorithm was implemented in the R library **Anthropometry** by Vinué et al (2015b). Three alternatives for the BUILD phase are considered in the R implementation. The first candidates are the nearest neighbors in Euclidean distance to the p archetypes, the so-called $cand_{ns}$ set. The second initial candidates, referred to as the $cand_{\alpha}$ set, are the cases with the maximum α value for each archetype j, i.e., the cases with the largest relative share for the respective archetype. The third set of candidates, the $cand_{\beta}$ set, consists of the observations with the maximum β value for each archetype j, i.e., the major contributors in the generation of the archetypes. From these three initial sets, after the SWAP phase, ADA returns three sets of archetypoids. The set with lowest RSS (often the same set is obtained from the three initializations) is the returned ADA solution.

An open question is the number p of archetypes or archetypoids to compute. Note that neither archetypes nor archetypoids are necessarily nested. It could be decided by the user or the elbow criterion could be used as made by Cutler and Breiman (1994); Eugster and Leisch (2009); Vinué et al (2015a) (the value p is selected as the point where the elbow on the RSS representation for a series of different p values is found).

2.2 Shape space and shape distances

The word "shape" is very commonly used in everyday language, usually referring to the visual appearance of a geometric object. More formally, shape can be defined as geometric information about the object that is invariant under a Euclidean similarity transformation, i.e., with respect to location (translation), orientation (rotation) and scale (size) (Dryden and Mardia (1998)). In this work, the shape of geometrical m-dimensional objects (usually $m = 2, 3$) is determined by a finite number of $k > m$ coordinate points, known as landmark points. Each object is then described by a $k \times m$ configuration matrix $X = (x_{ij})$ containing the m Cartesian coordinates x_{ij} of its k landmarks, i.e. each row represents a landmark and each column represents one Cartesian coordinate of that landmark.

However, a configuration matrix X is not a proper shape descriptor because it is not invariant to similarity transformations. For any similarity transformation, i.e., for any translation vector $b \in \mathbb{R}^m$, scale parameter $s \in \mathbb{R}^+$, and $m \times m$ rotation matrix R, the configuration matrix given by $sXR + 1_k b^T$ (where 1_k is the $k \times 1$ vector of ones, and the superscript T means transpose) describes the same shape as X.

Definition 1 The shape space Σ_m^k is the set of equivalence classes $[X]$ of $k \times m$ configuration matrices $X \in \mathbb{R}^{k \times m}$ under the action of Euclidean similarity transformations.

A representative of each equivalence class $[X]$ can be obtained by removing the similarity transformations one at a time. There are different ways to do that.

Let X be a configuration matrix. One way to remove the location effect consists of multiplying it by the Helmert sub-matrix ((Dryden and Mardia, 2016, p. 49-50)), H, i. e., $X_H = HX$.

To filter scale we can divide X_H by the centroid size, which is given by $S(X) = \|X_H\|$.

$$Y = \frac{X_H}{\|X_H\|} \tag{5}$$

is called the pre-shape of the configuration matrix X because all information about location and scale is removed, but rotation information remains.

Definition 2 The pre-shape space S_m^k is the set of all possible pre-shapes Y.

S_m^k is a hypersphere of unit radius in $\mathbb{R}^{m(k-1)}$ (a Riemannian manifold that is widely studied and known). Σ_m^k is the quotient space of S_m^k under rotations.

As a result, a shape $[X]$ is an orbit generated by the rotation group $SO(m)$ on the pre-shape space.

For $m = 2$, this quotient space is isometric with the complex projective space $\mathbb{CP}_{k-2}$, a familiar Riemannian manifold without singularities. For $m > 2$, which is the case of our application, Σ_m^k is not a familiar space, and it has singularities; however, the Riemannian structure of the non-singular part of Σ_m^k can be obtained taking into account that the quotient space $\Sigma_m^k/SO(m)$ is a Riemannian submersion; see Kendall et al (2009).

Different distances between shapes can be introduced in Σ_m^k. One of the most popular is the full Procrustes distance. Given two configuration matrices X_1 and X_2, the full Procrustes distance is a least-squares type metric between the corresponding pre-shapes Y_1 and Y_2, respectively.

Definition 3 The full Procrustes distance between configuration matrices X_1 and X_2 is defined by:

$$d_F(X_1, X_2) = \inf_{R \in SO(m), \beta \in \mathbb{R}} \|Y_2 - \beta Y_1 R\|, \tag{6}$$

$SO(m)$ being the orthogonal group of rotations.

The solution of this optimization problem is given by:

$$d_F(X_1, X_2) = \sqrt{1 - (\sum_{i=1}^{m} \lambda_i)^2},$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_{m-1} \geq | \lambda_m |$ are the square roots of the eigenvalues of $Y_1^T Y_2 Y_2^T Y_1$, and the smallest value λ_m is the negative square root if and only if $\det Y_1^T Y_2 < 0$ Dryden and Mardia (1998).

Alternative distances could be used in the shape space. In particular, we will also work with:

Definition 4 The Procrustes distance $\rho(X_1, X_2)$ is the closest (over rotations) great circle distance (the shortest distance between two points on the surface of a sphere, measured along the surface of the sphere) between Z_1 and Z_2 on the pre-shape space S_m^k.

The relationship between d_F and ρ is (Kendall et al (2009)): $d_F(X_1, X_2) = \sin \rho(X_1, X_2)$.

Based on the full Procrustes distance a concept of mean shape $[\widehat{\mu}]$ can be introduced in the Fréchet sense Fréchet (1948), i.e., one that minimizes the sum of squared distances from any shape in the set.

Definition 5 Given a set of configuration matrices $X_1, \ldots, X_n$, the full Procrustes mean in Σ_m^k is given by

$$[\widehat{\mu}] = \arg \inf_{\mu : S(\mu)=1} \sum_{i=1}^{n} d_F^2(X_i, \mu). \tag{7}$$

For two-dimensional data an explicit eigenvector solution of the optimization problem (5) is available (see p. 44 by Dryden and Mardia (1998)), but for $m = 3$ and higher dimensional data, an iterative procedure must be used.

2.3 AA and ADA for shapes

Let $X_1, \dots, X_n$ be n landmark configuration matrices, with k landmarks in dimension m, each X_i is the representative of a shape $[X_i]$, an element of the shape space Σ_m^k. Henceforth, in order to simplify the notation, we will use X to denote both a configuration matrix and its shape, provided that it is understood from the context.

Our objective is to find shapes $Z_j \in \Sigma_m^k$, $j = 1, .., p$ that generalize the settings (1) and (2) to this space. They cannot be used directly because shape space is not a vectorial space and the expressions $\sum_{j=1}^{p} \alpha_{ij} Z_j$ and $\sum_{i=1}^{n} \beta_{jl} X_l$ are not defined.

Instead we propose to project X_1, ..., X_n to an approximating linear space where this expression could be defined and then take it back to Σ_m^k.

In order to obtain this linear space we have to take into account that, as noted before, the non-singular part of Σ_m^k has a Riemannian manifold structure and so it is locally homeomorphic with a Euclidean space such that the local homeomorphisms can be smoothly patched together. This leads us to the definition of the tangent space at each point (pole) that is linear and contains all possible directions at the pole.

Additionally, the tangent space has the desirable property that the distance from the shape to the pole is preserved, i.e. the distance from a point in the manifold to the pole is equal to the Euclidean distance between its projections in the tangent space. The type of distance depends on the choice of the tangent; in our application the full Procrustes distance will be adopted (Dryden and Mardia (2016)).

As one moves away from the pole, Euclidean distances between some pairs of points in the tangent space are smaller than their corresponding shape distances. This distortion becomes larger as one considers points that are more distant from the pole. For this reason, the pole should be taken close to all of the points and the mean of the observed shapes is the best choice (Rohlf (1998); Slice (2001)).

The tangent space of the shape space at the Procrustes mean (7) is called the Procrustes tangent space.

As discussed before, if the data are fairly concentrated around this mean, the Euclidean distance in the Procrustes tangent space is a good approximation to d_F, and standard multivariate techniques based on Euclidean distance can be used in this space. For this reason, this approach is widely used for statistical inference on the shape space in many applications.

To know whether shape variation is sufficiently small in practical applications, Rohlf (1999) suggests comparing the Euclidean distances between all pairs of points in the tangent space (or simply the distance to the average shape) against their Procrustes distances in the shape space. Furthermore, Rohlf (1999) also points out that, at least in biological applications, the approximation would usually be good when there are more than just a few landmarks.

In terms of Riemannian manifolds, the maps that allow us to move from the manifold to the tangent space or vice versa are called logarithmic and exponential maps, respectively. For $m = 2$ their expressions are easy because, as noted before, in this case the shape space is isometric to the complex projective space. To obtain their expressions in the case of $m > 2$, which is the case of our application, we have to take into account that the mapping π that assigns to each preshape Y in S_m^k the corresponding element $\pi(Y)$ in Σ_m^k is a Riemannian submersion and it maps the horizontal subspace of the tangent space to the pre-shape sphere at Y isometrically onto the tangent space to the shape space at $\pi(Y)$. Using this result, the exponential and logarithm maps in Σ_m^k can be computed (Dryden and Mardia (2016)).

Before showing the calculus, it is necessary to introduce the vectorizing operator. The vectorizing operator of an $l \times m$ matrix A with columns $a_1, a_2, \ldots, a_m$ is defined as: $\mathrm{vec}(A) = (a_1^T, a_2^T, \ldots, a_m^T)^T$.

Let S be the pre-shape of the Procrustes mean μ of $X_1, ..., X_n$ and $Y_1, ..., Y_n$ their respective preshapes, obtained using equation (5). To obtain the expression of the projection onto the tangent plane at S of $X_1, ..., X_n$, the pre-shape Y_i is rotated to be as close as possible to S. We write the rotated pre-shape as $Y_i \hat{\Gamma}_i$ with the rotation matrix $\hat{\Gamma}_i$. The expression of $\hat{\Gamma}_i$ can be found on p. 61 of Dryden and Mardia (1998):$\hat{\Gamma}_i = U_i V_i^T$, where $U_i, V_i \in SO(m)$ are the left and right matrices of the singular value decomposition of $S^T Y_i$. Then, the projection of Y_i on the tangent space at S is:

$$\log_S(Y_i) = (I_{km-m} - \mathrm{vec}(S)\mathrm{vec}(S)^T)\mathrm{vec}(Y_i\hat{\Gamma}_i)\frac{trace(S^T Y_i \hat{\Gamma}_i)}{\sin(trace(S^T Y_i \hat{\Gamma}_i))}, \qquad (8)$$

where I_{km-m} is the $(km - m) \times (km - m)$ identity matrix.

The logarithmic tangent projection has the property that it preserves the Procrustes distance (Def. 4). Instead of using these coordinates, we propose to use:

$$v_i = \log_S(Y_i)\frac{\sin(trace(S^T Y_i \hat{\Gamma}_i))}{trace(S^T Y_i \hat{\Gamma}_i)}.$$

The v_i coordinates are called Kent's partial tangent coordinates. We chose this projection because instead of preserving the Procrustes distance it preserves the full Procrustes distance (Def. 6). The practical difference is that, using Kent's partial tangent coordinates the more extreme observations are pulled in more towards the pole Dryden and Mardia (2016).

We are now ready to introduce the definition of AA and ADA for shapes. Let $v_1, \ldots, v_n$ be the tangent coordinates of $X_1, \ldots, X_n$. The coordinates in the tangent space u_j $j = 1, .., p$ of the archetypes $Z_j \in \Sigma_m^k$, $j = 1, .., p$ are obtained by minimizing:

$$RSS = \sum_{i=1}^{n} \|v_i - \sum_{j=1}^{p} \alpha_{ij} u_j\|^2 = \sum_{i=1}^{n} \|v_i - \sum_{j=1}^{p} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} v_l\|^2, \qquad (9)$$

under the constraints

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

For the case of ADA, the definition is analogous, but constraint 2) is changed as in Eq. 4.

A significant difference between our methodology and the usual statistical methods for shape analysis is that once archetypes are computed in the tangent space, they are projected back into the configuration space. For this reason, usually this inverse function cannot be found in standard shape analysis software. The calculus of this map is again given by Dryden and Mardia (1998) and explained below. Given u_j $(j = 1, ..., p)$ on the tangent space at S obtained by solving the minimization problem (9), the corresponding points Y_{Z_j} $j = 1, ..., p$ on the pre-shape space are:

$$Y_{Z_j} = \mathrm{vec}^{-1}((1 - u_j^T u_j)^{1/2}\mathrm{vec}(S) + u_j). \tag{10}$$

Finally, the configuration matrices representing Z_j $j = 1, ..., p$ would be:

$$X_{Z_j} = H^T Y_{Z_j}. \tag{11}$$

As archetypoids correspond to concrete observations, besides inspecting their landmarks, if other information about the observation is available, it can also be inspected. For example, in many cases landmarks are extracted from images, so not only can the landmark configuration of archetypoids be visualized, but also the corresponding images from which the landmarks were extracted. This is the case of the problem in hand.

3 Comparison with other unsupervised methods - an example

Firstly, we illustrate AA and ADA with landmarks using digit 3 data, a well-known and simple database that appears as an example of PCA application by Dryden and Mardia (1998). It is available in the R package **shapes** Dryden (2015). It consists of $k = 13$ landmarks in $m = 2$ dimensions from $n = 30$ individuals (see Dryden and Mardia (1998) for details about this database). As landmarks are two-dimensional, results can be visualized more simply. We apply AA, ADA, PCA, sparse PCA (SPCA) and k-means to this database to better understand the differences between the various methodologies for obtaining representative data. Throughout this work, function *procGPA* (with 'partial' tangent coordinates) from the R package **shapes** is used for registering landmark configurations into optimal registration using translation, rotation and scaling, and matrix V is obtained. The same function is used to carry out PCA. AA, ADA and SPCA are applied to V. SPCA aims to produce easily interpreted models through sparse loadings, i.e. PC components are a linear combination of a subset of the original variables. An excellent explanation of SPCA for landmark-based shape analysis is given by Sjöstrand et al (2007); more examples of SPCA in medical shape modeling are shown by Sjöstrand et al (2006). For computing SPCA in the multivariate case, the R package **elasticnet** Zou and Hastie (2012) is used. For carrying out k-means clustering with shapes, we consider Lloyd's classic algorithm for k-means clustering adapted to the context of Shape Analysis, as made by Vinué et al (2016).

First of all, Procrustes distances to the mean are computed. The distance from the first digit to the mean is very high and is considered an outlier by Dryden and Mardia (1998). Therefore, the analyses are performed without this first digit as made by Dryden and Mardia (1998). In the interest of brevity and as an illustrative example we examine the results of $p = 4$ with AA (the RSS elbow is found at $p = 4$) and ADA. We also consider the first four PC components in PCA and SPCA (with 6, 25% of the variables, non-zero loadings in each component), and $k = 4$ clusters with k-means. The same number of representative objects is considered for all the techniques in order to better compare the results. The first four PC components obtained by PCA explained a total of 84.2% of the variability; 43.6%, 18.4%, 13.2% and 9.0%, respectively. While, the first four SPC components obtained by SPCA explained a total of 50.1% of the variability; 22.7%, 12.0%, 10.9%, 4.5%, respectively. Note the difference in the total variability explained by PCA and SPCA.

Figures 1 and 2 display the PCs for PCA and SPCA. Icons as explained by (Dryden and Mardia, 1998, Sect. 5.5) are plotted. Scores are standardized to make them easier to interpret.

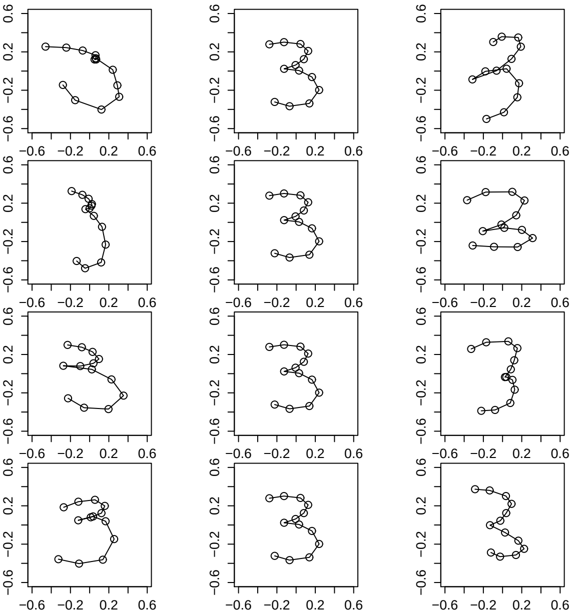


Fig. 1 Plots of the first four PCs of the digit 3s. In the ith row: *mean* - $3sd\ PC_i$, *mean*, *mean* + $3sd\ PC_i$ (where *mean* is the full Procrustes mean, sd is the standard deviation and $i = 1, 2, 3, 4$).

The first PC of PCA can be interpreted as mainly measuring the length of the central part, but also the length of the top loop, and the curl of the bottom loop. The second PC contrasts tall thin digits versus fat short digits, encompassing many characteristics of the digits. The third PC contrasts digits with wide top loop versus wide bottom loop digits, but also the tilt of loops. The interpretation of the fourth PC is not as clear as the previous ones, but it can be interpreted as measuring the angle at which both loops join and also the length of the loops. Although in the example in Dryden and Mardia (1998), the interpretation of the PCs is quite evident, it is more difficult for higher order PCs. In fact, the difficulty of interpreting them in many other problems leads to the use of SPCA. PCA usually produces holistic modes of variation, describing a series of effects at once. There are often many combined effects on each PC.

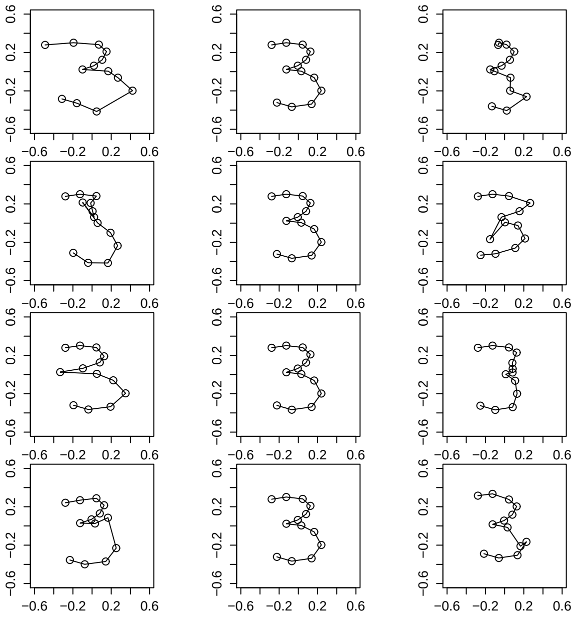


Fig. 2 Plots of the first four SPCs of the digit 3s. In the ith row: *mean* - $3sd\ SPC_i$, *mean*, *mean* + $3sd\ SPC_i$ (where *mean* is the full Procrustes mean, *sd* is the standard deviation and i = 1, 2, 3, 4).

Unlike PCA, each SPC concentrates on more or less separate effects. On the other hand, more SPCs are needed to collect a variety of effects and explain the variability of the data. As more components are needed, human understanding becomes a more difficult task. It is difficult for our brain to process the meaning of many components at the same time. The first SPC effect concentrates mainly on the following landmarks: extreme top left points and the point with the maximum curvature of the bottom arc. The second SPC effect concentrates on the point with the maximum curvature of the top arc and the point that marks the extreme end of the central protrusion in a vertical sense. However, the third SPC concentrates on this same point but in a horizontal sense. The fourth SPC reflects the different position of pseudo landmarks between the point with the maximum curvature of the bottom arc and the point marking the extreme end of the central protrusion. As can be seen, SPC effects are not as global as PC effects, but they are very localized. Note that for interpreting PCs or SPCs, the icons corresponding to extreme values of standardized PC scores were plotted. Note also that the values of the PC or SPC scores can take any value; they are not constrained as in AA or ADA to the $[0, 1]$ interval, neither do they add one. Therefore, their standardization facilitates their interpretation.

Figure 3 displays the centers of each cluster obtained by k-means. Some contrast can be found between the cluster centers corresponding to the top row, as regards the extreme top left point and the point marking the extreme end of the central protrusion. However, the cluster centers corresponding to the bottom row of Figure 3, especially the digit on the left of the bottom row, are typical. They have elegant calligraphy with no significant features. Furthermore, as k-means only return the assignments to each cluster, we cannot obtain information about deviations from the mean or where they occur.

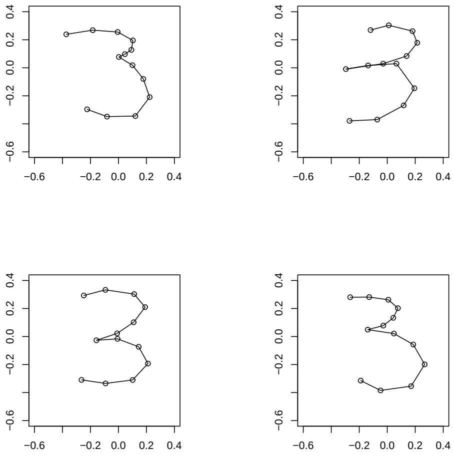


Fig. 3 Mean shapes for the 4 clusters obtained with k-means.

Archetypal shapes are precisely extreme shapes, which are easier for humans to interpret than central points. Figures 4 and 5 show the archetypes and archetypoids. The archetypal shapes obtained for AA and ADA are quite similar, so we only discuss the results for AA. Remember that archetypoids are concrete cases from the database. The most relevant characteristic of the first archetype is the long length of the central protrusion. By contrast, the central protrusion is almost missing in the second archetype, which also presents very long loops. The third archetype is a thin digit with a large bottom arc. The fourth archetype can be seen as the opposite of the third archetype.

Archetypoids do not have to be those cases with the highest Kendall's Riemannian distance ρ to the mean shape. In this example, the archetypoids are the cases with the first, second, fourth and ninth largest distances. The variability explained with $p = 4$ archetypes is 66%. To explain 84% of the variability, as with PCA, p $= 8$ archetypes would be necessary. However, it should be remembered that PC scores can take any value, so in order to interpret them the sign is very important. A positive sign in a PC score reflects different shapes to the same PC score with a negative sign. Therefore, although 4 PCs were computed, in Figure 1 eight plots (without considering the mean shapes) were necessary. In addition, the constraints for α in AA and ADA are a major advantage for understanding the data. Each

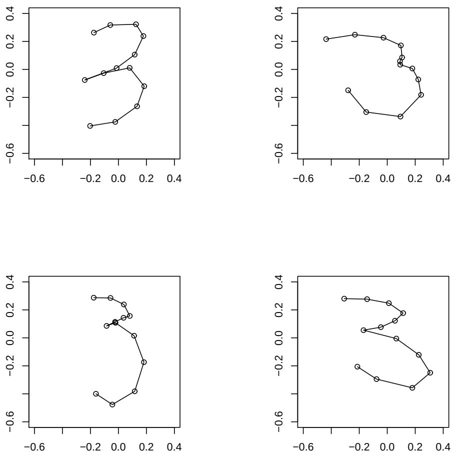


Fig. 4 Four archetypes of the digit 3s.

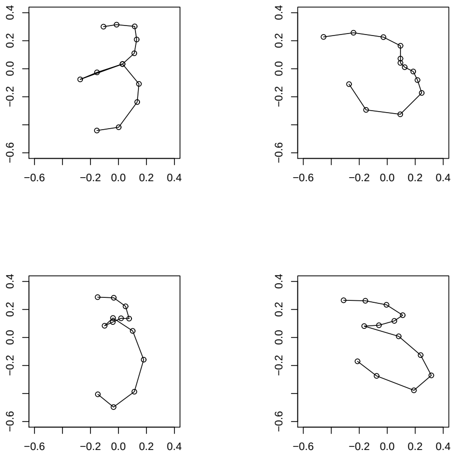


Fig. 5 Digit 3 example: The four archetypoids are the sample no. 2, 28, 15, 7 (after removing the first digit).

digit is approximated by a mixture of the archetypal shapes. In this problem, the second and third archetypoids are 100% explained by the respective archetype, while the first and fourth archetypoids are 97% and 99% explained, respectively. The other digits are approximated by mixtures of two, three or four archetypes. For example, the 10th digit of the data set is explained 82% by archetype 1 and 18% by archetype 2; the 19th digit is explained 14% by archetype 1, 50% by archetype

3 and 37% by archetype 4; the mean shape is explained 31% by archetype 1, 20% by archetype 2, 19% by archetype 3 and 31% by archetype 4.

Several kinds of plots for simplex visualization for classic multivariate AA are available in the R package **archetypes** (Eugster and Leisch (2009)) to represent the information in α. The left-hand panel in Figure 6 shows the plot with parallel coordinates, with α on the Y axis: the $n = 29$ α_j values of each archetype j are represented in vertical axes at $X = 1$, 2, 3 and 4, respectively. Note that it is difficult to appreciate this information. The more observations there are, the more difficult it is to see anything, as the lines are superimposed. The right-hand panel of Figure 6 shows a simplex plot visualization, where $A1$, $A2$, $A3$ and $A4$ represent the archetypes from $j = 1$ to 4. However, they do have their limitations, as information is projected in 2D. With more than three archetypes, the information in this plot can be misleading due to the non-uniqueness of the projections (see Seth and Eugster (2016b) for details about this issue).

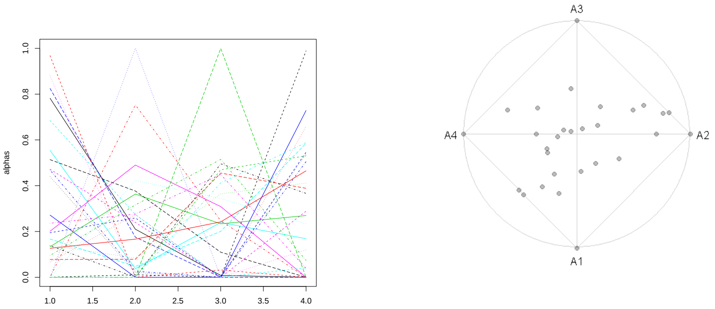


Fig. 6 α for AA of digit 3s. Left: parallel coordinates; Right: simplex visualization.

As the sample size of this database is small, we propose a scheme to show the information in α by exploiting the constraints in the definition of this matrix. This keeps all the information contained in this matrix visible, unlike the previous simplex visualization tools. First, we compute the Gini coefficient of the matrix α for each observation. The Gini coefficient is a concentration index that measures inequality. A Gini coefficient of one for an observation i indicates that only one archetype j explains that observation, i.e. its α_{ij} for that archetype is 1. On the other hand, a Gini coefficient of zero indicates that that observation is a mixture of all the archetypes; all its α_{ij} values are the same (i.e., $\alpha_{ij} = 1/p$). The higher the Gini coefficient, the less mixed the observation is. In this way, the purity of the observation is measured. Observations are ordered according to their Gini coefficient, i.e., from the purest cases to the most mixed cases. To the best of our knowledge, this is the first time a multivariate ordering based on the purity of the observations obtained by archetypes has been proposed. The most closely related approach would be the procedure proposed by D'Esposito and Ragozini (2008). They ranked multivariate performances based on the idea of the "worst-best" direction selected by applying AA with $p = 2$ archetypes (one of them is considered the “worst” case and the other is the “best” case). Once observations are ordered using the Gini coefficient, a barplot of α is carried out. This barplot for our example can be seen in Figure 7. More digits are similar to archetype 1, i.e., well explained by this archetype, than to other archetypes. Archetype 4 usually appears in the digits explained by mixtures consisting of only two archetypes (represented with two colors in the corresponding bar). Approximately half of the digits (the last digits represented in the barplot) in the database are explained as mixtures of three or four archetypes, with no high coefficients (less than 0.5) for any of them.

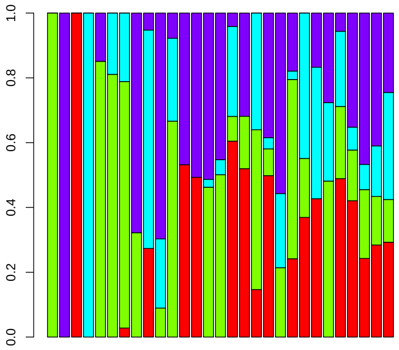


Fig. 7 α barplot for four archetypes of the digit 3s, ordered using the Gini coefficient. Red, green, cyan and blue colors for α values correspond to archetype 1, 2, 3 and 4, respectively.

4 AA and ADA with missing data

4.1 AA with missing values in the multivariate case

We follow the same strategy used in many basic functions in R R Development Core Team (2017) for handling missing data. Let X be an $n \times r$ matrix as before, but now allowing missing values (NAs). Let us suppose that there is no row or column with all its values missing; otherwise, that row or column would have to be removed. The missing values are excluded from the computations, but they are scaled up proportionally to the number of elements used. For example, if the mean of a column q has to be calculated and there are n_q missing values in that column, then the divisor for the mean computation should be $n - n_q$. This is equivalent to weighting the non-missing values by $n/(n - n_q)$ and using n as divisor for the mean computation. Analogously, if the squared norm of x_i has to be calculated and there are r_i missing values, then the non-missing values should be weighted by $r/(r - r_i)$. This is equivalent to $r/(r - r_i)$ times the squared norm of x_i computed with the non-missing values.

In summary, $Z = \beta X$ can be computed as follows. Let W be an $n \times r$ matrix with zeros in the positions where there are missing values in X, ones in the column q if there are not any missing values in X for that column, and $n/(n{-}n_q)$ otherwise. If we denote $.*$ as the element-wise multiplication and define $NA * 0$ as zero, then multiplying X and W element by element allows us to compute Z as $\beta X. * W$. Therefore, no archetype (Z) will have missing values. Analogously, $\|x_i - \hat{x}_i\|^2$ can be computed as follows. Let we_i be a vector of length r with zeros in the positions where there are missing values in x_i and $r/(r - r_i)$ otherwise. Note that if there are no missing values, then we is a vector of ones. In this way, the summands of RSS can be computed as $\|x_i. * we_i - \hat{x}_i\|^2$.

Unlike the previous attempt by Mørup and Hansen (2012) to compute AA with missing values, where the objective function in AA was modified, in our proposal the optimization problem with missing data is the same as Eq. 3. To solve AA problem with missing values, we modify the original alternating minimization algorithm described by Eugster and Leisch (2009) to handle the missing values. The outline of the algorithm is for a given number of archetypes p:

1. Preprocessing and initialization: build W and initialize β and α as made by Eugster and Leisch (2009). Calculate RSS as explained above.
2. Loop until RSS reduction is sufficiently small or the number of maximum iterations is reached.
   (a) Find best α for the given set of archetypes Z: solve n convex least squares problems ($i = 1, \ldots, n$). Note that each problem is independent of the rest, and its result can be computed by excluding the coordinates with missing values
   $$min_{\alpha_i} \|x_i - \alpha_i Z\|_2$$
   subject to $\alpha_i \geq 0$ and $\sum_{j=1}^{p} \alpha_{ij} = 1$.
   (b) Recalculate archetypes Z: solve the system of linear equations $X. * W = \alpha Z$.
   (c) Find best β for the given set of archetypes Z: solve p convex least squares problems ($j = 1, \ldots, p$).
   $$min_{\beta_j} \|z_j - \beta_j (X. * W)\|_2$$
   subject to $\beta_j \geq 0$ and $\sum_{l=1}^{n} \beta_{jl} = 1$.
   (d) Recalculate archetypes Z: $Z = \beta(X. * W)$
   (e) Calculate RSS.

Details about how to solve the numerical problems, such as the systems of linear equations and the convex least squares problems, can be found in Eugster and Leisch (2009). Note also that not only is RSS computation adapted to handle missing values, but instead of the spectral norm used by Eugster and Leisch (2009), the Frobenius norm is used as in the original AA definition (Cutler and Breiman (1994)).

In the supplementary material (Online Resource 1), we compare the results of our proposal versus those obtained using the methodology put forward by Mørup and Hansen (2012). We also carried out another comparison between the results with our proposal and those obtained by removing the cases with missing data and with imputations. The results show that our new procedure is the best alternative of those considered.

4.2 ADA with missing values in the multivariate case

Since archetypes (Z) do not have missing values, only the complete cases will be considered as possible archetypoids (z_j). In this way, RSS can be computed as explained in Section 4.1. In other words, l in equation (4) and the respective constraint 2) is only defined for the indices that correspond to the complete cases (ICC), i.e. $z_j = \sum_{l \in ICC} \beta_{jl} x_l$.

To solve ADA problem with missing values, we modify the original algorithm proposed by Vinué et al (2015a) to handle the missing values. The outline of the algorithm is for a given number of archetypoids p:

1. BUILD phase: look for a good initial set of p archetypoids from the complete data points.
2. SWAP phase: For each archetypoid a
   (a) For each non-archetypoid data point o from the complete data set
      i. Swap a and o and calculate the RSS of the configuration as explained in Section 4.1 (α coefficients must be calculated as indicated in Section 4.1).
3. Select the configuration with the lowest RSS.
4. Repeat steps 2 to 4 until there is no change in the archetypoids.

The initial set of archetypoids in the BUILD phase can be determined by the same strategies explained in Section 2.1, but taking into account that only the complete cases are considered as possible archetypoids.

4.3 AA and ADA with missing landmarks

Let X_1, ..., X_n be n landmark configuration matrices, with k landmarks in dimension m, as before, but this time allowing missing landmarks. Let us suppose that there is at least one complete configuration. The procedure proposed is very similar to that explained in section 2.3, but the Procrustes mean shape of the complete configurations is used as a pole, following the same idea as Arbour and Brown (2014).

Specifically the procedure is as follows:

Firstly, the mean shape and tangent coordinates of complete cases are obtained using centered coordinates instead of the Helmert matrix in order to remove location. We cannot use the Helmert matrix to remove location when we have missing landmarks, because the pre-shape resulting from applying this matrix would have all its landmarks missing.

Secondly, we obtain the coordinates of each case with missing landmarks on the tangent space at the mean shape of complete cases using the available landmarks (Rohlf (1999)). As a result, the matrix V with the vectors of Procrustes tangent coordinates now have missing values.

Next, multivariate AA with missing values is applied and the archetypes are computed in the tangent space.

Finally, they are projected back into the configuration.

The procedure is analogous for ADA, but considering multivariate ADA.

A comparative analysis of the impact of missing landmarks in AA for different alternatives is carried out in the supplementary material (Online Resource 1).

5 Application

The aim of this section is to show how the aforementioned methods can be used to identify extreme shapes by means of archetypal shapes. These will be the children with fitting problems in the actual sizing system and identifying them can be useful for the apparel design application.

A randomly selected sample of 502 Spanish children aged 6 to 12 years old was scanned using a Vitus Smart 3D body scanner from Human Solutions. The children were asked to wear a standard white garment in order to standardize the measurements. The body shape of each child in our data set was represented by 3075 3D landmarks, i.e., by a 3075×3 configuration matrix. From the 3D mesh, several anthropometric measurements were calculated semi-automatically by combining automatic measurements based on characteristic geometric points with a manual review.

Nowadays, a sizing system shows the range of body measurements for each key dimension. The body measurements that are covered by a standard garment sizing system differ from one country to another. The key dimensions that are most often used are chest girth, waist girth and height for men's garments; bust girth, waist girth, hip girth and height for women's garments; and apparel sizes for children are usually designated by sex and height or age. So, children aged between 6 and 12 are usually classified into three sizes per sex, and clothing designers use tables that list ranges of values of the main anthropometric measurements for each size. For most of the commercial tables these variables are: Sex, Age, Height, Chest girth, Waist girth and Hip girth, with the values shown in Table 1 according to Guerrero and ASEPRI (2000).

Table 1 Commercial measurements for sizes 8, 10 and 12, for boys and girls, respectively. Approximate age is given in years, and the other measurements are taken in cm.

Sex	Boys			Girls		
Age	6-8	8-10	10-12	6-8	8-10	10-12
Height	$[116, 128)$	$[128, 140)$	$[140, 152)$	$[116, 128)$	$[128, 140)$	$[140, 152)$
Chest	[60.0-64.0)	[64.0-68.0)	[68.0-74.0)	[60.0-64.0)	[64.0-68.0)	[68.0-73.0)
Waist	[56.0-59.0)	[59.0-62.0)	[62.0-66.0)	[54.0-56.0)	[56.0-59.0)	[59.0-62.0)
Hip	[66.0-71.0)	[71.0-76.0)	[76.0-81.0)	[66.0-71.0)	[71.0-76.0)	[76.0-81.5)

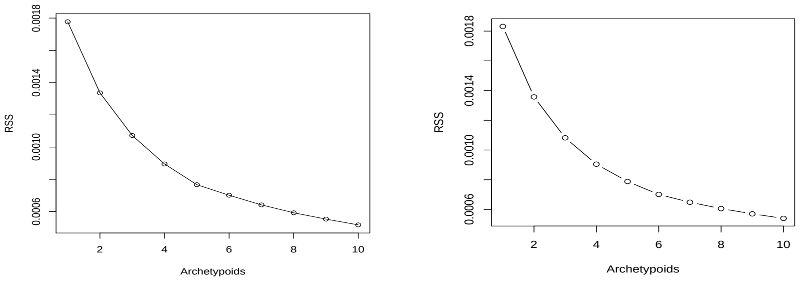
The children in our data set have been grouped into six groups according to their sex and the height intervals defined in the sizing system shown in Table 1. The sample size for each group is shown in Table 2. In order to illustrate our methodology, we are going to look for archetypoids in the youngest groups of boys and girls, where we expect to find a great variety of shapes, as this size covers children ranging from tall preschoolers to early pubescent individuals. Note that most children have a slimmer appearance during middle childhood than they did during preschool years.

Procrustes distances to the mean are computed. Some children could be considered outliers, but we have decided not to remove them, as they are part of the population variability. If we were more interested in the archetypes of the majority than of the totality, those outlier children could be removed before the analysis, as in Section 3. In the same way, if we want to accommodate a certain percentage

Table 2 Sample sizes.

Sex \ Height	$[116, 128)$ cm	$[128, 140)$ cm	$[140, 152)$ cm
Boy	61	116	74
Girl	84	99	68

of the population, then only an appropriate part of the sample should be used. As stated in Section 2.1 the number p of archetypes or archetypoids to compute is selected as the point where the elbow on the RSS representation for a series of different p values is found. As an illustration, Figure 8 shows the RSS representation for different numbers of archetypoids for the groups of younger boys and girls. In these figures, no clear elbow can be seen, so, according to a garment designer expert, four or five would be a reasonable number for design purposes (a large number of representative cases may overwhelm the designer and thus be counterproductive Epifanio et al (2013)). We decided to find five archetypoids per group. Although archetypoids are not necessarily nested, in this problem when the number of archetypoids is increased from $p = 4$ to $p = 6$, the patterns discovered with smaller p values have been kept, and increasing p has led to the discovery of new finer patterns. So, the decision about the number of archetypoids can simply be based on the most suitable option for the garment designer.



(a) (b)

Fig. 8 (a) RSS representations for the boys in height group $[116, 128)$ cm; (b) RSS representations for the girls in height group $[116, 128)$ cm.

The results for the first group (boys with heights ranging between $[116, 128)$ cm) are shown in Figure 9 and Table 3. The numbers in bold in Table 3 show measurements that lie outside the "normal" limits considered in the sizing system (Table 1). In Figure 10, the points representing the archetypoids have been labeled in red numbers, and green lines have been plotted to indicate the 'fit' ranges established by the sizing system that correspond to boys in this height range (Table 1).

In the boys group, the five archetypoids correspond to the boys labeled as $KID531$ (archetypoid B1), $KID435$ (archetypoid B2), $KID576$ (archetypoid B3), $KID532$ (archetypoid B4) and $KID773$ (archetypoid B5).

The measurements of archetypoid B5 in the main dimensions (chest, waist and hip contour) used for determining the size besides height are very large with

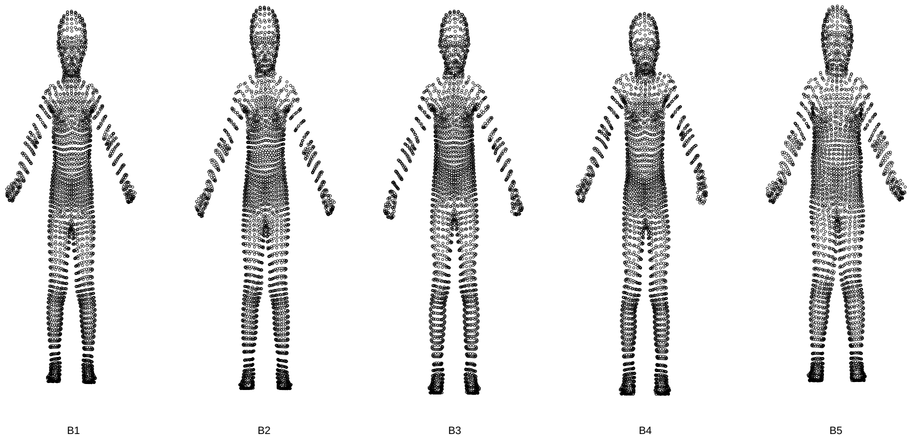


Fig. 9 Archetypoids found in group 1 (boys with height between $[116, 128)$ cm), labeled as B1, B2, B3, B4 and B5.

Table 3 Anthropometric variables commonly used in the clothing design process for the five archetypoids found in group 1 (boys with height between $[116, 128)$ cm).

	Boys				
Archetypoid	B1	B2	B3	B4	B5
Label	$KID531$	$KID435$	$KID576$	$KID532$	$KID773$
Age	7.4	6.9	8.6	7.5	6.6
Height	123.4	127.2	126.7	125.6	125.2
Chest circumf.	60.6	63.9	62.0	**59.0**	**79.8**
Waist circumf.	**53.3**	56.5	**54.9**	**54.0**	**76.5**
Hip circumf.	67.2	68.0	67.0	**65.9**	**79.5**

respect to the standards of his supposed size according to his height. Note that his chest, waist and hip contours are $79.8 > 64$, $76.5 > 59$ and $79.5 > 71$, respectively. His hip contour corresponds to two sizes above that suggested by his height (Table 1). According to his chest and hip contours, this boy would need to go three or four sizes up from the size indicated by his height. His clothes would therefore look extremely tight with a size 8. Figure 10 clearly shows that this child has the largest chest and waist contours of all the boys included in this height group, and one of the largest hip contours in the group, without being one of the tallest boys. Additionally, if the hip/waist ratio is computed, this child shows the lowest value in his group (Figure 11 (c)).

His opposite archetypoid would be archetypoid B4, whose main dimensions (besides height) are smaller than those supposed for his size according to his height (Figure 10). Therefore, for size 8, which corresponds to his height, the clothes will look loose on this child. Additionally, his shoulder length (8.4) is below the limits of this dimension for size 8.

On the other hand, both archetypoids B1 and B3, respectively, have a waist contour below the limits of their supposed size according to their height (Figure 10). This dimension is especially important for trousers. The chest and hip contours are within the limits. However, archetypoid B3 shows values of crotch height

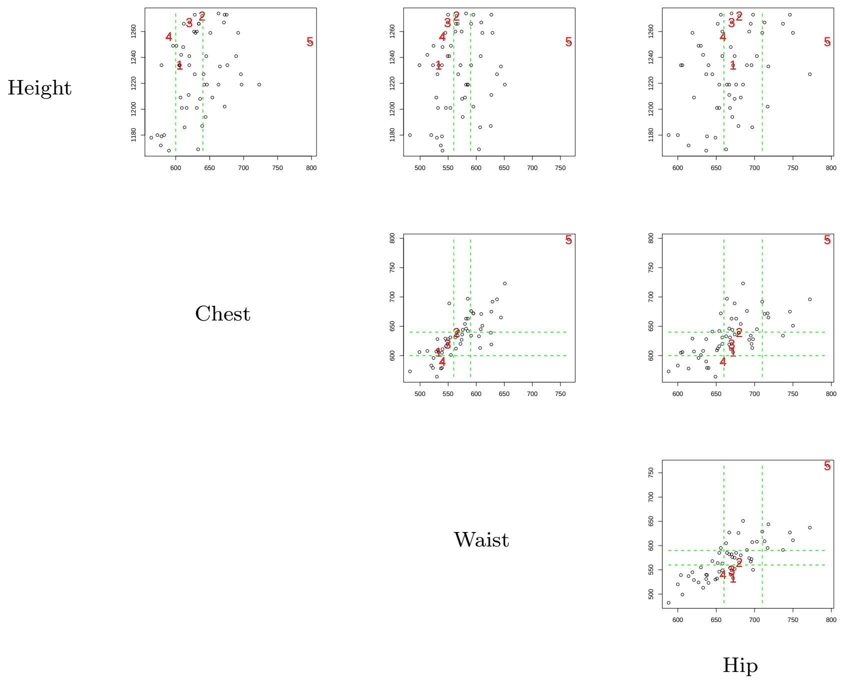
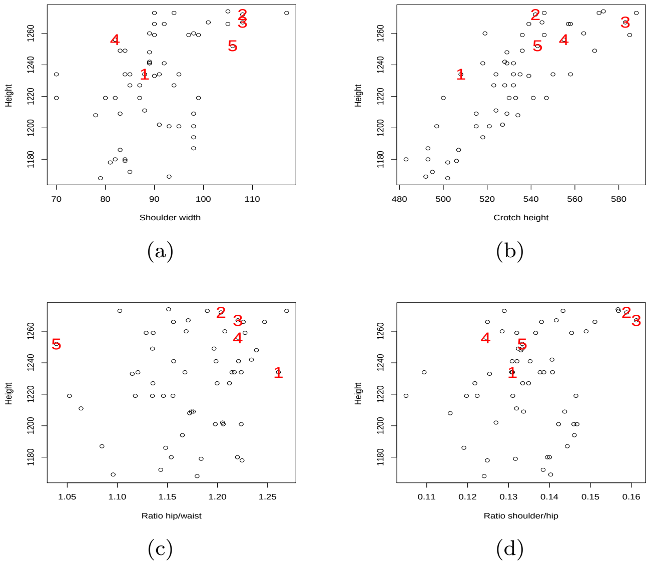


Fig. 10 Scatterplots of the main anthropometric measurements for designing clothes for all the boys with heights between [116, 128) cm. Points corresponding to archetypoid measurements are labeled in red numbers. Green lines indicate the ‘fit’ ranges established by the sizing system.

(58.3), knee height (35.1) and shoulder width (10.85) above the limits of his size according to his height and one of the largest values in these variables in his height group (Figure 11 (a) and (b)). So he is a long legged boy with broad shoulders. Archetypoid B1 presents one of the highest ratios between hip and waist in his height group (Figure 11 (c)), so he is one of the straightest boys with small differences between his hip and waist contour. Unlike B3, the crotch height (50.8), arm (40.2) and shoulder (89) lengths of B1 are below the limits of his supposed size according to his height.

Finally, for archetypoid B2, the main dimensions (chest, waist and hip contours) are within the limits of size 8, which corresponds to his height. However, although his basic dimensions are all within the limits, he has broad shoulders (above the limits of his size) in relation to his hip contour (Figure 11 (a) and (d)), like archetypoid B3, but with shorter legs (Figure 11 (b)).

Archetypoids found for the sample of girls with heights ranging between [116, 128) cm are shown in Figure 12 and Table 4. They are labeled as G1, G2, G3, G4 and G5. In Figure 13, the points representing the archetypoids have been labeled in red numbers, and green lines have been plotted to indicate the ‘fit’ ranges established by the sizing system that correspond to girls in this height range (Table 1). The numbers in bold in Table 4 show measurements that lie outside the “normal” limits considered in the sizing system (Table 1).



(a) (b)

(c) (d)

Fig. 11 Scatterplots of different measurements and ratios of measurements of all the boys with heights between [116, 128) cm. Points corresponding to archetypoid measurements are labeled in red numbers. (a) scatterplot between shoulder width and height; (b) scatterplot between leg length and height; (c) shows the relationship between hip/waist ratio and total height of each boy, and finally (d) shows the relationship between shoulder/hip ratio and total height.

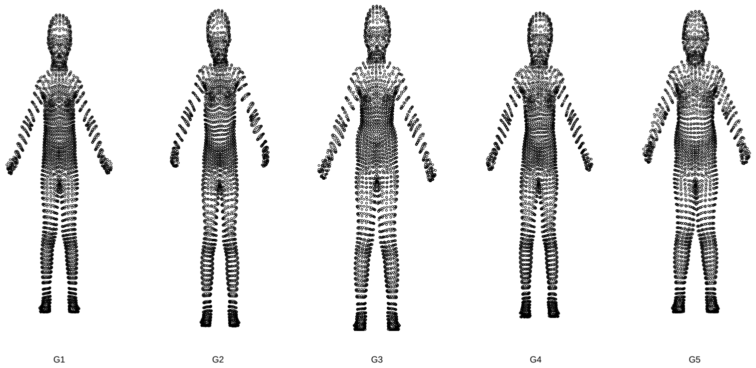


Fig. 12 Archetypoids found in the group of girls with height between [116, 128) cm.

In the girls group, archetypoids G1 and G5 are "short" girls (Figure 13), with short legs in realtion to their total height (Figure 14 (c)) but with different shapes. Archetypoid G1 has smaller chest and hip contours than those established for her height size, while archetypoid G5 has one of the largest chest, hip and waist contours in this group. Her measurements are very large with respect to the standards of her supposed size according to her height. This is clearly shown in Figure 13.

Table 4 Anthropometric measurements of archetypoids in group 2 (girls with height: $[116, 128)$ cm).

	Girls				
Archetypoid	G1	G2	G3	G4	G5
Label	$EKID156$	$KID333$	$KID553$	$KID037$	$KID220$
Age	7.1	6.2	8.2	7.2	6.3
Height	117.2	124.1	127.4	120.2	118.6
Chest circumf.	**58.3**	**57.8**	**67.2**	61.0	**78.1**
Waist circumf.	54.0	**53.8**	**60.6**	**50.4**	**75.3**
Hip circumf.	**64.0**	**62.4**	**74.5**	**65.2**	**78.2**

Three of the archetypoids (G1, G2 and G4) show girls with smaller chest, waist and/or hip contours than those established for their height interval, but with clear height differences. G2 is taller than G4, who is also taller than G1. In fact, G1 is one of the shortest girls in her group (Figure 13). G1 and G2 show narrow chest contours (Figure 13), while G4 is one of the girls with the narrowest waist contour and with the shortest arms in relation to her height (Figure 14 (c)). G4 has a very low waist/chest ratio (Figure 14 (b)), this is not the case with G1 and G2. There are also differences between these three archetypoids in crotch height/total height ratio (Figure 14 (c)). More differences between G1 and G2 are found in neck-related measurements.

On the other hand, G3 and G5 are girls with chest, waist and hip contours greater than those established for their height interval. Archetypoid G3 is one of the tallest girls in the group, and has long arms and long legs in relation to her total height (Figures 13 and 14 (c) and (d)), while archetypoid G5 is just the opposite, with low values in several non-basic dimensions such as crotch height/total height ratio and hip/waist ratio. The crotch height (49.8) for G5 is below the limits of her size according to her height, while it is above the limits for G3 (58.7).

We have found children with very different shapes who are supposedly in the same size. This causes garment fitting problems. Having identified the extremes of a size, and together with a central case that represents the basic proportions in a range of clothing, the apparel grading process within that size can begin. In this way, designers can know which adjustments are needed to accommodate individuals in a certain size. We propose to use the selected cases for saving costs. This strategy was also adopted for women by Vinué et al (2015a), but using dissimilarities between trunk forms to find the archetypoids. The importance of considering real people versus virtual people in sustainable sizing is highlighted by Robinette and Veitch (2016). Finally, note that the objective is not to find subsizes, but to accommodate children within a specific size. Clustering algorithms could be used to define sizes (Ibáñez et al (2012); Vinué et al (2016)).

6 Conclusions

It has been proposed that AA and ADA could be extended from the multivariate case to landmark-based data. A comparison with other unsupervised techniques commonly used in shape analysis (PCA, SPCA and clustering) has been carried out using the well-known digit 3s data set for illustration purposes. AA and ADA with landmarks can be applied for the same purposes as in the multivariate case.

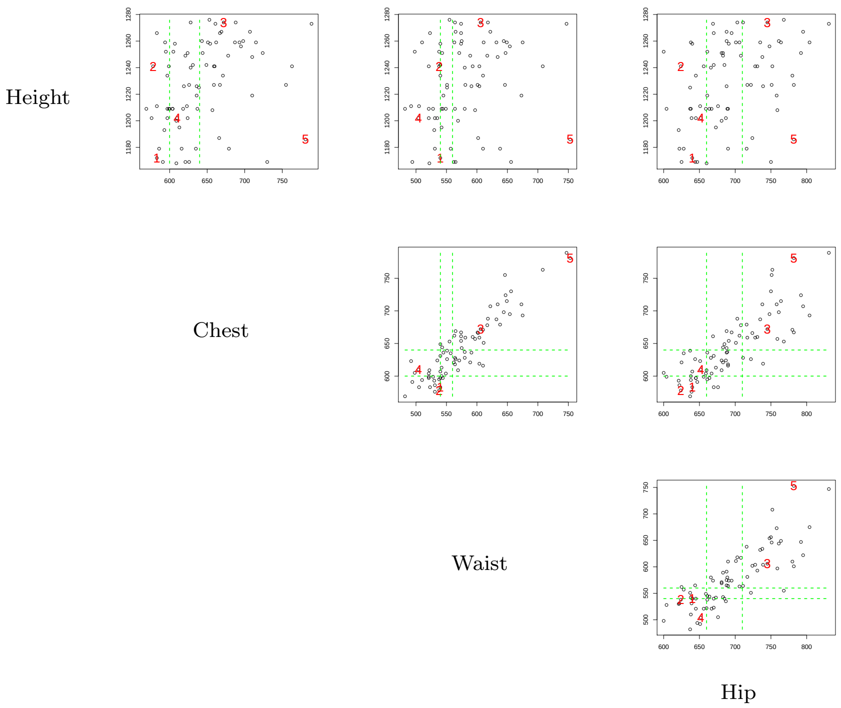
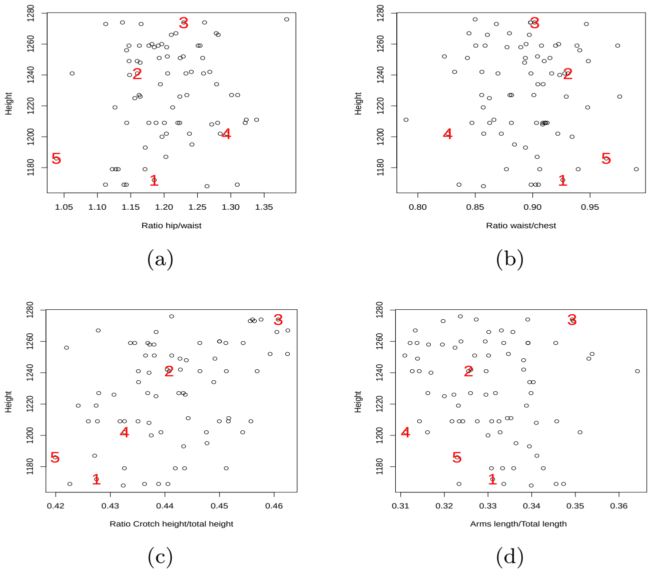


Fig. 13 Scatterplots of the main anthropometric measurements for designing clothes for all the girls with heights between $[116, 128)$ cm. Points corresponding to archetypoid measurements are labeled in red numbers. Green lines indicate the ‘fit’ ranges established by the sizing system.

A new tool for visualizing the information returned with archetypal shapes has been also proposed, based on a purity-based order. A procedure for computing AA and ADA with missing values has been also proposed in the multivariate case, and is also extended to the case of missing landmarks. A comparison with different alternatives has been carried out.

Archetypal shapes for children have been obtained for the apparel design application. However, they can also be obtained for the design of other products for children. Furthermore, instead of children, other populations could be of interest and the methodology could be applied to the appropriate database in ergonomic design or another application. In particular, AA and ADA with landmarks could have great potential in biological and medical applications, as discussed in Section 1.

More directions for future work could be as follows: firstly, in many situations landmarks are not the only descriptors of the observations, but also multivariate variables. For example, color is also important besides shapes by MacLeod (2015). In that case, the objective function in equations (3) and (4) should be modified to take in both sets of features. Once the shapes are represented in the tangent space, the information of both vectorial spaces could be (weighted) combined using an appropriate interior product to build the corresponding RSS. Secondly, a robust version for AA and ADA with landmarks could be used. In the illustrative example



(a)

(b)

(c)

(d)

Fig. 14 Scatterplots of different ratios of measurements for all the girls with heights between $[116, 128)$ cm. Points corresponding to archetypoid measurements are labeled in red numbers. (a) shows the ratio of hip and waist contours to height; (b) shows the ratio of waist and chest contours to height; (c) shows the relationship between leg length and total height of each girl, and finally (d) shows the ratio between arm length and total height.

with digit 3s, an outlier was removed prior to performing the analysis, as made by Dryden and Mardia (1998). Instead of this strategy, the methodology proposed by Eugster and Leisch (2011) could be followed. Thirdly, other multivariate techniques could be extended to shapes, such as fuzzy versions of k-means or k-medoids. Finally, we have extended AA to missing data without erasing incomplete cases or resorting to imputation, and similar strategies could be studied for other statistical techniques.

Acknowledgements We would like to thank the Biomechanics Institute of Valencia for providing us with the data set.

References

Arbour JH, Brown CM (2014) Incomplete specimens in geometric morphometric analyses. Methods in Ecology and Evolution 5(1):16–26

Ayala G, Epifanio I, Simó A, Zapater V (2006) Clustering of spatial point patterns. Computational Statistics & Data Analysis 50(4):1016–1032

Bookstein F (1978) Lecture notes in biomathematics. In: The measurement of biological shape and shape change, Springer-Verlag

Brown CM, Arbour JH, Jackson DA (2012) Testing of the effect of missing data estimation and distribution in morphometric multivariate data analyses. Systematic Biology 61(6):941–954

Canhasi E, Kononenko I (2013) Multi-document summarization via archetypal analysis of the content-graph joint model. Knowledge and Information Systems pp 1–22, DOI 10.1007/s10115-013-0689-8

Canhasi E, Kononenko I (2014) Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. Expert Systems with Applications 41(2):535 – 543

Chan B, Mitchell D, Cram L (2003) Archetypal analysis of galaxy spectra. Monthly Notices of the Royal Astronomical Society 338

Claude J (2008) Morphometrics with R. Springer, New York

Cutler A, Breiman L (1994) Archetypal Analysis. Technometrics 36(4):338–347

Davis T, Love B (2010) Memory for category information is idealized through contrast with competing options. Psychological Science 21(2):234–242

D'Esposito MR, Ragozini G (2008) A New R-Ordering Procedure to Rank Multivariate Performances. Quaderni di Statistica 10:22–40

D'Esposito MR, Palumbo F, Ragozini G (2012) Interval Archetypes: A New Tool for Interval Data Analysis. Statistical Analysis and Data Mining 5(4):322–335

Dryden IL (2015) shapes: Statistical Shape Analysis. URL `https://CRAN.R-project.org/package=shapes`, R package version 1.1-11

Dryden IL, Mardia KV (1998) Statistical Shape Analysis. John Wiley & Sons, Chichester

Dryden IL, Mardia KV (2016) Statistical Shape Analysis: With Applications in R. John Wiley & Sons, Chichester

Dryden IL, Zempléni A (2006) Extreme shape analysis. Journal of the Royal Statistical Society Series C 55(1):103–121

Du J, Dryden IL, Huang X (2015) Size and shape analysis of error-prone shape data. Journal of the American Statistical Association 110(509):368–379

Eirola E, Doquire G, Verleysen M, Lendasse A (2013) Distance estimation in numerical data sets with missing values. Information Sciences 240:115 – 128

Eneh S (2015) Showroom the future of online fashion retailing 2.0: Enhancing the online shopping experience. Master's thesis, University of Borås, Faculty of Textiles, Engineering and Business

Epifanio I (2016) Functional archetype and archetypoid analysis. Computational Statistics & Data Analysis 104:24 – 34

Epifanio I, Vinué G, Alemany S (2013) Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem. Computers & Industrial Engineering 64(3):757–765

Eugster MJ, Leisch F (2009) From Spider-Man to Hero - Archetypal Analysis in R. Journal of Statistical Software 30(8):1–23

Eugster MJA (2012) Performance profiles based on archetypal athletes. International Journal of Performance Analysis in Sport 12(1):166–187

Eugster MJA, Leisch F (2011) Weighted and robust archetypal analysis. Computational Statistics & Data Analysis 55(3):1215–1225

Fréchet M (1948) Les éléments aléatoires de nature quelconque dans un espace distancié. Annales de l'Institut Henri Poincaré Probabilités et Statistiques 10(4):215–310

Goodall C (1991) Procrustes methods in the statistical analysis of shape. Journal of the Royal Statistical Society Series B (Methodological) pp 285–339

Guerrero J, ASEPRI (2000) Estudio de tallas y medidas de la población infantil internacional. Asociación Española de Fabricantes de Productos para la Infancia

(ASEPRI)

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Data mining, inference and prediction. 2nd ed., Springer-Verlag, New York

Hinrich JL, Bardenfleth SE, Roge RE, Churchill NW, Madsen KH, Mørup M (2016) Archetypal analysis for modeling multisubject fMRI data. IEEE Journal on Selected Topics in Signal Processing 10(7):1160–1171

Ibáñez MV, Vinué G, Alemany S, Simó A, Epifanio I, Domingo J, Ayala G (2012) Apparel sizing using trimmed PAM and OWA operators. Expert Systems with Applications 39(12):10,512 – 10,520

Kaufman L, Rousseeuw PJ (1990) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, New York

Kendall D (1984) Shape manifolds, Procrustean metrics, and complex projective spaces. London Math Soc 16:81–121

Kendall DG, Barden D, Carne T, Le H (2009) Shape and shape theory. John Wiley & Sons, Chichester

Lawson CL, Hanson RJ (1974) Solving Least Squares Problems. Prentice Hall, Englewood Cliffs

Li S, Wang P, Louviere J, Carson R (2003) Archetypal Analysis: A New Way To Segment Markets Based On Extreme Individuals. In: ANZMAC 2003 Conference Proceedings, pp 1674–1679

MacLeod N (2015) Proceedings of the Third International Symposium on Biological Shape Analysis, World Scientific, Singapore, chap The direct analysis of digital images (eigenimage) with a comment on the use of discriminant analysis in morphometrics, pp 156–182

Midgley D, Venaik S (2013) Marketing strategy in MNC subsidiaries: pure versus hybrid archetypes. In: P. McDougall-Covin and T. Kiyak, Proceedings of the 55th Annual Meeting of the Academy of International Business, pp 215–216

Mørup M, Hansen LK (2012) Archetypal analysis for machine learning and data mining. Neurocomputing 80:54–63

Pennec X (2006) Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. Journal of Mathematical Imaging and Vision 25(1):127–154

Porzio GC, Ragozini G, Vistocco D (2008) On the use of archetypes as benchmarks. Applied Stochastic Models in Business and Industry 24:419–437

R Development Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, ISBN 3-900051-07-0

Ragozini G, D'Esposito MR (2015) Archetypal networks. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ACM, New York, NY, USA, pp 807–814

Ragozini G, Palumbo F, D'Esposito MR (2017) Archetypal analysis for data-driven prototype identification. Statistical Analysis and Data Mining: The ASA Data Science Journal 10(1):6–20

Robinette KM, Veitch D (2016) Sustainable sizing. Human Factors: The Journal of the Human Factors and Ergonomics Society 58:657–664

Rohlf FJ (1998) On applications of geometric morphometrics to studies of ontogeny and phylogeny. Systematic Biology 47(1):147–158

Rohlf FJ (1999) Shape statistics: Procrustes superimpositions and tangent spaces. Journal of Classification 16(2):197–223

Seth S, Eugster MJA (2016a) Archetypal analysis for nominal observations. IEEE Trans Pattern Anal Mach Intell 38(5):849–861

Seth S, Eugster MJA (2016b) Probabilistic archetypal analysis. Machine Learning 102(1):85–113

Sjöstrand K, Stegmann MB, Larsen R (2006) Sparse principal component analysis in medical shape modeling. In: International Symposium on Medical Imaging 2006, San Diego, CA, USA, The International Society for Optical Engineering (SPIE), vol 6144

Sjöstrand K, Rostrup E, Ryberg C, Larsen R, Studholme C, Baezner H, Ferro J, Fazekas F, Pantoni L, Inzitari D, Waldemar G (2007) Sparse decomposition and modeling of anatomical shape variation. IEEE Transactions on Medical Imaging 26(12):1625–1635

Slice DE (2001) Landmark coordinates aligned by Procrustes analysis do not lie in Kendall's shape space. Systematic Biology 50(1):141–149

Stoyan LA, Stoyan H (1995) Fractals, Random Shapes and Point Fields. John Wiley and Sons, Chichester

Theodosiou T, Kazanidis I, Valsamidis S, Kontogiannis S (2013) Courseware usage archetyping. In: Proceedings of the 17th Panhellenic Conference on Informatics, ACM, New York, NY, USA, PCI '13, pp 243–249

Thøgersen JC, Mørup M, Damkiær S, Molin S, Jelsbak L (2013) Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. BMC Bioinformatics 14:279

Thurau C, Kersting K, Wahabzada M, Bauckhage C (2012) Descriptive matrix factorization for sustainability: Adopting the principle of opposites. Data Mining and Knowledge Discovery 24(2):325–354

Tsanousa A, Laskaris N, Angelis L (2015) A novel single-trial methodology for studying brain response variability based on archetypal analysis. Expert Systems with Applications 42(22):8454 – 8462

Vinué G (2017) Anthropometry: An R package for analysis of anthropometric data. Journal of Statistical Software 77(6):1–39

Vinué G, Epifanio I (2017) Archetypoid analysis for sports analytics. Data Mining and Knowledge Discovery pp 1–35, DOI 10.1007/s10618-017-0514-1

Vinué G, Epifanio I, Alemany S (2015a) Archetypoids: A new approach to define representative archetypal data. Computational Statistics & Data Analysis 87:102 – 115

Vinué G, Epifanio I, Simó A, Ibáñez M, Domingo J, Ayala G (2015b) Anthropometry: An R Package for Analysis of Anthropometric Data. R package version 1.5

Vinué G, Simó A, Alemany S (2016) The k-means algorithm for 3D shapes with an application to apparel design. Advances in Data Analysis and Classification 10(1):103–132

Viscosi V, Cardini A (2011) Leaf morphology, taxonomy and geometric morphometrics: A simplified protocol for beginners. PLOS ONE 6(10):1–20

Zapater V, Martínez-Costa L, Ayala G, Domingo J (2002) Classifying human endothelial cells based on individual granulometric size distributions. Image and Vision Computing 20(11):783–791

Zou H, Hastie T (2012) elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. URL `http://CRAN.R-project.org/package=elasticnet`, R package version 1.1