



---

**Título artículo / Títol article:** Archetypoids: A new approach to define representative archetypal data

**Autores / Autors** Guillermo Vinué Visús, Irene Epifanio López, Sandra Alemany Mut

**Revista:** Computational Statistics & Data Analysis

**Versión / Versió:** Versión Pre-print

**Cita bibliográfica / Cita bibliogràfica (ISO 690):** VINUÉ VISÚS, Guillermo; EPIFANIO LÓPEZ, Irene; ALEMANY MUT, Sandra. Archetypoids: A new approach to define representative archetypal data. 2015. *Computational Statistics & Data Analysis*, 2015, vol. 87, p. 102-115

**url Repositori UJI:** <http://hdl.handle.net/10234/116382>

# Archetypoids: a new approach to define representative archetypal data <sup>☆</sup>

Guillermo Vinué<sup>a</sup>, Irene Epifanio<sup>b,\*</sup>, Sandra Alemany<sup>c</sup>

<sup>a</sup>*Department of Statistics and O.R., University of Valencia, 46100 Burjassot, Spain*

<sup>b</sup>*Dept. Matemàtiques. Campus del Riu Sec. Universitat Jaume I, 12071 Castelló, Spain*

<sup>c</sup>*Biomechanics Institute of Valencia, 46022 Valencia, Spain*

---

## Abstract

The new concept archetypoids is introduced. Archetypoid analysis represents each observation in a dataset as a mixture of actual observations in the dataset, which are pure type or archetypoids. Unlike archetype analysis, archetypoids are real observations, not a mixture of observations. This is relevant when existing archetypal observations are needed, rather than fictitious ones. An algorithm is proposed to find them and some of their theoretical properties are introduced. It is also shown how they can be obtained when only dissimilarities between observations are known (features are unavailable). Archetypoid analysis is illustrated in two design problems and several examples, comparing them with the archetypes, the nearest observations to them and other unsupervised methods.

*Keywords:* Archetype, Convex hull, Unsupervised learning, Extremal point, Non-negative matrix factorization.

---

## 1. Introduction

There are problems where it is fundamental to find the extreme individuals of a sample. Archetypal analysis is a very useful tool for this pur-

---

<sup>☆</sup>A pdf file is included with supplementary material. The code and data for reproducing the examples are available at <http://www.uv.es/vivigui/software.html> and they form part of the R package **Anthropometry**.

\*Tel.: +34 964728390; fax: +34 964728429.

*Email addresses:* [Guillermo.Vinue@uv.es](mailto:Guillermo.Vinue@uv.es) (Guillermo Vinué), [epifanio@uji.es](mailto:epifanio@uji.es) (Irene Epifanio), [sandra.alemany@ibv.upv.es](mailto:sandra.alemany@ibv.upv.es) (Sandra Alemany)

pose. Archetypes were defined in [Cutler and Breiman \(1994\)](#) and they have been applied in different fields such as market research ([Li et al. \(2003\)](#), [Porzio et al. \(2008\)](#), [Midgley and Venaik \(2013\)](#)), biology ([D’Esposito et al. \(2012\)](#)), genetics ([Thøgersen et al. \(2013\)](#)), sports ([Eugster \(2012\)](#)), industrial engineering (ergonomic design and evaluation) ([Epifanio, Vinué, and Alemany \(2013\)](#)), the evaluation of scientists ([Seiler and Wohlrabe \(2013\)](#)), astrophysics ([Chan et al. \(2003\)](#), [Richards et al. \(2012\)](#)), e-learning ([Theodosiou et al. \(2013\)](#)), multi-document summarization ([Canhasi and Kononenko \(2013, 2014\)](#)) and different machine learning problems ([Mørup and Hansen \(2012\)](#), [Stone \(2002\)](#)). The archetypes returned by archetypal analysis are a convex combination of the sampled individuals, but they are not necessarily observed individuals. However, in certain problems it is crucial that the archetypes are real subjects, that is, observations of the sample. For example, [Seiler and Wohlrabe \(2013\)](#) considered the case of finding archetypal economists and, using archetypal analysis, found that in some cases: “the identified archetypes are artificial, i.e. no economist in our sample fits this archetype to 100%.”. A new archetypal concept is introduced to tackle this problem: the archetypoid, which is a real (observed) archetype.

Human modeling is widely used in automotive engineering, aviation, manufacturing and defense industries, amongst others. 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. A case represents a set of body dimensions we plan to accommodate in design. There are three types of cases ([HFES 300 Committee \(2004\)](#)) according to the location of the cases: central (located toward the center of the distribution of the body dimensions selected), boundary (located toward the edges of the distribution) and distributed cases (spread throughout the distribution of the body dimensions). The objective in our design problems is to obtain the boundary cases. We are assuming that the adjustable components (for instance, the seat, rudder, etc. in an aircraft cockpit design) in each problem can be adjusted in sufficiently small increments. Therefore, we are assuming that the accommodation of boundaries ensures the accommodation of interior points. One of the advantages of considering boundary cases is that a large range of accommodation is achieved while using a relatively small number of cases. For example, [Bittner et al. \(1987\)](#) showed that by using only 17 cases (16 boundary and 1 centroid) they were able to achieve the same population accommodation percentage as with 400 distributed cases.

When designing a workspace it has typically been a requirement that between 90 and 95 percent of the relevant population are accommodated. In traditional workspace evaluation, at a later stage in the design process, a mock-up of the workspace is built and assessed with mannequins or, even better, with “live” test subjects (Rothwell and Hickey (1986)). If any problems are found, the mock-up has to be modified or even a new mock-up has to be built. The cost of a mock-up is extremely expensive, in terms of both time and economic costs (Blanchonette (2010)). Note that when a new car model is being developed, many mock-ups have traditionally been built, and each mock-up costs between \$500,000 and \$1,000,000 (Brown (1999)). Furthermore, a large number of individuals are used for assessing the mock-ups (about 30 individuals in a cockpit design (Kennedy and Zehner (1995))). If the “hard-to-fit” extreme individuals (the boundary cases) were previously identified, the design could be improved at the beginning and the time and cost of the design process would be reduced, as well as the number of “live” individuals needed for assessing the mock-ups. In this problem, we also need to identify the archetypoids.

Another real and immediate application of archetypoids in ergonomic product design is related to workplace adaptation in manufacturing and production companies. These companies maintain comprehensive databases with information about their employees. When they perform an ergonomic study aimed at designing a new production line, a small sample of workers is selected to support the system validation. The computation of archetypoids allows us to select this small number of representative workers.

In some situations (Hastie, Tibshirani, and Friedman (2009, Sect. 16.5)) we only know the dissimilarities between the observations, i.e. features (variables) are unavailable. In this case, it is imperative that the archetypes are observations of the sample, otherwise, we cannot define a mixture of objects without having access to feature vectors. The archetypoids always exist, even when the data are only a collection of dissimilarities. With abstract data objects, like proteins or images, archetypoids have a practical advantage over archetypes: an archetypoid is one of the observations, and can be displayed, which aids their interpretation.

The outline of the paper is as follows: In Section 2 we review archetype analysis, we introduce archetypoid analysis and the algorithm for computing archetypoids and we explain how to calculate archetypoids when features are unavailable. In Section 3 we discuss some of the properties of archetypoids and carry out several comparisons with other unsupervised methods.

In Section 4, our proposal is applied to a cockpit design problem used in [Epifanio et al. \(2013\)](#) and to an apparel design problem, focusing on the case where features are not available. The implementation of our proposal is written in R ([R Development Core Team \(2013\)](#)) and is available together with the code for reproducing the examples at <http://www.uv.es/vivigui/software.html> and forms part of the R package **Anthropometry** ([Vinué et al. \(2014a\)](#)). Section 5 contains conclusions and some ideas for future work.

## 2. Methodology

We aim to find extremal observations or pure types, which must be specific individuals of the database. To do this, we build upon the archetype analysis algorithm presented in [Cutler and Breiman \(1994\)](#). The archetype algorithm is implemented in the R package **archetypes** ([Eugster and Leisch \(2009\)](#)). We will now summarize the theoretical foundation of archetype analysis and the proposed archetypoid analysis.

### 2.1. Archetype and Archetypoid analysis

Let  $\mathbf{X}$  be an  $n \times m$  matrix that represents a multivariate dataset with  $n$  observations and  $m$  variables. The goal of archetype analysis is to find a  $k \times m$  matrix  $\mathbf{Z}$  that characterizes the archetypal patterns in the data, such that data can be represented as mixtures of those archetypes. Specifically, archetype analysis is aimed at obtaining the two  $n \times k$  coefficient matrices  $\alpha$  and  $\beta$  which minimize the residual sum of squares that arises from combining the equation that shows  $\mathbf{x}_i$  as being approximated by a linear combination of  $\mathbf{z}_j$ 's (archetypes) and the equation that shows  $\mathbf{z}_j$ 's as linear combinations of the data:

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

under the constraints

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

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

On the one hand, constraint 1) tells us that the predictors of  $\mathbf{x}_i$  are finite mixtures of archetypes,  $\hat{\mathbf{x}}_i = \sum_{j=1}^k \alpha_{ij} \mathbf{z}_j$ . Each  $\alpha_{ij}$  is the weight of the archetype  $j$  for the individual  $i$ , that is to say, the  $\alpha$  coefficients represent how much each archetype contributes to the approximation of each individual. On the other hand, constraint 2) implies that archetypes  $\mathbf{z}_j$  are convex combinations of the data points,  $\mathbf{z}_j = \sum_{l=1}^n \beta_{jl} \mathbf{x}_l$ . According to this definition, the archetypes need not be observed individuals of the database. The archetypes would correspond to specific individuals when  $\mathbf{z}_j$  is an observation of the sample, that is to say, when only one  $\beta_{jl}$  is equal to 1 in constraint 2) for each  $j$ . As  $\beta_{jl} \geq 0$  and the sum of constraint 2) is 1, this implies that  $\beta_{jl}$  should only take on the value 0 or 1. In the analysis of archetypoids, the original continuous optimization problem therefore becomes:

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

under the constraints

- 1)  $\sum_{j=1}^k \alpha_{ij} = 1$  with  $\alpha_{ij} \geq 0$  and  $i = 1, \dots, n$  and
- 2)  $\sum_{l=1}^n \beta_{jl} = 1$  with  $\beta_{jl} \in \{0, 1\}$  and  $j = 1, \dots, k$  i.e.  $\beta_{jl} = 1$  for one and only one  $l$  and  $\beta_{jl} = 0$  otherwise.

In summary, archetypes are points  $\mathbf{z}_j$ ,  $j = 1, \dots, k$  that are a mixture of data, such that each data point  $\mathbf{x}_i$  can be expressed as  $\mathbf{x}_i = \sum_{j=1}^k \alpha_{ij} \mathbf{z}_j$ , with the constraint that  $\alpha_{ij}$  are positive and add up to one. Archetypoids add the constraint that the  $\mathbf{z}_j$  must be some point in the dataset. Archetypoids could be considered extreme points in the data.

## 2.2. Computing archetypoids

Various different alternatives were evaluated in order to solve this new mixed-integer optimization problem. We considered branch and bound and

genetic algorithms, but they have a high computational cost when the sample size increases. Besides, the results provided by the genetic algorithm did not satisfy the constraints of the archetypoid analysis problem. Another more naive possibility is to calculate the archetypoids with an exhaustive search, that is to say, to obtain the set of archetypoids that produces the minimum value of the objective function after trying all the possible combinations. This will be called the combinatorial or true solution. However, when the sample size of the database is large, this possibility has a very high computational cost. Because none of these approaches are useful for calculating archetypoids, we decided to develop an algorithm based on the Partitioning Around Medoids (PAM) clustering algorithm, which is very well described in [Kaufman and Rousseeuw \(1990\)](#). An outline of the PAM algorithm can be seen in the supplementary material. In the next section we detail our proposal.

### 2.3. Archetypoid algorithm

The outline of the archetypoid algorithm (the goal is to minimize the  $RSS = \|\mathbf{X} - \alpha\beta'\mathbf{X}\|_2$ , where  $'$  denotes transpose) is (each step is explained below):

1. BUILD phase: look for a good initial set of  $k$  archetypoids from the  $n$  data points.
2. SWAP phase: For each archetypoid  $a$ 
  - (a) For each non-archetypoid data point  $o$ 
    - i. Swap  $a$  and  $o$  and compute the RSS of the configuration ( $\alpha$  coefficients must be calculated).
3. Select the configuration with the lowest RSS.
4. Repeat steps 2 to 4 until there is no change in the archetypoids.

Our algorithm is made up of two phases, a BUILD phase and a SWAP phase, as PAM. In the BUILD step, an initial set of archetypoids is determined. Although we could simply randomly select (without replacement)  $k$  of the  $n$  data points as the initial set, we propose the four, from our point of view, most meaningful ways of trying to find a good initial set which shortens the SWAP phase. The first possibility consists in computing the Euclidean distance between the  $k$  archetypes and the individuals and choosing the nearest ones, as [Epifanio et al. \(2013\)](#) did. From now on, we refer to this candidate set as  $can d_{ns}$ . The second choice identifies the individuals with the maximum  $\alpha$  value for each archetype, i.e. the individuals with the largest relative

share for the respective archetype. We refer to this set as  $cand_\alpha$  and is used in Eugster (2012) and Seiler and Wohlrabe (2013). The third choice identifies the individuals with the maximum  $\beta$  value for each archetype, i.e. the major contributors in the generation of the archetypes. We refer to this set as  $cand_\beta$ . A fourth possibility consists of using the stepwise *FURTHESTSUM* initialization procedure, a clever way to compute possible candidates for archetypes proposed by Mørup and Hansen (2012). We refer to this set as  $cand_{FS}$ . Our initial set of archetypoids is therefore  $cand_{ns}$ ,  $cand_\alpha$ ,  $cand_\beta$  or  $cand_{FS}$ . In our implementation, the archetypes for the first three choices are computed by the R package **archetypes** (the best ones are selected after running the algorithm several times), whereas the fourth choice is computed with the Matlab software available at <http://www.imm.dtu.dk/mm/downloads/PCHA.zip>, which implements the methods described by Mørup and Hansen (2012).

The idea behind the SWAP phase of our algorithm is the same as that of the SWAP phase of PAM, and it is much more computer intensive than the BUILD phase. However, the objective function of our optimization problem is different. This is because PAM is aimed at clustering from  $k$  central points and our algorithm is aimed at finding  $k$  representatives that characterize the extreme types in the data. Specifically, our SWAP phase attempts to improve the set of archetypoids by exchanging selected individuals for unselected individuals and by checking whether these replacements reduce the objective function of the equation 1. In the inner loop, for each given set of archetypoids,  $S$ , the  $\alpha$  coefficients are updated in order to calculate the effect of the swap. The corresponding RSS is then calculated. If this RSS is lower than the previous RSS,  $S$  is the new initial vector of archetypoids. This second phase is repeated until there is no change in any archetypoid. Note that as all potential swaps are considered, the results of the algorithm do not depend on the order of the objects in the database. As mentioned, the  $\alpha$  coefficients need to be computed. In our implementation, we solve  $n$  convex least squares problems as in the algorithm implemented in **archetypes**. In order to solve those  $n$  convex least squares problems, a penalized version of the non-negativity least squares algorithm by Lawson and Hanson (1974) is used, in such a way that the convexity constraints (being non-negative and adding up to one) are fulfilled (see specifically (Eugster and Leisch, 2009, point 2.1 on page 3)). Furthermore, we do not update the  $\beta$  coefficients in the same way as **archetypes** (in the inner loop by solving  $k$  convex least squares problems). In our algorithm, the  $\beta$  coefficients are “updated” in the sense that for the individuals considered as archetypoids, their  $\beta$  is equal to

1, and the  $\beta$  for the other unselected individuals is equal to 0. The RSS is calculated with the 2-norm or spectral norm, which is computed as the largest singular value of the matrix, as in [Eugster and Leisch \(2009\)](#), although other matrix norms, such as the Frobenius norm, could be used.

The archetype algorithm alternates between finding the best  $\alpha$  for given archetypes  $\mathbf{Z}$  and finding the best archetypes  $\mathbf{Z}$  for given  $\alpha$ . However, with our algorithm, we only focus on finding the best  $\alpha$  for given archetypoids  $\mathbf{Z}$ . This is because the difference between archetypes and archetypoids is that archetypes are not necessarily observed points, but archetypoids are.

As we said before, our algorithm is based on the PAM clustering method. This type of algorithm aims to find good solutions in a short period of time, although not necessarily the best solution. Otherwise, the global minimum solution could always be obtained using as much time as necessary with the combinatorial solution, but this would be computationally very inefficient.

As regards standardization of the data, it should be mentioned that standardizing the data depends on the nature of the data. The variables should be standardized in cases where they measure different dimensions. Standardization is suitable if the scales are not comparable and especially if the ranges of variables are very different. This is the case of the aircraft pilot database. In other circumstances it makes sense to work with the data as they stand. This is the case of the database used in the apparel design example, where we work with the configuration that aims to reproduce the original dissimilarities between trunks. Therefore the variables have an absolute meaning.

Our procedure was therefore as follows: First, depending on the problem, it must be decided whether or not the data should be standardized, then the archetypes must be calculated and, finally, the archetypoids must be calculated with our algorithm, beginning from the initial sets, for several values of  $k$ . As in [Cutler and Breiman \(1994\)](#) and [Eugster and Leisch \(2009\)](#), we select the  $k$  where the elbow on the RSS representation is found.

#### *2.4. Archetypoids when features are unavailable*

In some problems, especially those where multidimensional scaling (MDS) applies, such as psychology, economy, etc., only dissimilarities are available, for example in studies of perception. In those cases, we cannot approximate the data directly as mixtures of archetypoids or archetypes. However, if the dissimilarities are Euclidean distances, they can be represented exactly in at most  $n - 1$  dimensions ([Mardia, Kent, and Bibby \(1979, Theorem 14.4.1\)](#)) by means of classical multidimensional scaling (cMDS). cMDS takes a set of

dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities, since the dimension of the space which the data are to be represented in is usually less than  $n - 1$ . We can use these features to find the archetypoids. Note that the archetypes can also be computed in this new space, but we cannot establish a correspondence with the original subjects or create artificial subjects, for which only the dissimilarities were available. If the dissimilarities are a distance but not an Euclidean distance, cMDS can be used as an approximation (and it is optimal for a kind of discrepancy measure (Mardia et al., 1979, Theorem 14.4.2), or we can use the h-plot (Epifanio (2013)), a recent alternative method that also works when the dissimilarity is not a distance, or any other MDS method.

Let us detail the phases for obtaining the archetypoids when features are unavailable. Let  $\mathbf{D}$  be the  $n \times n$  matrix where  $d_{ij}$  denotes the dissimilarity between the observations  $i$  and  $j$ .

1. Compute an MDS method for finding a representation in  $\mathbb{R}^m$  that preserves the pairwise dissimilarities (the information of  $\mathbf{D}$ ) in some way. Depending on the method, a goodness of fit measure can be used to choose  $m$ . See the supplementary material for more details. Note that the greater  $m$  is, the more variables are available, and the computation time increases with the number of variables (Eugster and Leisch (2009)).

2. Compute the archetypoids of the  $n \times m$  matrix  $\mathbf{X}$ , the matrix returned by the MDS method. This matrix has the coordinates of the points computed to represent the dissimilarities. These archetypoids correspond to a specific set of observations, with a direct correspondence with the original observations. Note that we also obtain the  $\alpha$  coefficients, indicating the contribution of each original archetypoid to each original observation. However, the predictors ( $\sum_{j=1}^k \alpha_{ij} \mathbf{z}_j$ ) of each original observation cannot be represented with the original information (the dissimilarities), in the same way that archetypes cannot be represented in that space.

We use a well-known database in MDS to make our ideas clearer. Wish (1971) asked 18 students to rate the similarity between 12 nations. Let us call  $\mathbf{S}$  the matrix with the mean similarity ratings. The standard transformation from  $\mathbf{S}$  to a distance matrix  $\mathbf{D}$  is defined by  $d_{ij} = (s_{ii} - 2s_{ij} + s_{jj})^{1/2}$  (Mardia et al., 1979, Definition 14.2.14). Note that  $s_{ij} = s_{ji}$  and  $s_{ij} \leq s_{ii}$  for all  $i$  and  $j$ , therefore the quantity under the square root is non-negative and  $d_{ii} = 0$ . As  $\mathbf{S}$  is positive definite,  $\mathbf{D}$  is Euclidean (Mardia et al., 1979, Theorem 14.2.2). A typical application of cMDS is to consider a two-dimensional MDS configuration of the distances in order to interpret the data. Ex-

tremes are better than central points for human interpretation (Thureau et al. (2012)). In Fig. 1 we display the 4 archetypes from the cMDS representation. Archetypoids beginning from  $cand_\alpha$ ,  $cand_\beta$  and  $cand_{FS}$  (these three sets are identical: Brazil, China, Congo, and USA) and  $cand_{ns}$  (Brazil, Congo, USA and Yugoslavia) are: China, Congo, USSR, and USA, which coincides with the combinatorial (true) solution. Note that one archetype is between China and Yugoslavia, which makes it difficult to interpret. Actual data points are more easily interpretable by individuals. None of the initial sets ( $cand_{ns}$ ,  $cand_\alpha$ ,  $cand_\beta$  or  $cand_{FS}$ ) coincide with the true solution. Assuming that the respondents use our knowledge of those countries in the ‘70’-s, we can interpret the archetypoids as: China was communist and more economically underdeveloped than the USSR (communist and economically developed), whereas the USA was economically developed and noncommunist, and Congo was underdeveloped. Congo and Brazil appear in the initial sets, and also archetypes, when both countries had the same profile: neither of them were highly industrialized or extreme political alignment countries. With archetypoids the contrast between countries in the ‘Political Alignment and Ideology’ and ‘Economic Development’ dimensions is clearer.

We have also computed the KERNEL-AA proposed by Mørup and Hansen (2012), which generalizes the archetype analysis to kernel representations, with  $\mathbf{S}$ . KERNEL-AA returns archetypes, not actual data points, so the interpretation is not as clear as archetypoids. The first archetype is a convex combination of Cuba (10%), China (32%), USSR (28%) and Yugoslavia (30%), which could correspond to a profile of communist countries. The second archetype is a convex combination of France (12%), Israel (25%), Japan (26%) and USA (37%), which could correspond to a profile of economically developed and non-communist countries. The third archetype is constituted by the weighted sum of Congo (30%), Egypt (38%) and India (32%), which could correspond to countries that are not economically developed. The fourth archetype is formed by the combination of Brazil (80.3%), Congo (1.3%) and Cuba (18.4%), which could also correspond to countries that are not economically developed. As  $\mathbf{D}$  is Euclidean, the configuration of cMDS in 11 dimensions exactly reproduces the interpoint distances. The archetypoids in this representation beginning from  $cand_{ns}$  and  $cand_\alpha$  are as follows (coinciding with the true solution): Congo, Egypt, USSR and USA. They are the same archetypoids obtained with cMDS in 2D, except China is replaced by Egypt. Egypt was an Arab country classed as not economically developed. Note that the paper by Wish (1971) was published before

Egypt's Cold War allegiance switched from the USSR to the USA in 1972. Again, it seems that with archetypoids the contrast between countries in the 'Political Alignment and Ideology' and 'Economic Development' dimensions is clearer.

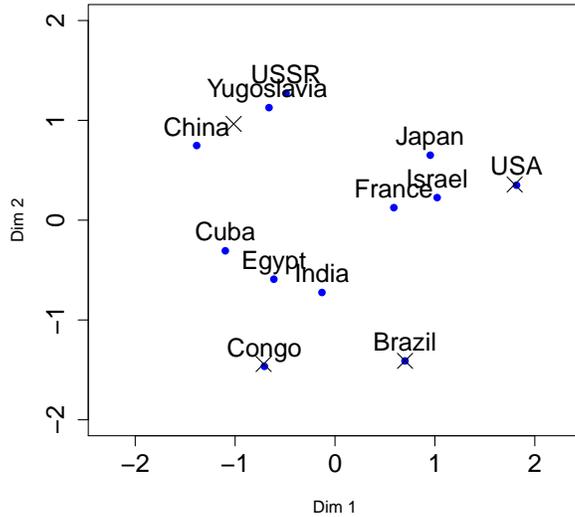


Figure 1: 2D cMDS representation of similarity ratings of nations. Archetypes are represented by crosses.

### 3. Location of the archetypoids

Let  $Conv(\mathbf{X})$  be the convex hull of the  $n$  observations in  $\mathbb{R}^m$  of the set  $\mathbf{X}$ . As the number of points in  $\mathbf{X}$  is finite,  $Conv(\mathbf{X})$  is a convex polytope, which is the convex hull of its vertices. A vertex of  $Conv(\mathbf{X})$  is an observation  $\mathbf{x}_i$  of  $\mathbf{X}$  for which  $\mathbf{x}_i$  does not belong to  $Conv(\mathbf{X} \setminus \{\mathbf{x}_i\})$ . A vertex of  $Conv(\mathbf{X})$  is also called an extremal point of  $\mathbf{X}$ . Let  $\mathbf{V}$  be the set of vertices of  $Conv(\mathbf{X})$ , and  $N$  be the number of vertices.

Let us see the location of the archetypoids and the differences with the archetype locations for different values of  $k$ .

1. If  $k = 1$ , the archetypoid is the medoid (with one cluster) of  $\mathbf{X}$  considering the squared Euclidean distance as dissimilarity, since the minimization of  $RSS$  coincides with the definition of the medoid (Kaufman and Rousseeuw

(1990)) (the medoid is that object of the cluster for which the average dissimilarity to all the objects of the cluster is minimal). In case of archetypes the  $k = 1$  solution is the sample mean (Cutler and Breiman (1994)).

2. If  $k = N$  (or  $> N$ ), the archetypoids are  $\mathbf{V}$  (or  $\mathbf{V}$  plus any other observation), as  $RSS = 0$ , since  $Conv(\mathbf{V}) = Conv(\mathbf{X})$ .

3. If  $1 < k < N$ , we cannot state that the archetypoids are on the boundary of  $Conv(\mathbf{X})$  as the archetypes are, as we can see in the following artificial example 1. It depends on the distribution of the observations, although for Normal distributions archetypoids seem to be vertices, as can be seen in the example 2, where we reproduce the example of (Hastie et al., 2009, Fig. 14.35) and (Cutler and Breiman, 1994, Fig. 14).

**Example 1** Fig. 2a shows the location of 7 points in  $\mathbb{R}^2$ . Archetypes and archetypoids are computed for  $k = 2$  (note that with  $k = 4$  the  $Conv(\mathbf{X})$  is the square formed by vertices 1, 2, 3 and 4, and these are the archetypes and archetypoids). Note that the archetypoids are not vertices. The nearest points to archetypes coincide with the archetypoids in this example. The archetypes are built as a weighted mean of the pair 2 and 3, and the pair 1 and 4. If we compute the RSS for archetypes and archetypoids, the elbow is at  $k = 4$ . In Fig. 2b another artificial example is displayed with  $k = 2$ . In this case, the nearest points to the archetypes are 1 and 4, when the archetypoids are 7 and 8 (not vertices).

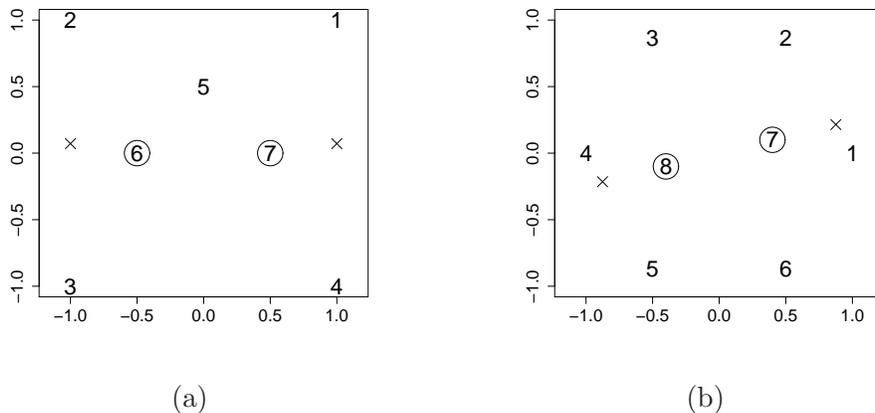


Figure 2: Two examples where the two archetypoids obtained by the combinatorial method (circles) are not on the boundary of  $Conv(\mathbf{X})$  as the two archetypes (crosses) are.

**Example 2** A sample of size 50 has been generated from  $N(\mu, \Sigma)$ , where  $\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  and  $\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ . Figs. 3a, 3b and 3c show the archetypes and archetypoids (computed with our algorithm beginning from the  $cand_{ns}$  set) for  $k = 2, 4$  and  $8$ , respectively (in this example  $N$  is 7). Note that like archetypes, the archetypoids do not nest (as more archetypoids are found, the existing ones can change to better capture the shape of the dataset). In the supplementary material, we present an expanded analysis of this example, which has been repeated 100 times.

The stability (if the solution does not change much when the data are modified slightly) of archetypoids is also studied in the supplementary material. They are very stable, especially when compared with the medoids of PAM.

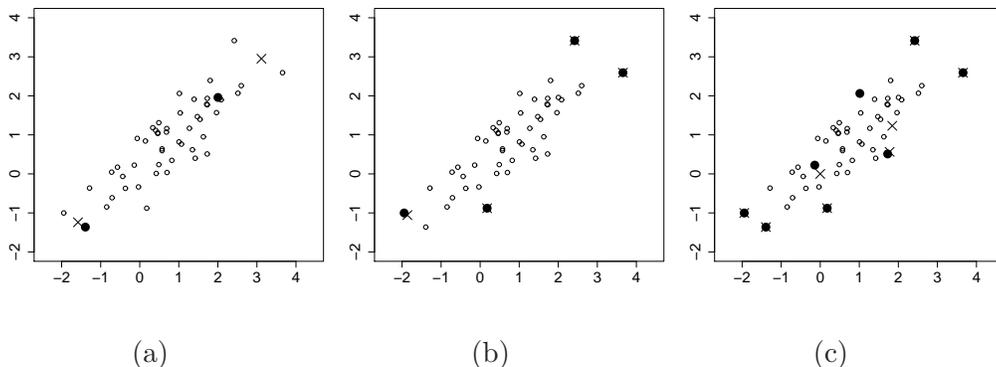


Figure 3: Archetypes (with crosses) and archetypoids (with solid circles) for simulated Bivariate Normal Data, with  $k = 2$  (a),  $4$  (b) and  $8$  (c) respectively. The archetypoids beginning from  $cand_{ns}$  returned with our algorithm coincide with the combinatorial solutions. Although RSS for archetypes should theoretically be smaller than for archetypoids because the range of possible solutions is larger, in (c) the RSS for archetypoids is  $3.3e-15$  (zero), but  $9.764514e-4$  with archetypes. In fact, the archetype algorithm does not recover  $\mathbf{V}$ , it converges to a local minimum, even considering 500 random starts.

### 3.1. Comparison with other unsupervised methods

In the matrix notation used in Mørup and Hansen (2012), the objective of archetypoid analysis is to find the optimal matrices  $\beta$  and  $\alpha'$  minimizing some measure of distortion  $D(\mathbf{X}'|\mathbf{X}'\beta\alpha')$  (for example,  $\|\mathbf{X}' - \mathbf{X}'\beta\alpha'\|_2$  or  $\|\mathbf{X}' - \mathbf{X}'\beta\alpha'\|_F^2$ ). As an extension of Mørup and Hansen (2012), Table 1

shows the relationship between archetypoid analysis and different unsupervised methods (seen as a linear mixture type representation of data with various constraints) in terms of possible values of  $\beta$  and  $\alpha$  (note that  $\mathbf{X}'\beta$  are the feature vectors, while  $\alpha'$  gives the weights for the predictors of  $\mathbf{X}$ ).

Table 1: Relationship between archetypoid analysis and several unsupervised methods, as in [Mørup and Hansen \(2012\)](#): Principal component analysis (PCA), Non-negative matrix factorization (NMF), Convex NMF (CNMF), Archetype analysis (AA), Archetypoid analysis (ADA), Soft  $k$ -means (i.e. fuzzy  $k$ -means or the EM-algorithm for clustering),  $k$ -means and  $k$ -medoids.  $\mathbb{B}$  represents the set  $\{0, 1\}$ .

PCA	$\beta \in \mathbb{R}$
	$\alpha \in \mathbb{R}$
NMF	$\mathbf{X}'\beta \geq 0$
	$\alpha \geq 0$
CNMF	$\beta \geq 0$
	$\alpha \geq 0$
AA	$ \beta_k _1 = 1, \beta \geq 0$
	$ \alpha_n _1 = 1, \alpha \geq 0$
ADA	$ \beta_k _1 = 1, \beta \in \mathbb{B}$
	$ \alpha_n _1 = 1, \alpha \geq 0$
Soft $k$ -means	$\beta_{k,n} = \frac{\alpha_{k,n}}{\sum_{\tilde{n}} \alpha_{k,\tilde{n}}}$
	$ \alpha_n _1 = 1, \alpha \geq 0$
$k$ -means	$ \beta_k _1 = 1, \beta \geq 0$
	$ \alpha_n _1 = 1, \alpha \in \mathbb{B}$
$k$ -medoids	$ \beta_k _1 = 1, \beta \in \mathbb{B}$
	$ \alpha_n _1 = 1, \alpha \in \mathbb{B}$

Other authors have previously compared archetypal representation with other unsupervised methods. For example, ([Hastie et al., 2009](#), Sec. 14.6.1) compared archetypal analysis with  $k$ -means clustering and NMF. They also applied AA, PCA and ICA (independent component analysis) to the same database. [Mørup and Hansen \(2012\)](#) analyzed several databases with AA, PCA, NMF, ICA and  $k$ -means, and [Canhasi and Kononenko \(2013\)](#) also

compared AA with PCA, NMF and  $k$ -means and other multi-document summarization methodologies. There is a clear difference between archetypal analysis and clustering. The former focuses on extremes in the data, while traditional clustering algorithms, like  $k$ -means or PAM, segment subjects based on centroids (averages) or medoids (the points obtained with PAM). Furthermore, the main objective of archetypal analysis is to obtain the archetypes or archetypoids, but the main objective of clustering focuses on membership in each cluster.

Thurau et al. (2012) introduced the Simplex Volume Maximization (SiVM) algorithm. They formulated the same problem as ADA, but they seek to minimize the Frobenius norm, and our algorithm can consider any matrix norm, although in our implementation the 2-norm is considered. However, they assumed that archetypoids are vertices, when we have shown in Example 1 that it is not necessary true. Therefore SiVM cannot return the true solutions in that example. SiVM aims to select sequentially the  $j + 1$  vertex that maximizes the simplex (polytope which is the convex hull of its vertices) volume given the first  $j$  vertices. For Fig. 1, SiVM returns the same countries as the  $cand_\alpha$ ,  $cand_\beta$  and  $cand_{FS}$  sets, which was not the best solution. Due to its efficiency (low running times), SiVM gives a reasonable approximation in the case of very large databases.

In order to better understand the differences between the different methodologies for obtaining representative data (most of them are clustering methods) and archetypoid analysis, the same data as in Fig. 3b will be used. Fig. 4 shows the representatives using different methods. Specifically, we have used: a) SiVM; b) the Sparse Modeling Representative Selection method developed by Elhamifar et al. (2012)(SMRS); c) the Affinity Propagation algorithm (AP) by Frey and Dueck (2007); d) the HOTTOPIXX (a new approach for non-negative matrix factorization (NMF)) (Bittorf et al. (2012)), using the code developed by Gillis (2013); e) a Bayesian partial membership model (BPM) (Mohamed et al. (2014)), in which we have represented the points with the highest membership in each group; and f) PAM,  $k$ -means and fuzzy  $k$ -means.

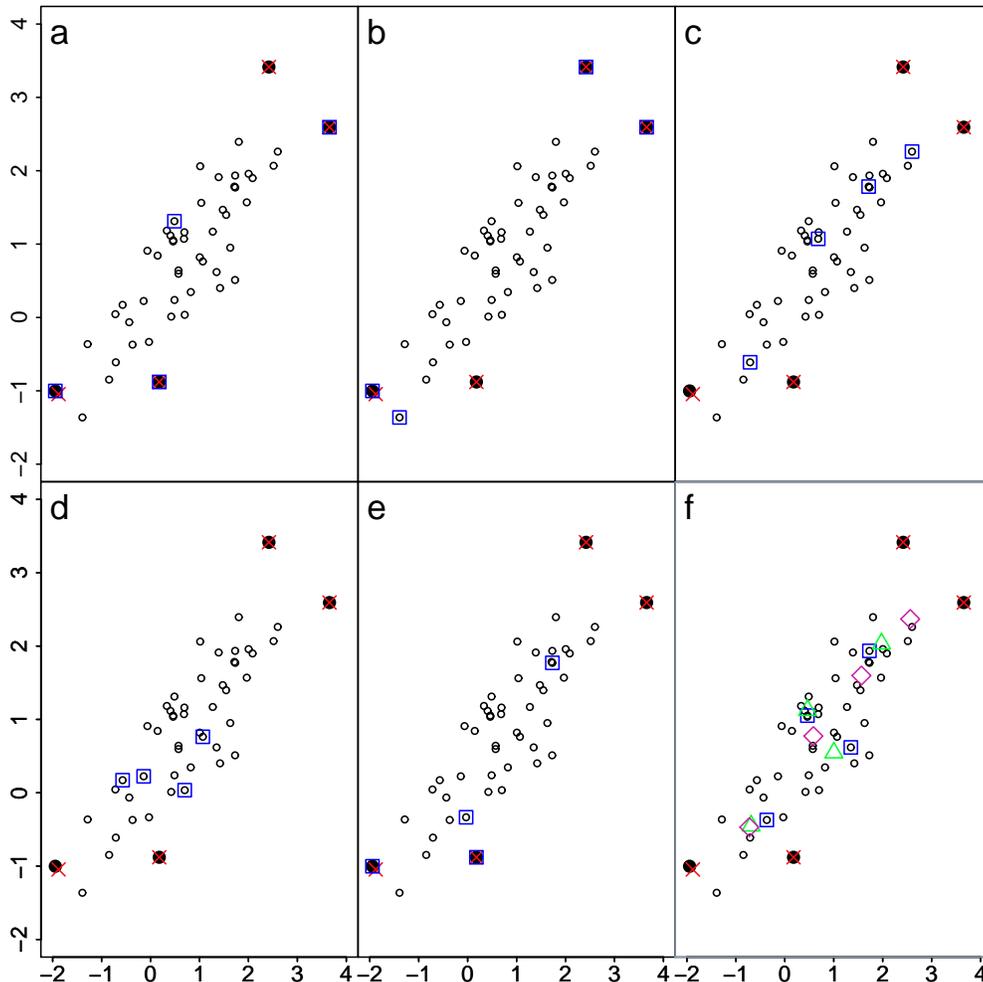


Figure 4: Archetypes (with red crosses) and archetypoids (with solid black circles) for simulated Bivariate Normal Data, with  $k=4$ , together with the four representatives (with blue squares) obtained for the following methods, respectively: (a) SiVM, (b) SMRS, (c) AP, (d) HOTTOPIX, (e) BPM and (f) classical clustering algorithms (PAM with blue squares,  $k$ -means with green triangles and fuzzy  $k$ -means with magenta diamonds).

Except for SiVM and SMRS, the other methods return representatives that are mainly in the middle of the data rather than the boundaries, as we are seeking. On the one hand, if we consider the convex hull generated with the points obtained with SiVM (the rhombus built joining those points), many data points fall outside this rhombus (greedy algorithms are

fast and often give good solutions, but a certain selection in a determined iteration could prevent a good solution being found because they do not reconsider their selections). This fact indicates that RSS (with 2-norm) for SiVM ( $2.990284e-2$ ) is larger than using ADA ( $8.064882e-3$ ), which was the best solution. Note that one of the points obtained with SiVM is not a vertex of  $Conv(\mathbf{X})$ , maybe due to their implementation and their parameter selection for computational efficiency (SiVM with  $k = 8$  does not recover the  $N = 7$  vertices of  $\mathbf{V}$ ). On the other hand, unlike ADA (where the points are approximated by a mixture of archetypoids and the coefficients therefore add up to one and are positive), with SMRS each point in the dataset is approximated by an affine combination of the representatives, meaning that the coefficients can be negative (in fact, for this example several coefficients are negative; the maximum value of the coefficients is 0.669, only 6 coefficients are above 0.5 and the majority of non-zero values are between 0.2 and 0.3). This makes their intuitive interpretation difficult. The RSS for SMRS is  $1.704842e-2$ . Furthermore, with SMRS it is not possible to select exactly how many representatives have to be obtained. In this example, only two representatives were returned by the algorithm, since the other two representatives were below a certain threshold, and only four representatives were extracted in total (without considering the threshold). We could not have obtained five or more representatives in this dataset with SMRS. We have also considered the fast and robust recursive algorithm for separable NMF by [Gillis and Vavasis \(2014\)](#), but it only returned two points (the two uppermost points). Note that their stopping criterion does not fix the number of points to extract a priori.

## 4. Applications and results

### 4.1. Cockpit design problem

The dataset of this problem comes from the 1967 United States Air Force (USAF) Survey. From the total variables, we select six anthropometric measurements for 2420 Air Force personnel, which are the same as those selected by [Epifanio et al. \(2013\)](#). These six dimensions are the so-called cockpit dimensions because they are the most important dimensions when designing aircraft cockpits. A description of each of them can be found in the supplementary material. As in [Epifanio et al. \(2013\)](#), the variables are standardized and subjects outside the 95% density contour are discarded, as in the expanded analysis of example 2 in the supplementary material. The

screepLOTS in Fig. 5 suggest that 3 archetypes and archetypoids should be chosen. In the interests of brevity and as an illustrative example we examine the results of 3 archetypes and archetypoids. However, in a real situation it would be up to the analyst to decide how many representative cases to choose. Table 2 shows the RSS associated with this number of archetypes, with the initial sets and with the same number of archetypoids. The smallest RSS in Table 2 is for the archetypes. This could be expected because its set of possible solutions is the largest. However, the RSS of the  $cand_{ns}$ ,  $cand_{\alpha}$ ,  $cand_{\beta}$  and  $cand_{FS}$  archetypoids (archetypoids beginning with  $cand_{ns}$ ,  $cand_{\alpha}$ ,  $cand_{\beta}$  and  $cand_{FS}$ , respectively) are quite close to that (in particular with  $cand_{\alpha}$  archetypoids). In addition, the RSS of the archetypoids decrease the corresponding RSS of the initial sets. Although not dramatic, this reduction is notable. Furthermore, it may be the case that the nearest individuals are not plausible individuals (as in fact occurs in Seiler and Wohlrabe (2013), where the nearest “economists are a mixture of different types”). In that case, it would be necessary to look for archetypoids.

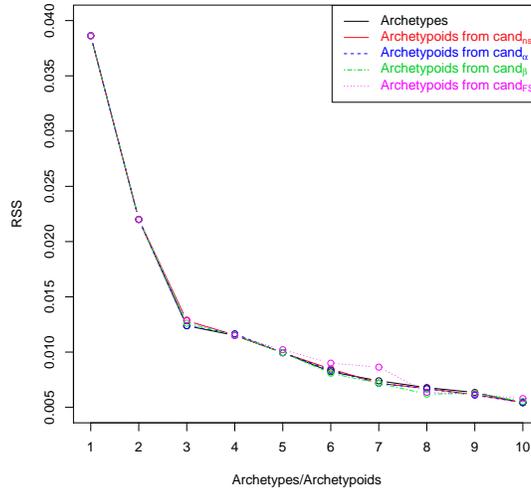


Figure 5: Screenshot of the RSS of the archetypes and archetypoids for the aircraft pilots. The elbow is at 3 in all the cases.

Fig. 6 shows the percentiles of the archetypoids beginning with  $cand_{\alpha}$ . The percentiles of each archetypoid are represented by each set of bars, where a bar represents a different variable, from dark gray (thumb tip reach) to light

Table 2: RSS of archetypes, initial sets and archetypoids for the aircraft pilots.

	RSS
3 archetypes	0.01238078
$cand_{ns}$ (511,314,1691)	0.01824692
$cand_{\alpha}$ (1421,314,1691)	0.01947072
$cand_{\beta}$ (2027,611,1114)	0.01418717
$cand_{FS}$ (187,1114,1560)	0.01417640
3 archetypoids from $cand_{ns}$ (2177,2240,1691)	0.01283086
3 archetypoids from $cand_{\alpha}$ (1632,1822,52)	0.01238504
3 archetypoids from $cand_{\beta}$ (107,757,110)	0.01261483
3 archetypoids from $cand_{FS}$ (1946,1319,1114)	0.01290036

gray (shoulder height sitting), as in [Epifanio et al. \(2013\)](#). The first  $cand_{\alpha}$  archetypoid is small in all measurements. The second  $cand_{\alpha}$  archetypoid is high in the six variables. The third  $cand_{\alpha}$  archetypoid has high percentiles for the first three variables (corresponding to limb dimensions), and small percentiles for the last three variables (corresponding to torso dimensions). The percentiles of the three  $cand_{\alpha}$  archetypoids are not very extreme. Finally, it should be mentioned that given the large sample size of this database, it was not possible to obtain the combinatorial solution in a reasonable time.

As explained in Sect. 1, the archetypoids are human live models, which are used in the design and fit evaluation (see ([HFES 300 Committee, 2004](#), Ch. 6) for details about transitioning cases to products, or the ISO 15537:2004 standard ([International Organization for Standardization \(2004\)](#)) for determining representative subjects of the target population applicable to the testing of industrial products and designs). This procedure and its benefits (producing an effective design, while simultaneously minimizing cost and maximizing accommodation) compared to the virtual evaluation of products using theoretical digital models was described in [Robinette and Hudson \(2006\)](#). Note that only live models can accurately represent postures, tissue deformation, painful pressures or forces, fatigue, and strength in marginal reach zones. An additional possibility for using archetypoids is explained in [Veitch et al. \(2013\)](#). It consists of obtaining boundary cases (archetypoids) for designing a single cockpit. This reasoning is based on defining the anthropometric dimensions of all pilots who can fly each aircraft model. There are aircraft models with archetypoids that are not recommended because serious

risks (safety criteria include escape clearance, minimal operation clearances including the ability to reach emergency controls or external visual field) have been detected in ergonomic testing with real subjects who represent each archetypoid.

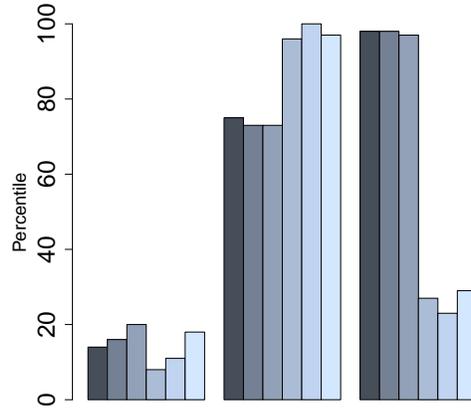


Figure 6: Percentiles for the six variables of the 3 archetypoids beginning with  $cand_\alpha$  for the aircraft pilots. Each one is represented by a different shade of gray, from the lightest to the darkest shade in the same order as the variables are shown in Table 3 of the supplementary material.

#### 4.2. Apparel design problem

A national 3D anthropometric survey of the female population was conducted in Spain in 2006 by the Spanish Ministry of Health. The aim of this survey was to generate anthropometric data about the female population for the clothing industry (Alemany et al. (2010)). In this study, a sample of 10415 Spanish females from 12 to 70 years old were randomly selected. Images are captured by several cameras and a triangulation is generated using associated software supplied by the scanner manufacturers, providing knowledge about the 3D spatial location of a large number of points on the surface of the body. A 3D binary image of the trunk of each woman (white pixel if it belongs to the body, otherwise black) is produced from the collection of points located on the surface of each woman scanned, as explained in Ibáñez et al. (2012). The location is removed by translating each image to the origin in such a way that its centroid coincides with the origin. Each trunk is also

rotated to make its principal inertia axis coincide with the canonical axis of coordinates.

We can compute the dissimilarity between trunk forms and build a dissimilarity matrix  $\mathbf{D}$  between women. Let  $A$  and  $B$  be two binary images associated with the trunk of two women and defined in a lattice  $\Lambda$ . There are several metrics for measuring the differences between  $A$  and  $B$ . We use the simplest one, which is the misclassification error:  $d(A, B) = \frac{nu(A\Delta B)}{nu(\Lambda)}$ , where  $\Delta$  is the set symmetric difference and  $nu$  counts the number of pixels in that set, that is to say, the volume of the set.

We selected 470 women: women between 25 and 45 years old with a bust circumference between 86 and 90 cm, not including pregnant and lactating women. This age range represents an important potential group for the apparel market and, at the same time, includes a high variability of body shapes (Alemay et al. (2010)). As a result, women with the same size (86-90 cm bust for upper garments) may have very different body shapes (De Raeve et al. (2012)), causing fitting problems when a garment is designed to fit a body prototype perfectly. Thus, different classifications of body types have been proposed for apparel sizing and design (Simmons et al. (2004), Rasband and Liechty (2006), Faust and Carrier (2009), Hsu (2009)).

Within this context, it is proposed that archetypoids should be used to identify subjects who represent the fittings problems of the target population. Having identified the extremes of a size, and together with the central case that represents the basic proportions in a range of clothing, the apparel grading process within that size could begin. The designer may increase or decrease the base pattern to ensure that each new pattern is adapted to the measurements of the extremes. Several boundary cases can be used in conjunction with the central case to make the adjustments needed to accommodate both the boundaries and all the individuals between the boundaries. Note that we are not seeking to find sub-sizes, but to accommodate women within a specific size. Distributed and central cases and clustering algorithms should be used to define sizes (Ibáñez et al. (2012), Vinué et al. (2014b)).

The methodology explained in Section 2.4 has been used with  $\mathbf{D}$ , which describes the dissimilarity between the 470 women. We have made a simulation study (see the supplementary material), and based on this we have chosen to use the h-plot with  $m = 4$ . The screeplots in Fig. 7 suggest that 3 or 6 archetypes and archetypoids should be chosen. In the interests of brevity and as an illustrative example we examine the results of 3 archetypes and archetypoids. However, in a real situation it would be up to the analyst to

decide how many representative cases to choose.

Table 3 shows the RSS associated with this number of archetypes, with the initial sets and with the archetypoids. The smallest RSS is for the archetypoids (all archetypoids agree). Although the RSS for archetypes should be theoretically smaller than for archetypoids because the range of possible solutions is larger, the archetype algorithm converges to a local optimum despite the 20 random starts. The RSS associated with the archetypoids is smaller than the RSS corresponding to the respective initial sets. Because the sample size is not too large (470 women), we were able to calculate the combinatorial solution. It consists of the same individuals obtained with our algorithm. The best set of three archetypoids was obtained after 20 days of computation, using a forward sequential search procedure run on a single computer. Our algorithm only needed a few minutes starting with the initial sets to obtain the same vector.

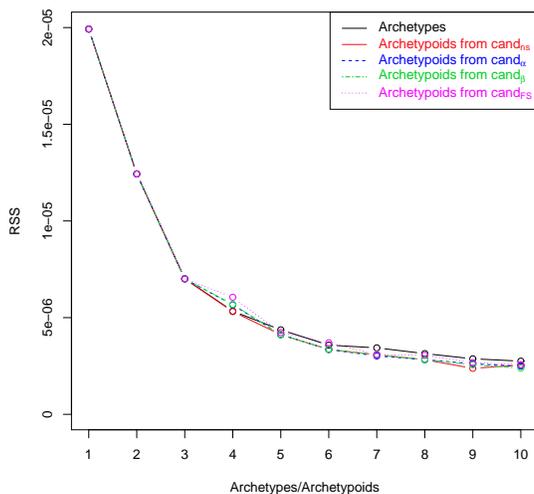


Figure 7: Screeplots of the RSS of the archetypes and archetypoids for Spanish women.

In addition, Table 4 describes the archetypoid women according to certain easily recognized variables: weight, height, waist circumference and hip circumference. Finally, Fig. 8 shows the archetypoids. It should be remembered that the bust circumference was between 86 and 90 cm in our sample. We can see that MADY179 has similar bust, waist and hip circumferences.

Her trunk is cylindrical and she is short. On the other hand, for PEGO137 her bust and waist measurements are similar, but her hip circumference is bigger. She is bell-shaped and overweight. Finally, MADY123 has a small waist, with an hourglass shape. She is tall and thin.

Table 3: RSS of archetypes, initial sets and archetypoids for the Spanish women.

	RSS
3 archetypes	7.003166e-06
$cand_{ns}$ (287,397,459)	7.085687e-06
$cand_{\alpha}$ (267,287,414)	8.799861e-06
$cand_{\beta}$ (171,217,287)	7.088453e-06
$cand_{FS}$ (170,383,414)	9.839477e-06
3 archetypoids from $cand_{ns}$ (287,394,397)	7.002191e-06
3 archetypoids from $cand_{\alpha}$ (287,394,397)	7.002191e-06
3 archetypoids from $cand_{\beta}$ (287,394,397)	7.002191e-06
3 archetypoids from $cand_{FS}$ (287,394,397)	7.002191e-06

Table 4: Anthropometric measurements of women archetypoids.

Woman code	Weight	Height	Waist circumf.	Hip circumf.
MADY123	59.0	1684	698	995
MADY179	48.8	1537	796	897
PEGO137	69.0	1620	865	1130

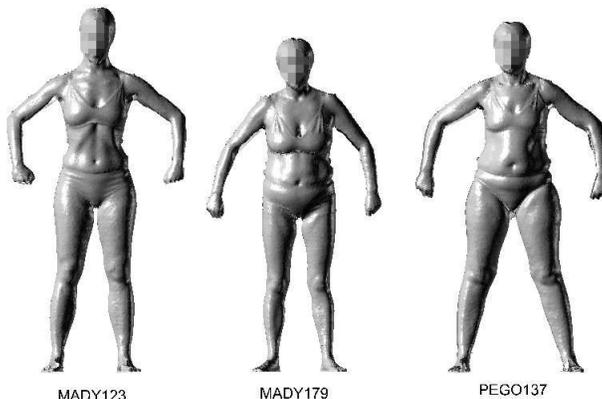


Figure 8: Three archetypal women: MADY123, MADY179 and PEGO137.

## 5. Conclusions

Archetypal analysis is widely used today in problems where the goal is to define extreme representative data. An important drawback of archetypal analysis is that the archetypes do not necessarily correspond to observed individuals. However, in some cases it is critical that the archetypes are real subjects. Within this context, a new archetypal concept has been proposed: the archetypoid. In addition, an algorithm has been developed to obtain them quickly and efficiently in terms of computational complexity (see the supplementary material for a comparison of our algorithm with other algorithms).

In the cockpit and apparel design problems the RSS of archetypoids is really decreased to the same level as the archetype-RSS (in fact, for the apparel problem the RSS of archetypoids was smaller than the RSS of archetypes). It is not possible to say which initial set is the best to start with. The true solution in Fig. 2a is only returned by beginning with the  $can_{ns}$  option. The true solution in Fig. 3a is only returned by beginning with the  $can_{ns}$  and  $can_{\beta}$  sets. The archetypoids obtained by beginning with all the options are the best solution for Figs. 1, 3b and 3c and the apparel problem, while the  $can_{\alpha}$  alternative offered the local minimum for the cockpit design problem. Neither of the options returns the true solution for Fig. 2b. All options must be checked, although  $can_{FS}$  usually begins with a higher RSS than the other options. Obviously, this issue need not be taken into account if the initial sets coincide.

There are various possibilities for future work: from a practical perspective, a study about the computational complexity of archetypoids could be carried out based on the ideas of [Eugster and Leisch \(2009, Sect. 4\)](#), in addition to the implementation of the archetypoid algorithm for very large databases. For very large databases, this algorithm is not practical. In that case, an algorithm using samples of the data following the idea of the Clustering LARge Applications (CLARA) ([Kaufman and Rousseeuw \(1990\)](#)) would be more suitable. From a theoretical point of view, it would be interesting to carry out a numerical simulation with data from different probability distributions to study the location of the archetypes and archetypoids and to study their accuracies by means of randomization techniques. Another direct extension would be to try to define weighted and robust archetypoids, similarly to [Eugster and Leisch \(2011\)](#), or to consider missing values by modifying the objective function analogously, as [Mørup and Hansen \(2012\)](#) did with AA.

Furthermore, archetypoid analysis can be used beyond multivariate vectors or dissimilarity matrices. For example, it is suitable for use with functional data, interval data, images ([Thureau and Bauckhage, 2009](#)), etc. The calculus of archetypoids can be successfully applied in many fields such as computer vision, neuroimaging, chemistry, text mining, collaborative filtering, etc. ([Mørup and Hansen \(2012\)](#)). In fact, archetypal analysis is currently quite popular and several applications have emerged in recent years.

We intend to look in more depth at the apparel design problem by considering more body sizes and age populations. We have determined three archetypoids in our approach. However, it may be more interesting to consider a greater number of representative individuals in order to achieve a better fit of garments. For example, as we said before, we could also have considered six archetypoids. Lack of fit is one of the main complaints about clothing for both customers and apparel companies. Several sociological studies have shown that a high percentage of customers have problems finding the right size and style. In the target group we analyzed, nearly 30% of women have difficulty finding clothes that fit them. In the whole Spanish anthropometric study this percentage is as much as 40%. We expect that archetypoid analysis will serve as a satisfactory approach to tackle this problem.

## 6. Acknowledgements

The authors would like to thank Juan Domingo from the University of Valencia for providing the binary images of women’s trunks. They would also like to thank the Biomechanics Institute of Valencia for providing them with the dataset and the Spanish Ministry of Health and Consumer Affairs for having promoted and coordinated the “Anthropometric Study of the Female Population in Spain”. The authors are also grateful to the Associate Editor and two reviewers for their very constructive suggestions, which have led to improvements in the manuscript. This work has been partially supported by Grant DPI2013-47279-C2-1-R.

## References

- Aleman, S., González, J. C., Nácher, B., Soriano, C., Arnáiz, C., Heras, H., 2010. Anthropometric survey of the Spanish female population aimed at the apparel industry. In: Proceedings of the 2010 Intl. Conference on 3D Body scanning Technologies. Lugano, Switzerland, pp. 1–10.
- Bittner, A., Glenn, F., Harris, R., Iavecchia, H., Wherry, R., 1987. CADRE: A family of mannikins for workstation design. In: Asfour, S.S. (ed.) Trends in Ergonomics/Human Factors IV. North Holland. pp. 733–740.
- Bittorf, V., Recht, B., Re, C., Tropp, J. A., 2012. Factoring nonnegative matrices with linear programs. In: Advances in Neural Information Processing Systems (NIPS). pp. 1214–1222.
- Blanchonette, P., 2010. Jack human modelling tool: A review. Tech. Rep. DSTO-TR-2364, Defence Science and Technology Organisation (Australia). Air Operations Division.
- Brown, A., 1999. Role models. *Mechanical Engineering* 121 (7), 44–49.
- Canhasi, E., Kononenko, I., 2013. Multi-document summarization via archetypal analysis of the content-graph joint model (doi: 10.1007/s10115-013-0689-8). *Knowledge and Information Systems*, 1–22.
- 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.
- Cutler, A., Breiman, L., November 1994. Archetypal Analysis. *Technometrics* 36 (4), 338–347.
- De Raeve, A., De Smedt, M., Bossaer, H., 2012. Mass customization, business model for the future of fashion industry. In: 3rd Global Fashion International Conference. pp. 1–17.
- D’Esposito, M. R., Palumbo, F., Ragozini, G., 2012. Interval Archetypes: A New Tool for Interval Data Analysis. *Statistical Analysis and Data Mining* 5 (4), 322–335.
- Elhamifar, E., Sapiro, G., Vidal, R., 2012. See all by looking at a few: Sparse modeling for finding representative objects. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8.
- Epifanio, I., 2013. h-plots for displaying nonmetric dissimilarity matrices. *Statistical Analysis and Data Mining* 6 (2), 136–143.
- 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, M. J., Leisch, F., April 2009. From Spider-Man to Hero - Archetypal Analysis in R. *Journal of Statistical Software* 30 (8), 1–23.  
URL <http://www.jstatsoft.org/>
- Eugster, M. J. A., 2012. Performance profiles based on archetypal athletes. *International Journal of Performance Analysis in Sport* 12 (1), 166–187.
- Eugster, M. J. A., Leisch, F., 2011. Weighted and robust archetypal analysis. *Computational Statistics and Data Analysis* 55 (3), 1215–1225.
- Faust, M. E., Carrier, S., 2009. 3D body scanning’s contribution to the use of apparel as an identity construction tool. In: Proceedings of the 2nd International Conference on Digital Human Modeling. pp. 19–28.
- Frey, B. J., Dueck, D., 2007. Clustering by passing messages between data points. *Science* 315, 972–976.

- Gillis, N., 2013. Robustness analysis of Hottopixx, a linear programming model for factoring nonnegative matrices. *SIAM Journal on Matrix Analysis and Applications* 34 (3), 1189–1212.
- Gillis, N., Vavasis, S. A., 2014. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (4), 698–714.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning. Data mining, inference and prediction.* 2nd ed., Springer-Verlag.
- HFES 300 Committee, 2004. *Guidelines For Using Anthropometric Data In Product Design.* Human Factors and Ergonomics Society.
- Hsu, C.-H., 2009. Data mining to improve industrial standards and enhance production and marketing: An empirical study in apparel industry. *Expert Systems with Applications* 36, 4185–4191.
- Ibáñez, M. V., Simó, A., Domingo, J., Durá, E., Ayala, G., Alemany, S., Vinué, G., Solves, C., 2012. A statistical approach to build 3D prototypes from a 3D anthropometric survey of the Spanish female population. In: *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods.* pp. 370–374.
- Ibáñez, M. V., 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), 10512 – 10520.
- International Organization for Standardization, 2004. *ISO 15537:2004: Principles for selecting and using test persons for testing anthropometric aspects of industrial products and designs.*
- Kaufman, L., Rousseeuw, P. J., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, New York.
- Kennedy, K., Zehner, G., 1995. Assessment of anthropometric accommodation in aircraft cockpits. *SAFE Journal* 25 (1), 51–57.
- Lawson, C. L., Hanson, R. J., 1974. *Solving Least Squares Problems.* Prentice Hall.

- Li, S., Wang, P., Louviere, J., Carson, R., December 2003. Archetypal Analysis: A New Way To Segment Markets Based On Extreme Individuals. In: ANZMAC 2003 Conference Proceedings. pp. 1674–1679.
- Mardia, K., Kent, J., Bibby, J., 1979. *Multivariate Analysis*. Academic Press.
- 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.
- Mohamed, S., Heller, K. A., Ghahramani, Z., 2014. *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC, Boca Raton, Florida, Ch. A simple and general exponential family framework for partial membership and factor analysis.
- Mørup, M., Hansen, L. K., 2012. Archetypal analysis for machine learning and data mining. *Neurocomputing* 80, 54–63.
- Porzio, G. C., 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, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.  
URL <http://www.R-project.org>
- Rasband, J. A., Liechty, E. L., 2006. *Fabulous Fit: Speed Fitting and Alteration*. 2nd ed., Fairchild Publication, New York, USA.
- Richards, J., Lee, A., Schafer, C., Freeman, P., 2012. Prototype selection for parameters in complex models. *The Annals of Applied Statistics* 6 (1), 383–408.
- Robinette, K. M., Hudson, J. A., 2006. *Handbook of Human Factors and Ergonomics*. John Wiley & Sons, New York, Ch. 12. Anthropometry, pp. 322–339.
- Rothwell, P., Hickey, D., 1986. Three-dimensional computer models of man. In: Proceedings of the Human Factors Society 30th Annual Meeting. pp. 216–220.

- Seiler, C., Wohlrabe, K., 2013. Archetypal scientists. *Journal of Informetrics* 7 (2), 345–356.
- Simmons, K., Istook, C. L., Devarajan, P., 2004. Female figure identification technique (FFIT) for apparel. Part I: Describing female body shapes. *Journal of Textile and Apparel, Technology and Management* 4, 1–16.
- Stone, E., 2002. Exploring archetypal dynamics of pattern formation in cellular flames. *Physica D* 161, 163–186.
- Theodosiou, T., Kazanidis, I., Valsamidis, S., Kontogiannis, S., 2013. Courseware usage archotyping. In: *Proceedings of the 17th Panhellenic Conference on Informatics. PCI '13*. ACM, New York, NY, USA, pp. 243–249.
- Thøgersen, J. C., 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.
- Thureau, C., Bauckhage, C., 2009. Archetypal images in large photo collections. In: *Proceedings of the 3rd IEEE International Conference on Semantic Computing*. pp. 129–136.
- Thureau, 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.
- Veitch, D., Fitzgerald, C., et al., 2013. *Sizing Up Australia - The Next Step*. Canberra: Safe Work.
- Vinué, G., Epifanio, I., Simó, A., Ibáñez, M. V., Domingo, J., Ayala, G., 2014a. *Anthropometry: An R Package for Analysis of Anthropometric Data*. R package version 1.0.
- Vinué, G., León, T., Alemany, S., Ayala, G., 2014b. Looking for representative fit models for apparel sizing. *Decision Support Systems* 57 (0), 22–33.
- Wish, M., 1971. *Attitude Research Reaches New Heights*. Chicago: American Marketing Association, Ch. Individual Differences in Perception and Preferences Among Nations, pp. 312–328.