SORT 44 (1) January-June 2020, 39-66 DOI: 10.2436/20.8080.02.94

Finding archetypal patterns for binary questionnaires

Ismael Cabero[1] and Irene Epifanio[2]

Abstract

Archetypal analysis is an exploratory tool that explains a set of observations as mixtures of pure (extreme) patterns. If the patterns are actual observations of the sample, we refer to them as archetypoids. For the first time, we propose to use archetypoid analysis for binary observations. This tool can contribute to the understanding of a binary data set, as in the multivariate case. We illustrate the advantages of the proposed methodology in a simulation study and two applications, one exploring objects (rows) and the other exploring items (columns). One is related to determining student skill set profiles and the other to describing item response functions.

MSC: 62H99, 62P25, 97D60.

Keywords: Dichotomous item test, archetypal analysis, functional data analysis, item response theory, skill profile.

1 Introduction

Mining binary survey data is of utmost importance in social sciences. Many raw data from exams, opinion surveys, attitude questionnaires, etc. come in the form of a binary data matrix, i.e. examinees' responses are coded as 0/1 (1 if examinee i answers item h correctly, otherwise 0). The binary matrix can be viewed from two points of views. In the first, the interest lies in the rows, i.e. in the people, while in the second, the interest lies in the columns that contain the items or variables. In both cases, exploratory data analysis (EDA) aims to find information in data and generate ideas (Unwin, 2010). In order to be useful as a tool for EDA on data sets, the tool should be simple and easy to use, with few parameters, and reveal the salient features of the data in such a way that humans can visualize them (Friedman and Tukey, 1974).

For the first time, we propose the use of the exploratory tool Archetypoid Analysis (ADA) for this kind of data in order to understand, describe, visualize and extract information that is easily interpretable, even by non-experts. ADA is an unsupervised

[1] Departamento de Didáctica de las Matemáticas, Universidad Internacional de La Rioja, Spain. Email: ismael.cabero@unir.net

[2] Departament de Matemàtiques-IMAC-IF, Universitat Jaume I, Castelló 12071, Spain. Email: epifanio@uji.es

Received: March 2019
Accepted: February 2020

statistical learning technique (see Hastie, Tibshirani and Friedman, 2009, Chapter 14) for a complete review of unsupervised learning techniques). Its objective is to approximate sample data by a convex combination (a mixture) of k pure patterns, the archetypoids, which are extreme representative observations of the sample. Being part of the sample makes them interpretable, but also being extreme cases facilitates comprehension of the data. Humans understand the data better when the observations are shown through their extreme constituents (Davis and Love, 2010) or when features of one observation are shown as opposed to those of another (Thurau et al., 2012).

ADA was proposed by Vinué et al. (2015) for real continuous multivariate data as a derivative methodology of Archetype Analysis (AA). AA was formulated by Cutler and Breiman (1994), and like ADA, it seeks to approximate data through mixtures of archetypes, also for real continous multivariate data. However, archetypes are not actual cases, but rather a mixture of data points. Recently, Seth and Eugster (2016b) proposed a probabilistic framework of AA (PAA) to accommodate binary observations by working in the parameter space.

AA and ADA have been applied to many different fields, such as astrophysics (Chan, Mitchell and Cram, 2003), biology (D'Esposito, Palumbo and Ragozini, 2012), climate (Steinschneider and Lall, 2015; Su et al., 2017), developmental psychology (Ragozini, Palumbo and D'Esposito, 2017), e-learning (Theodosiou et al., 2013), finance (Moliner and Epifanio, 2019), genetics (Thøgersen et al., 2013), human development (Epifanio 2016; Epifanio, Ibáñez and Simó, 2020), industrial engineering (Epifanio et al., 2013; Epifanio, Ibáñez and Simó, 2018; Millán-Roures, Epifanio and Martínez, 2018; Alcacer et al., 2020), machine learning (Mørup and Hansen, 2012; Seth and Eugster, 2016a,b; Ragozini and D'Esposito, 2015; Cabero and Epifanio, 2019), market research (Li et al., 2003; Porzio, Ragozini and Vistocco, 2008; Midgley and Venaik, 2013), multidocument summarization (Canhasi and Kononenko, 2013, 2014), nanotechnology (Fernandez and Barnard, 2015), neuroscience (Tsanousa, Laskaris and Angelis, 2015; 2016) and sports (Eugster, 2012; Vinué and Epifanio, 2017, 2019).

Archetypal analysis techniques lie somewhere in between two well-known unsupervised statistical techniques: Principal Component Analysis (PCA) and Cluster Analysis (CLA). In data decomposition techniques, a data set is viewed as a linear combination of several factors to find the latent components. Different prototypical analysis tools arise depending on the constraints on the factors and how they are combined (Mørup and Hansen, 2012; Vinué, Epifanio and Alemany, 2015). The factors with the least restrictions are those produced by PCA, since they are linear combinations of variables. One of the advantages is that this helps explain the variability of the data; however, the interpretability of the factors is compromised. Instead, the greatest restrictions are found in cluster tools, such as k-means or k-medoids. Their factors are readily interpreted because they are centroids (means of groups of data) or medoids (concrete observations) in the case of k-means and k-medoids, respectively. The price that clustering tools pay for interpretability is their modeling flexibility due to the binary assignment of data to the clusters. Archetypal tools, on the other hand, enjoy higher modeling flexibility than

cluster tools but without losing the interpretability of their factors. A table summarizing the relationship between several unsupervised multivariate techniques is provided by Mørup and Hansen (2012) and Vinué et al. (2015).

AA and ADA were originally thought of for real-valued observations. The aim of this work is to extend archetypal tools to binary data. For AA, as the factors (archetypes) are a mixture of data, they would not necessarily be binary vectors, and as a consequence they would not be interpretable. In ADA though, the factors (archetypoids) are actual cases, so ADA can be applied to binary data without losing the interpretability of the factors. So, among the possible archetypal techniques (AA, PAA and ADA), we propose to use ADA for binary data.

To perform a sanity check and provide insight we analyze the solutions obtained by AA, PAA and ADA through a simulation study, where ADA shows its appropriateness versus AA or PAA for binary data sets. Furthermore, we present two real applications and compare ADA solutions with those of other established unsupervised techniques to illustrate the advantages of ADA in educational and behavioral sciences, when used as another useful tool for data mining in these fields (Slater et al., 2017). In the first application, we are interested in rows, while in the second application in columns.

The outline of the paper is as follows: In Section 2 we review AA and ADA for real-valued multivariate and functional data and PAA, besides other multivariate techniques used in the comparison. In Section 3 we introduce the analysis for binary multivariate data. In Section 4, a simulation study with binary data compares the different strategies for obtaining archetypal patterns. In Section 5, our proposal is applied to two real data sets and compared to the results of other well-known unsupervised statistical learning techniques. Section 6 contains conclusions and some ideas for future work.

The data sets and code in R (R Development Core Team, 2019) for reproducing the results for both artificial and real data are available at http://www3.uji.es/ epifanio/ RESEARCH/adaedu.rar.

2 Preliminary

2.1 AA and ADA in the real-valued multivariate case

Let $\mathbf{X}$ be an $n \times m$ real-valued matrix with n observations and m variables. Three matrices are established in AA: a) the k archetypes $\mathbf{z}_j$, which are the rows of a $k \times m$ matrix $\mathbf{Z}$; b) an $n \times k$ matrix $\alpha = (\alpha_{ij})$ with the mixture coefficients that approximate each observation $\mathbf{x}_i$ by a mixture of the archetypes ($\hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$); and c) a $k \times n$ matrix $\beta = (\beta_{jl})$ with the mixture coefficients that characterize each archetype ($\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$). To figure out these matrices, we minimize the following residual sum of squares (RSS) with the respective constraints ($\|\cdot\|$ denotes the Euclidean norm for vectors):

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

under the constraints

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

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

As previously mentioned, archetypes do not necessarily match real observations. Indeed, this will only happen when one and only one β_{jl} is equal to one for each archetype, i.e. when each archetype is defined by only one observation. So, in ADA the previous constraint 2) is substituted by the following one, and as a consequence in ADA a mixed-integer problem is optimized instead of the AA continuous optimization problem:

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

As regards the location of archetypes, they are on the boundary of the convex hull of the data if $k > 1$ (see Cutler and Breiman, 1994), although this does not necessarily happen for archetypoids (see Vinué et al., 2015). Nonetheless, the archetype is equal to the mean and to the medoid in case of the archetypoid (Kaufman and Rousseeuw, 1990), if $k = 1$.

We want to emphasize that archetypal analysis is an EDA technique based on a geometric formulation (no distribution of data is assumed). It is not an inferential statistical technique, i.e. it is not about fitting models, parameter estimation, or testing hypotheses. Nevertheless, a field to study in the future would be to view archetypal analysis as a feature extraction method (Hastie et al., 2009, Ch. 5), where the raw data are preprocessed and described by α, which can be used as inputs into any learning procedure for compositional data (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015).

2.1.1 Computation of AA and ADA

The estimation of the matrices in the AA problem can be achieved by means of an alternating minimizing algorithm developed by Cutler and Breiman (1994), where the best α for given archetypes $\mathbf{Z}$ and the best archetypes $\mathbf{Z}$ for a given α are computed by turns. To solve the convex least squares problems, a penalized version of the non-negative least squares algorithm by Lawson and Hanson (1974) is used. Eugster and Leisch (2009) implemented that algorithm in the R package **archetypes**, although with some changes. Specifically, the data are standardized and the spectral norm in equation

1 is used instead of Frobenius norm for matrices. In our R implementation those changes were annulled, i.e. the data are not standardized by default and the objective function to minimize is defined by equation 1.

With respect to the estimation of the matrices in the ADA problem, it can be achieved using the algorithm developed by Vinué et al. (2015). It is composed of two steps: the BUILD step and the SWAP step. The objective of the BUILD step is to determine an initial set of archetypoids that will be upgraded during the following step. The objective of the SWAP step is to improve the primitive set by exchanging the selected instances for unselected observations and checking whether these replacements decrease the RSS. Vinué (2017) implemented that algorithm in the R package **Anthropometry** with three possible original sets in the BUILD step: $cand_{ns}$, $cand_{\alpha}$ and $cand_{\beta}$. These sets correspond to the nearest neighbor observations in Euclidean distance to the k archetypes, the cases with the maximum α value for each archetype j and the observations with the maximum β value for each archetype j, respectively. Then three possible solutions are obtained once these three sets go through the SWAP step, but only the solution with lowest RSS (often the same final set is returned from the three initializations) is chosen as the ADA solution.

One important point is the selection of k, since archetypes are not necessarily nested and neither are archetypoids. If the user has prior knowledge of the structure of the data, the value of k can be chosen based on that information. Otherwise, a simple but effective heuristic (Cutler and Breiman, 1994; Eugster and Leisch, 2009; Vinué et al., 2015; Seth and Eugster, 2016b) such as the elbow criterion can be used. With the elbow criterion, we plot the RSS for different k values and the value of k is selected as the point where the elbow is located.

2.1.2 Illustrative example

In Figure 1 a toy two-dimensional data set is used to illustrate what archetypoids mean and the differences compared with CLA and PCA, as well as to provide some intuition on what these pure and extreme patterns imply in behavioral sciences. Two numeric variables are considered from the data set personality-1.0 of the R package **neuropsychology** (Makowski, 2016), which contains personality traits data from an online questionnaire: Empathy.Agreeableness and Honesty.Humility. We apply k-means and ADA with $k = 3$, i.e. we find 3 clusters and archetypoids. We also apply PCA. Archetypoids are people with extreme values, which have clear profiles: archetypoid 1 is characterized by a very low Empathy.Agreeableness value together with a high Honesty.Humility value (1, 5.25), archetypoid 2 has the maximum values for both Empathy.Agreeableness and Honesty.Humility (7,7), while the third archetypoid has a very high Empathy.Agreeableness value together with the lowest Honesty.Humility value (6,0). Archetypoids are the purest people. The rest of the individuals are expressed as mixtures (collected in alpha coefficients) of these ideal people. For example, an individual with values of 6.25 and 0.75 for Empathy.Agreeableness and Honesty.Humility, respectively, is explained by 11% of

archetypoid 2 plus 89% of archetypoid 3.

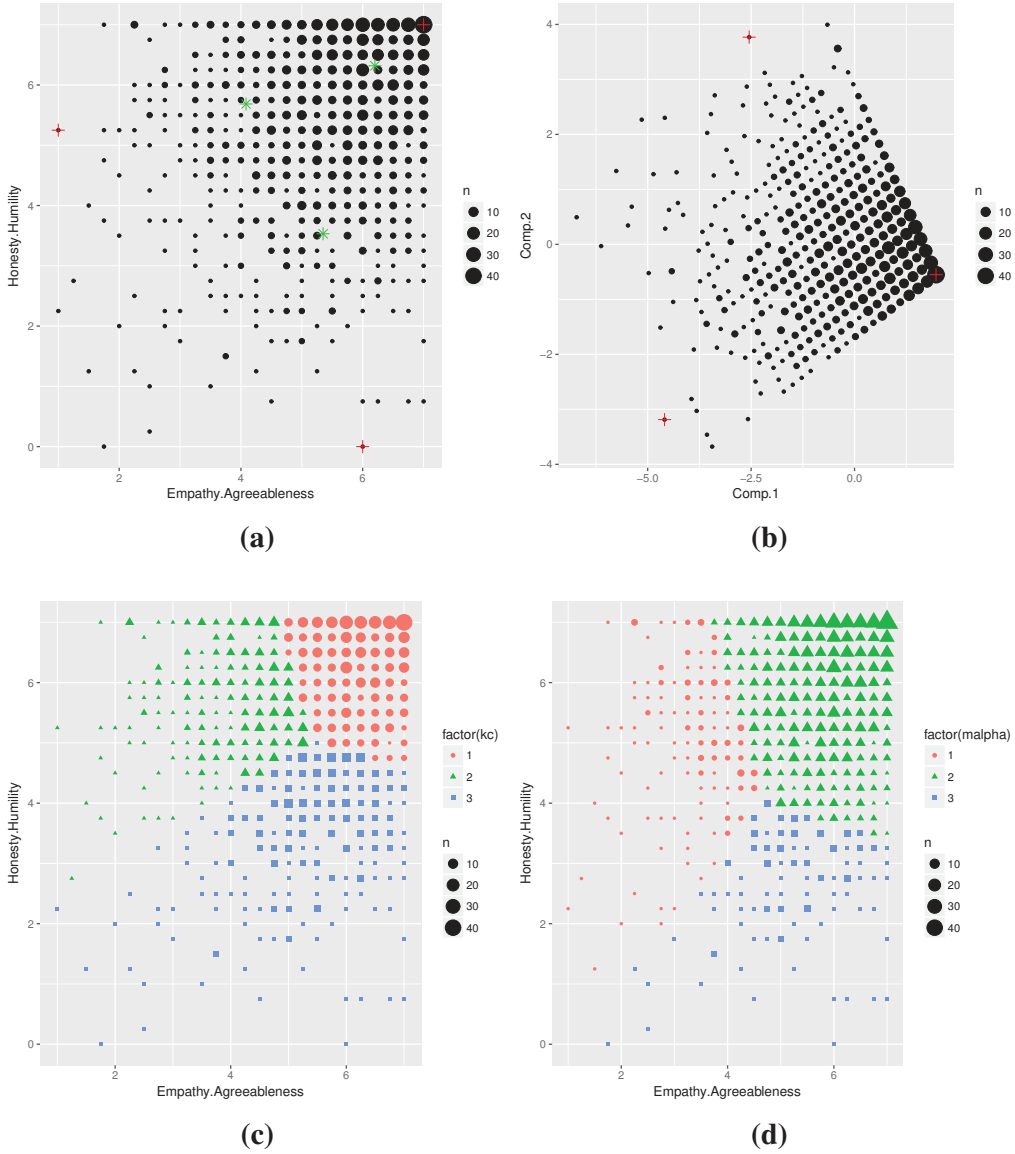

Figure 1: *(a) Plot of the toy example. The size of the points depends on their frequency. The red crosses represent the archetypoids, while the green stars represent the centroids of each cluster; (b) PC scores. Projected archetypoids are represented by red crosses; (c) k-means cluster assignments; (d) ADA assignments by the maximum alpha, i.e. assigned to the archetypoid that best explains the corresponding observation.*

This is compatible with the natural tendency of humans to represent a group of objects by its extreme elements (Davis and Love, 2010). Figure 1 d) shows the partition of the set generated by assigning the observations to the archetypoid that best explains

each individual. However, when we apply k-means to this kind of data set, without differentiated clusters, the centroids are in the middle of the data cloud. Centroid profiles are not as differentiated from each other as archetypoid profiles. This happens because centroids have to cover the set in such a way that the set is partitioned by minimizing the distance with respect to the assigned centroid (see Wu, Kamar and Horvitz, 2016) about the connection between set partitioning and clustering). On the one hand, this means that the set partition generated by k-means and ADA would be different (Figures 1 c) and d)). On the other hand, centroids are not the purest, and therefore their profiles are not as clear as those of archetypoids. For example, centroids 2 and 3 have values (4.1, 5.7) and (5.3, 3.5), which are not as intuitively interpretable as archetypoids. If we look again at the individual with values (6.25, 0.75) from the clustering point of view this individual is clearly assigned to cluster 3, with centroid (5.3, 3.5), but clustering does not say anything about the distance of this point with respect to the assigned centroid, or in which direction they are separated. In fact, (6.25, 0.75) is quite far from (5.3, 3.5). This happens because the objective of clustering is to assign the data to groups, not to explain the structure of the data more qualitatively. Finally, note that archetypoids do not coincide with the individuals with the most extreme PC scores (see Figure 1 b)).

In summary, depending on our objective, the appropriate analysis should be selected. The objective of PCA is to reduce data dimension. Although PCA returns the location of the observations in the new dimensions by PC scores, there is no guarantee that the principal components are interpretable. In other words, observations are expressed in a new base, but in general the PCA base is not easily interpretable. However, the objective of CLA is to segment the data, i.e. to make groups of data by finding modes in data. Although the modes can be easily interpretable, CLA does not return an expression about the location of each observation with respect to each mode. On the other hand, finding extreme profiles, which are easily interpretable, is not the objective of PCA or CLA, but that of AA or ADA. These techniques also return the location of the observations as a function of the extreme profiles, in fact as a mixture (a convex combination), which is more easily interpretable than a linear combination. This provides a complete overview of the data set, generally supported by visual methods, i.e. this allows data to tell us more beyond the formal modeling or hypothesis testing task.

2.2 Probabilistic archetype analysis

The idea underlying PAA is to work in a parameter space instead of the observation space, since the parameter space is often vectorial even if the sample space is not. The key is to assume that data points come from a certain distribution (from the Bernoulli distribution in the case of binary observations). Then the maximum likelihood estimates of the parameters of the distributions are seen as the parametric profiles that best describe each observation, and archetypal profiles are computed in the parameter space by maximizing the corresponding log-likelihood under the constraints for α and β. In sum-

mary, probabilistic archetypes lie in the parameter space, whereas classical archetypes lie in the observation space. Thus, archetypal profiles for binary data are the probability of a positive response. Details can be found in Seth and Eugster (2016b).

2.3 AA and ADA in the functional case

In Functional Data Analysis (FDA) each datum is a function. Therefore, the sample is a set of functions $\{x_1(t),\ldots,x_n(t)\}$ with $t \in [a,b]$, i.e. the values of the m variables in the standard multivariate context are replaced by function values with a continuous index t. We assume that these functions belong to a Hilbert space, satisfy reasonable smoothness conditions and are square-integrable functions on that interval. Simplistically, the sums are replaced by integrals in the definition of the inner product.

AA and ADA were extended to functional data by Epifanio (2016). In the functional context, functions from the data set are approximated by mixtures of archetypal functions. In functional archetype analysis (FAA), we seek k archetype functions that approximate the functional data sample by their mixtures. In other words, the objective of FAA is the same as AA, but now both archetypes ($z_j(t)$) and observations ($x_i(t)$) are functions. As a consequence, RSS is now calculated with a functional norm instead of a vector norm. We consider the L^2-norm, $\|f\|^2 =< f,f >= \int_a^b f(t)^2 dt$. The interpretation of matrices α and β is the same as in the classical multivariate case.

Analogously, FADA is also a generalization of ADA, where k functional archetypoids, which are functions of the sample, approximate the functions of the sample through the mixtures of these functional archetypoids. Again, vectors are replaced by functions and vector norms by functional norms, and the matrices are interpreted is the same way as before.

To obtain FAA and FADA in a computationally efficient way (Epifanio (2016)), functional data are represented by means of basis functions (see Ramsay and Silverman (2005) for a detailed explanation about smoothing functional data). Let B_h ($h = 1,\ldots,m$) be the basis functions and $\mathbf{b}_i$ the vector of coefficients of length m such that $x_i(t) \approx \sum_{h=1}^{m} b_i^h B_h(t)$. Then, RSS is formulated as (see Epifanio (2016) for details):

$$RSS = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} z_j\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l\|^2 = \sum_{i=1}^{n} \mathbf{a}_i^{\mathsf{T}} \mathbf{W} \mathbf{a}_i, \tag{2}$$

where $\mathbf{a}_i^{\mathsf{T}} = \mathbf{b}_i^{\mathsf{T}} - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} \mathbf{b}_l^{\mathsf{T}}$ and $\mathbf{W}$ is the order m symmetric matrix with the inner products of the pairs of basis functions $w_{m_1,m_2} = \int B_{m_1} B_{m_2}$. If the basis is orthonormal, for instance the Fourier basis, $\mathbf{W}$ is the order m identity matrix and FAA and FADA can be estimated using standard AA and ADA with the basis coefficients. If not, $\mathbf{W}$ has to be calculated previously one single time by numerical integration.

2.4 Other unsupervised learning techniques

The following well-known multivariate analysis tools for binary data are used in the comparison. We use homogeneity analysis (multiple correspondence analysis) using the R package **homals** (de Leeuw and Mair, 2009) (HOMALS). HOMALS can be considered as an equivalent to PCA in the case of categorical data. For CLA we use Partitioning Around Medoids (PAM) from the R package **cluster** (Maechler et al., 2018; Kaufman and Rousseeuw, 1990), since it returns representative objects or medoids among the observations of the data set. The pairwise dissimilarities between observations in the data set needed for PAM are computed with the *daisy* function from the R package **cluster** (Maechler et al., 2018), specifically using Gower's coefficient (Gower, 1971) for binary observations. Other popular clustering methods (Flynt and Dean, 2016) are also used in the comparison: latent class analysis (LCA) from the R package **poLCA** (Linzer and Lewis, 2011), which is a finite mixture model clustering for categorical data, and classical k-means clustering (Lloyd, 1982). It is used in the literature (Henry et al., 2015), despite not being recommended for binary data (IBM Support, 2016). For that reason, we also consider PAM, since it is a robustified version of k-means (Steinley, 2006) that can be used with distances other than Euclidean, and observations, rather than centroids, serve as the exemplars for each cluster.

3 Archetypal analysis for binary data

Let $\mathbf{X}$ be an $n \times m$ binary matrix with n observations and m variables. The idea behind archetypal analysis is that we can find a set of archetypal patterns, and that data can be expressed as a mixture of those archetypal patterns. In the case of binary data, on the one hand the archetypal patterns should also be binary data, as the population from which data come. For example, if pregnancy was one of the binary variables, it would not make sense to consider as an archetypal observation a woman who was pregnant 0.7. In other words, archetypal patterns should be binary in order to have a clear meaning and not lose their interpretability, which is the cornerstone of archetypal techniques, i.e. they should not be 'mythological', but rather something that might be observed. On the other hand, in order to describe data as mixtures, we should assume that observations exist in a vector space, i.e. that observations can be multiplied by scalars (in this case in the interval $[0, 1]$) and added together.

A solution that meets all these ideas is to apply ADA to $\mathbf{X}$, since the feasible archetypal patterns belong to the observed sample. In fact, ADA was originally created as a response to the problem in which pure non-fictitious patterns were sought (Vinué et al., 2015).

Instead, the archetypes returned by applying AA or PAA do not need to be binary, i.e. they do not need to belong to the feasible set of solutions. In fact, Seth and Eugster (2016b) binarized the archetypes obtained by AA or PAA in experiments. However,

using a continuous optimization problem to solve a problem whose feasible solutions are not continuous can fail badly (Fletcher, 2000, Ch. 13). Indeed, there is no guarantee that this approach will provide a good solution, even by examining all the feasible binary solutions in a certain neighborhood of the continuous solution.

Therefore, we propose to use ADA to handle binary observations.

4 Simulation study

We have carried out a simulation study to assess all the alternatives in a controlled scenario. The design of the experiment has been based on simulation studies that appear in Vinué et al. (2015) and Seth and Eugster (2016b). We generate $k = 6$ archetypes, ζ_i, with $m = 10$ binary variables by sampling them from a Bernoulli distribution with a probability of success $p = 0.7$, $A = [\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5, \zeta_6]$. Given the archetypes, we generate $n = 100$ observations as the binarized version of $x_i = \tilde{A}_i h_i + E_i$, where $\tilde{A}_i$ contains the archetypes after adding salt and pepper noise to them, h_i is a random vector sampled from a Dirichlet distribution with $\alpha = (0.8, 0.8, 0.8, 0.8, 0.8, 0.8)$, and E_i is a 10-dimensional random vector of Gaussian white noise with a mean of zero and standard deviation of 0.1. The binarized versions are obtained by replacing all values above 0.5 with 1 and others with 0. The noise density added to A is 0.05 (the default value used in MATLAB). With salt and pepper noise, a certain amount of the data is changed to either 0 or 1. To ensure that $\tilde{A}_i$'s are archetypes, we chose $\alpha = 0.8$, a value near to but less than one.

We compute PAA, AA and ADA. The archetypes returned by PAA and AA are binarized for comparison with the true ones, A. We calculated the Hamming distance (Manhattan distance between binary vectors), which is the same as the misclassification error used with binary images, between each archetypal solution and the true archetypes after permuting the columns of each archetypal solution to match the true archetypes in such a way that the least error with the city block distance is provided.

This was repeated 100 times. The first 10 times are displayed in Figure 2. The solutions returned by all the methods are quite similar to the true archetypes, i.e. the number of errors (a zero in the solution where the true value is 1, or vice versa) is very small. Nevertheless, there are differences between the methods, which are more evident in columns 5 and 6. For columns 5 and 6, the number of errors for PAA is 5 and 5, it is 4 and 2 for AA, but only 2 and 2 for ADA. Table 1 shows a the summary of the

Table 1: *Summary of misclassification errors of the archetype profiles for each method over 100 simulations.*

Method	PAA	AA	ADA
Mean (Std. dev.)	4.20 (1.86)	3.59 (1.99)	3.19 (1.88)

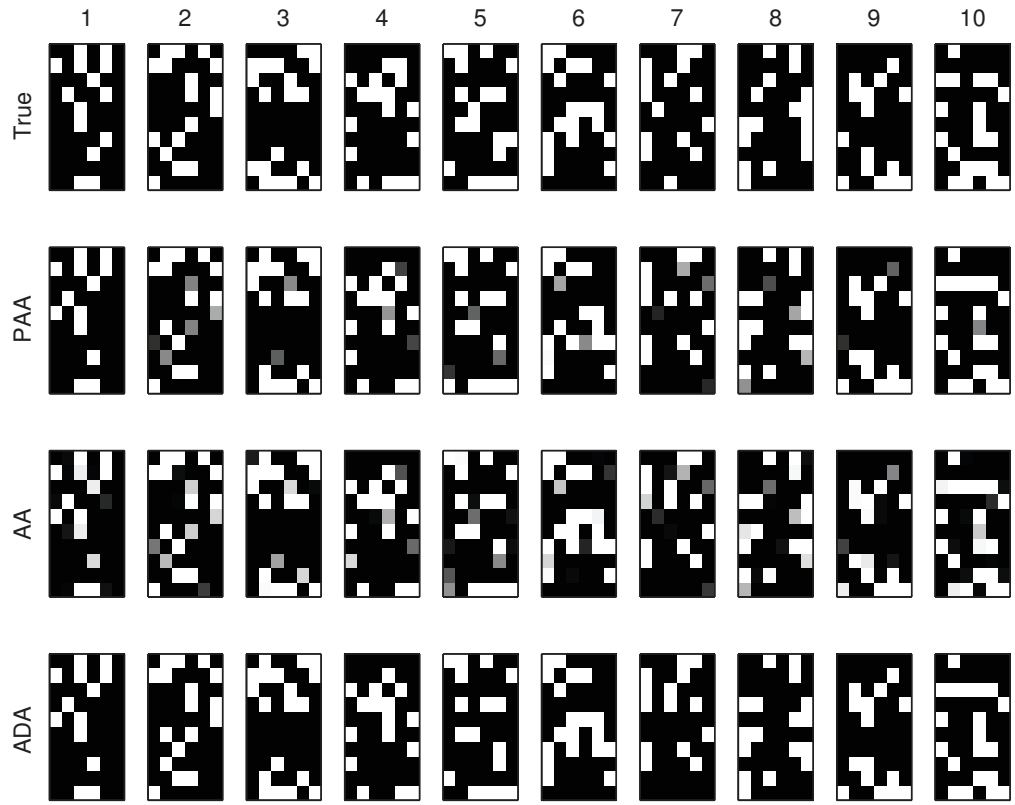

Figure 2: *Comparison between true archetypes and those returned by PAA, AA and ADA, respectively. The 10 columns represent the first 10 repetitions of the simulation. Black represents 0 and white 1.*

misclassifications. The archetypoids returned by ADA match the true archetypes better than those returned by AA or PAA, in this order, i.e. ADA provides the smallest mean misclassification error.

5 Applications

5.1 An initial mathematical skills test for first-year university students

5.1.1 Data

The first application corresponds to the first point of view of the binary matrix (analysis of the rows). We analyze the data set described by Orús and Gregori (2008), which was obtained through the application of a test on the initial mathematics skills of 690 first-year students of the College of Technology and Experimental Sciences at Jaume I University (Spain) at the beginning of the 2003-04 academic year. The test consisted

of 17 questions corresponding to 21 single items, the answers to which were coded as 0 (incorrect or unanswered) or 1 (correct). The items of the test were selected in order to ascertain some given didactic hypotheses on the didactic discontinuities between mathematics at pre-university and university levels. It is not a test designed to rank the students and return a unique score. The complete description of the questions can be seen in Orús and Gregori (2008). With ADA, we could obtain students' skill set profiles. In this way, students can be grouped by their similar mastery of skills. For instance, students showing consistently high levels of aptitude may be selected for an advanced class or students with similar difficulties could receive extra instruction as a group and also teaching strategies could also be adapted to suit their level. A classical way to group student skill set profiles is by using a clustering method, as carried out by Dean and Nugent (2013), but in terms of human interpretability, the central points returned by clustering tools do not seem as favorable as the extreme points returned by ADA. Results from different exploratory tools are compared.

5.1.2 Results and discussion

We would like to estimate the skill set profiles hidden in the data set. In other words, we would like to discover the data structure. Our intuition tells us that skill sets vary continuously across students, i.e. we do not expect there to be clearly differentiated (separate) groups of students with different abilities. Even so, CLA has been used to generate groups of students with similar skill set profiles (Chiu, Douglas and Li, 2009; Dean and Nugent, 2013). Here, we are going to consider the raw binary data and let the data speak for themselves, as ADA is a data-driven method. We compare the ADA solution with others from well-established unsupervised techniques introduced in Section 2.4 to highlight the information about the quality understanding of data provided by ADA.

For the sake of brevity and as an illustrative example, we examine the results of $k = 3$. The RSS elbow for ADA and the Bayesian Information Criterion (BIC) elbow for LCA are found at $k = 3$ (see Figure 3). According to the silhouette coefficient (a method of interpretation and validation of consistency within clusters of data, see Kaufman and Rousseeuw (1990) for details), the optimal number of clusters are $k = 2$ and $k = 3$ for PAM. However, the highest value of the silhouette coefficient is 0.22 (for $k = 2$ and $k = 3$ clusters), which means that no substantial cluster structure was found, as we predicted. We perform an h-plot (a multidimensional scaling method that is particularly suited for representing non-Euclidean dissimilarities, see Epifanio (2013) for details) on the dissimilarities used by PAM to graphically summarize the data set and to visualize the obtained clusters by PAM in two dimensions (see Figure 4). Effectively, separate clusters do not seem to exist.

This is also corroborated by Figure 5, where the students' scores from HOMALS are plotted in two dimensions. As regards the interpretation of the dimensions of HOMALS,

the loadings are displayed in Figure 6 and Table 2 shows their exact values, together with the number of correct answers. As also happens with PCA, their interpretation is not always easy and immediate. For the first dimension, all the coefficients are positive (as a measure of size), which can indicate a kind of sum score. The highest coefficients more or less correspond to the last questions of the test, which fewer students answered correctly. The second dimension compares, above all, questions 4, 5, 6a and 6b (with high positive coefficients) with 13a and 13b (with low negative values), while in the third dimension, questions 1, 3, 7, 8 and 10 (with high positive coefficients) are compared with 14a and 14b (with low negative values). However, we do know how the meaning of these contradistinctions is interpreted.

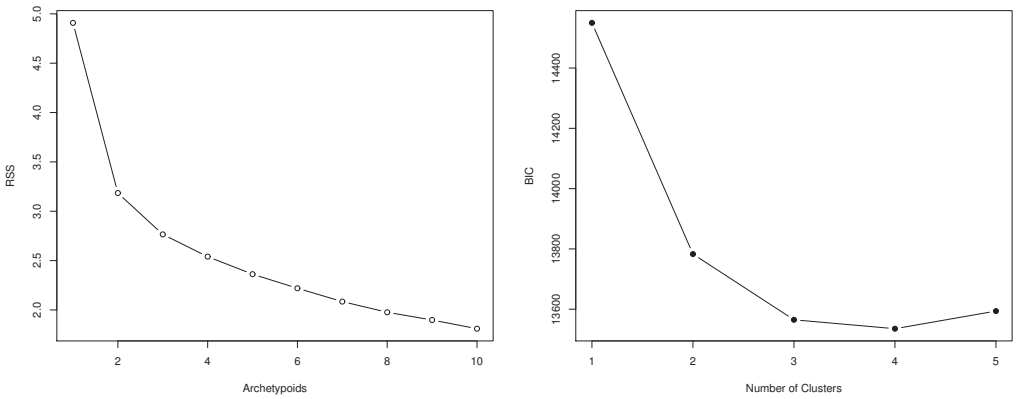

Figure 3: *Initial mathematical skills test data: Screeplot of ADA (left-hand panel); screeplot of LCA (right-hand panel).*

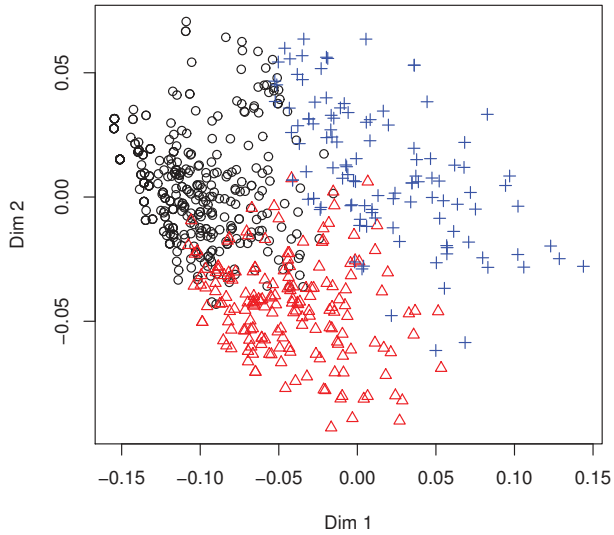

Figure 4: *H-plot of dissimilarities for the initial mathematical skills test data. We perform PAM. The black circles represent data points assigned to the first cluster, the red triangles to the second cluster and the blue crosses to the third cluster.*

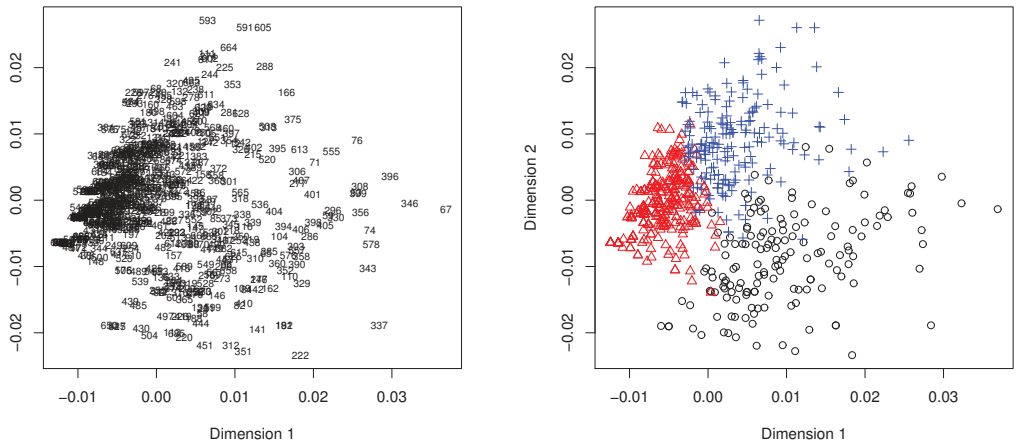

Figure 5: *HOMALS of the initial mathematical skills test data. Plot of students' scores. The numbers indicate the code of each of the 690 students (left-hand panel). We perform LCA. The black circles represent data points assigned to the first cluster, the red triangles to the second cluster and the blue crosses to the third cluster (right-hand panel).*

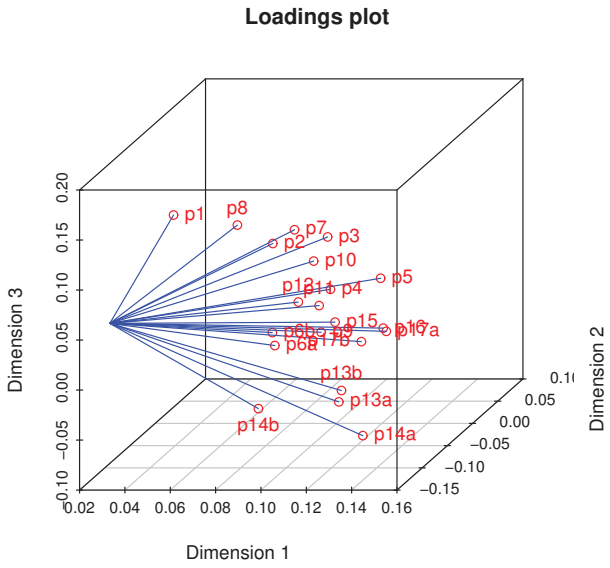

Figure 6: *HOMALS of the initial mathematical skills test data. Loadings plot.*

LCA returns the conditional item response probabilities by outcome variable for each class. Table 3 lists these probabilities for correct answer. The predicted classes for each student are shown in Figure 5, since the profiles of cluster 1 and 3 are mainly differentiated in questions 4, 5, 13a and 13b, which are the most relevant variables of dimension 2 of HOMALS.

Table 2: *Number of correct answers and loadings of the first three dimensions by HOMALS for the initial mathematical skills test data.*

Question	No. correct answers	D1	D2	D3
1	621	0.02863	−0.00232	0.10936
2	589	0.05797	0.06268	0.05200
3	301	0.09098	0.02291	0.07621
4	233	0.07597	0.09596	−0.00863
5	253	0.09922	0.09102	0.00465
6a	231	0.05414	0.08397	−0.05947
6b	105	0.05230	0.08735	−0.04804
7	270	0.07408	0.03324	0.07881
8	140	0.0601	−0.01729	0.10596
9	109	0.07749	0.07059	−0.04028
10	202	0.09541	−0.02469	0.07325
11	71	0.07485	0.07870	−0.01717
12	329	0.08006	0.01380	0.01518
13a	177	0.12934	−0.12748	−0.02185
13b	132	0.12953	−0.12312	−0.01241
14a	114	0.11951	−0.03559	−0.09643
14b	22	0.07565	−0.04544	−0.06506
15	183	0.10749	−0.03665	0.01754
16	236	0.12062	−0.00018	−0.00465
17a	47	0.12116	0.00354	−0.00952
17b	62	0.10884	0.00955	−0.02249

Table 3 also lists the profiles of the medoids, centroids of k-means and the archetypal profiles for AA, PAA and ADA. For medoids and archetypoids, the code of the corresponding observation is also displayed. To facilitate the analysis we also show the binarized profiles of AA and PAA, referred as BAA and BPAA, respectively.

As a simple summary of the profiles, we compute the percentage of correct answers for each profile. For PAM, the percentages are 9.5%, 33.3% and 57.1%; for binarized LCA, 38.1%, 9.5% and 33.3%; for binarized k-means, 38.1%, 9.5% and 42.9%; for BAA, 9.5%, 47.6% and 61.9%; for BPAA, 9.5%, 42.9% and 57.1%; and for ADA, 57.1%, 52.4% and 9.5%, respectively. Note that the median of the percentage of correct answers in the data set is 28.6% (the minimum is 0, the first quartile is 19.1%, the third quartile is 38.1%, while the maximum is 95.2%).

One profile is repeated in all the methods, a student who only answers questions 1 and 2 correctly, i.e. a student with a serious lack of competence. We therefore concentrate the analysis on the other two profiles for each method.

Table 3: *Profiles for the initial mathematical skills test data, for PAM, LCA, k-means (k-M), AA (and binarized, BAA), PAA (and binarized, BPAA) and ADA, with k = 3. The numbers in brackets for PAM and ADA indicate the code of the representative student.*

Methods	1	2	3	4	5	6a	6b	7	8	9	10	11	12	13a	13b	14a	14b	15	16	17a	17b
PAM (661)	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PAM (586)	1	1	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
PAM (162)	1	1	1	0	1	0	0	1	0	0	1	0	1	1	1	1	0	1	1	0	0
LCA 1	0.93	0.88	0.57	0.36	0.48	0.37	0.17	0.49	0.32	0.20	0.51	0.14	0.62	1.00	0.85	0.43	0.10	0.50	0.55	0.18	0.19
LCA 2	0.88	0.79	0.28	0.17	0.15	0.24	0.07	0.30	0.14	0.03	0.15	0.02	0.32	0.05	0	0.03	0	0.13	0.14	0	0
LCA 3	0.91	0.94	0.60	0.62	0.65	0.48	0.27	0.47	0.23	0.35	0.36	0.23	0.63	0.04	0	0.20	0.03	0.31	0.53	0.09	0.17
k-M 1	0.91	0.95	0.64	0.70	0.74	0.43	0.25	0.53	0.24	0.32	0.38	0.22	0.63	0.05	0.00	0.19	0.02	0.33	0.52	0.10	0.15
k-M 2	0.88	0.78	0.26	0.13	0.11	0.27	0.09	0.27	0.13	0.05	0.14	0.02	0.32	0.06	0.01	0.03	0.01	0.12	0.15	0.01	0.01
k-M 3	0.93	0.91	0.59	0.34	0.49	0.37	0.17	0.50	0.33	0.20	0.54	0.14	0.64	1	0.88	0.46	0.10	0.52	0.58	0.18	0.19
AA 1	0.85	0.68	0.02	0	0	0.05	0	0.04	0.05	0	0.01	0	0.16	0	0	0	0	0.01	0	0	0
AA 2	0.90	1	0.87	1	1	1	0.63	0.82	0.19	0.52	0.24	0.38	0.65	0	0	0.16	0	0.07	0.43	0.09	0.15
AA 3	1	1	0.89	0.32	0.53	0.19	0.06	0.71	0.58	0.26	1	0.17	1	1	1	0.67	0.18	1	1	0.36	0.37
BAA 1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BAA 2	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0
BAA 3	1	1	1	0	1	0	0	1	1	0	1	0	1	1	1	1	0	1	1	0	0
PAA 1	0.86	0.72	0.13	0	0	0	0	0.12	0.07	0	0	0	0.23	0	0	0	0	0	0	0	0
PAA 2	0.90	1	0.78	1	1	1	0.61	0.73	0.27	0.40	0.38	0.31	0.66	0	0	0	0	0	0.43	0	0
PAA 3	0.99	1	0.82	0.36	0.57	0.25	0	0.66	0.44	0.27	0.86	0.12	0.85	1	1	0.73	0.15	1	1	0.32	0.42
BPAA 1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BPAA 2	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
BPAA 3	1	1	1	0	1	0	0	1	0	0	1	0	1	1	1	1	0	1	1	0	0
ADA (182)	1	1	1	0	0	0	0	0	1	0	1	0	1	1	1	1	0	1	1	0	1
ADA (274)	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	1	0	0
ADA (1)	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

In contrast with the third archetypoid, i.e. the student with very poor skills, the first and second archetypoids correspond to students with very high percentages of correct answers. In fact, the first archetypoid corresponds to the 92nd percentile of the data set, while the second archetypoid corresponds to the 88th percentile. Nevertheless, both profiles are quite different. In fact, the Hamming distance between archetypoids 1 and 2 is 13, which means that although they answered a lot of items correctly, these correctly answered items do not coincide. In other words, archetypoids 1 and 2 are somehow complementary. Both answered items 1, 2, 3, 12 and 16 correctly, which were among the most correctly answered items. Neither of them answered items 11, 14b and 17a correctly, which were among the least correctly answered items. On the one hand, the items that archetypoid 1 answered correctly, but archetypoid 2 did not are 8, 10, 13a, 13b, 14a, 15 and 17b. These items are about nonlinear systems and linear functions. On the other hand, the items that archetypoid 2 answered correctly, but

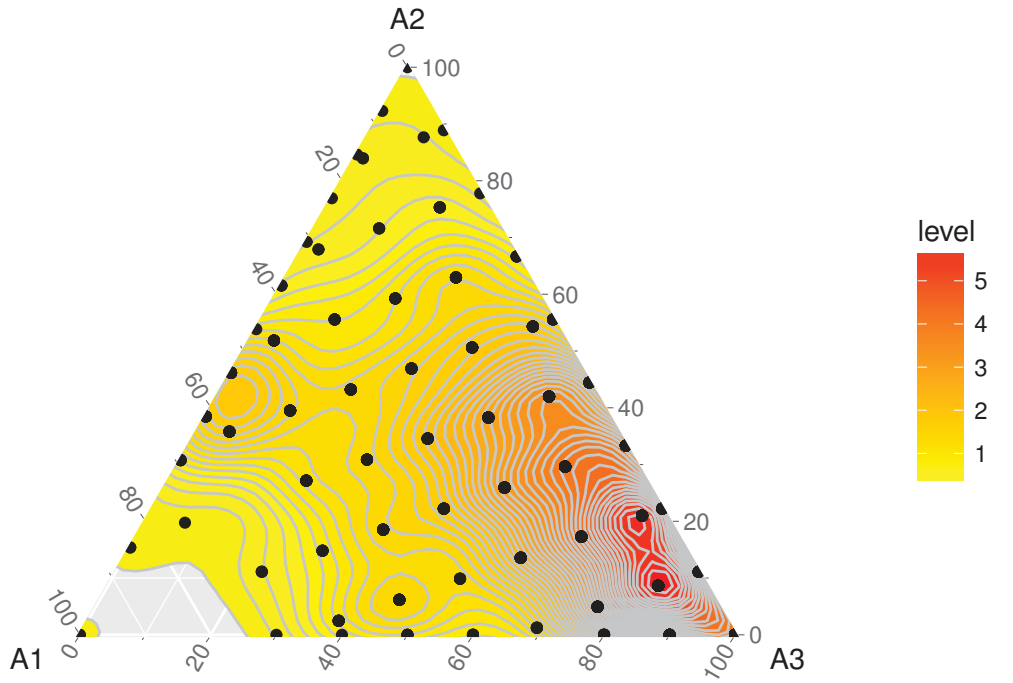

Figure 7: *Ternary plot of αs of ADA together with a plot density estimate for the initial mathematical skills test data.*

archetypoid 1 did not are 4, 5, 6a, 6b, 7 and 9. These items are about the calculation of derivatives and integrals and algebraic interpretation. The skills of these archetypoids are clear and different to each other.

We can use the alpha values for each of the students to learn about their relationship to the archetypoid profiles. The ternary plot in Figure 7 displays the alpha values that provide further insight into the data structure. Note that the majority of the data is concentrated around archetypoid 1, i.e. the one with very poor skills. If we wanted to form three groups using the alpha values, we could assign each student to the group in which their corresponding alpha is the maximum, as we did in Figure 1 (d). In this way, the number of students similar to archetypoid 1 is 113, to archetypoid 2 it is 110 and to archetypoid 3 it is 467.

The profiles of medoids 2 and 3 are not as complementary as the previous archetypoids. In fact, medoid 2 corresponds to the 56th percentile, while medoid 3 corresponds to the 92nd percentile. In this case, the percentage of correct answers for medoid 2 is not high. The Hamming distance between the two medoids is only 7. On the one hand, both answered items 1, 2, 3, 5, 7 and 12 correctly, which are the most correctly answered items. On the other hand, both failed items 6a, 6b, 8, 9, 11, 14b, 17a and 17b, many more items than in the case of ADA. The only item that medoid 2 answered correctly but

medoid 3 did not is item 4. The items that medoid 3 answered correctly but medoid 2 did not are 10, 13a, 13b, 14a, 15 and 16. It seems as if the cluster definition was guided by the number of correct answers rather than by the kind of item answered correctly. This is the reason why PAM selects medoid 2 in the middle of the data cloud. PAM, and usual clustering methods, tries to cover the set in such a way that every point is near to one medoid or one cluster center. The number of students belonging to each cluster is 398, 179 and 113, respectively. Note that the size of the cluster of students with poor skills is smaller than in the case of ADA, because some of those students are assigned to the cluster of medoid 2.

The binarized profile of LCA 1, corresponding to the 75th percentile, is similar to medoid 3, but with a lower number of correct answers (5, 7, 14a and 15), while the binarized profile of LCA 3, corresponding to the 56th percentile, is similar to medoid 2, only differentiated by two items (7 and 16). Therefore, they are even less complementary than the previous medoids. The Hamming distance between both LCA-profiles is only 5. The number of students belonging to each cluster is 155, 352 and 183, respectively. Note that the size of the cluster of students with poor skills is smaller than in the case of PAM.

The binarized profile of the first centroid of k-means, corresponding to the 75th percentile, is similar to medoid 2, only differentiated by item 16, while the binarized profile of the third centroid, corresponding to the 82nd percentile, is similar to medoid 3, but with a lower number of correct items (5, 7 and 14a). The Hamming distance between both centroids is 7. The level of complementarity between both centroids is similar to that of the medoids of PAM, but the number of correct answers of medoid 3 is higher than binarized centroid 3. The number of students belonging to each cluster is 196, 349 and 145, respectively. Note that the size of the cluster of students with poor skills is smaller than in the case of PAM, but larger than in the case of PAM for cluster 3, which in both clustering methods corresponds to the students with more correct answers.

In the clustering methods, the profiles of each cluster are not as extreme as archetypoids. Archetypoids are also more complementary, which makes it clearer to establish which kinds of features distinguish one group from another. Remember also that clustering is limited to assign each student to a group but alpha values of ADA allows to know the composition, i.e. ADA returns a richer information.

The profiles of BAA2 and BAA3 and BPAA2 and BPAA3 are quite similar to the profiles of archetypoid 2 and 1, respectively, but with slight differences. The percentiles corresponding to correctly answered items are also high, although for one of the archetypes not as high as for archetypoids. The percentiles are the 82nd and 94th for BAA2 and BAA3, and the 75th and 92nd for BPAA2 and BPAA3, respectively. Therefore, the archetypoids are more extreme than the binarized archetypes of AA and PAA. Although the profiles for BAA and BPAA are also complementary, they are not as complementary as the two archetypoids. The Hamming distance between BAA2 and BAA3 is 11, and 9 between BPAA2 and BPAA3. Archetypoids therefore manage to find more complementary profiles.

5.2 An American College Testing (ACT) Mathematics Test

5.2.1 Data

This application corresponds to the second point of view of the binary matrix (analysis of the columns). We use the same data and approach followed by Ramsay and Silverman (2002, Ch. 9) and Rossi, Wang and Ramsay (2002), although another strategy could be considered (Ramsay and Wiberg, 2017). The data used are the 0/1 (incorrect/correct) responses of 2115 males from administration of a version of the ACT Program 60-item Mathematics Test. Unlike the test introduced in Section 5.1.1, the objective of the test is to relate a student's ACT score with probability of him or her earning a college degree, i.e. to rank students. It seeks that the difficulty of questions increases as you get to higher question numbers.

Although this binary matrix does not seem curvaceous at first sight, by making the simplifying assumption that the probabilities P_{ih} (probability that examinee h gets item i right) vary in a smooth one-dimensional way across examinees, we can estimate the ability space curve that this assumption implies. Then, we can work with item response functions (IRFs) $P_i(\theta)$ as functional data (Ramsay and Silverman, 2005), where θ is the charting variable that measures out positions along the ability space curve. Or rather, we can work with the log odds-ratio functions $W_i(\theta)$, since these transformations of the item response functions have the unconstrained variation that we are used to seeing in directly observed curves. Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) used functional PCA (FPCA) to study variations among these functions. Instead, we propose to use functional ADA (FADA), which reveals very interesting patterns that were not discovered with FPCA.

Note that in the literature, we find other terms for IRFs, such as option characteristic curves, category characteristic curves, operating characteristic curves, category response functions, item category response functions or option response functions (Mazza, Punzo and McGuire, 2014).

5.2.2 Results and discussion

As mentioned previously, we used the same data and approach followed by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) to estimate IRFs, $P_i(\theta)$, and their logit functions, $W_i(\theta) = \log(P_i(\theta)/(1 - P_i(\theta)))$. In particular, a penalized EM algorithm was used and functions were expanded by terms of 11 B-spline basis functions using equally spaced knots. Figure 8 displays the estimated IRFs, $\exp(W_i(\theta))/(1 + \exp(W_i(\theta)))$, and their log odds-ratio functions $W_i(\theta)$ for the 60 items. As expected, this kind of graphs with superimposed curves is largely uninformative and aesthetically unappealing (Jones and Rice, 1992).

To explore a set of curves Jones and Rice (1992) proposed the use of functions with extreme principal component scores. This could be viewed as finding the archetypoid functions. Nevertheless, the aim of PCA is not to recover extreme patterns. In fact,

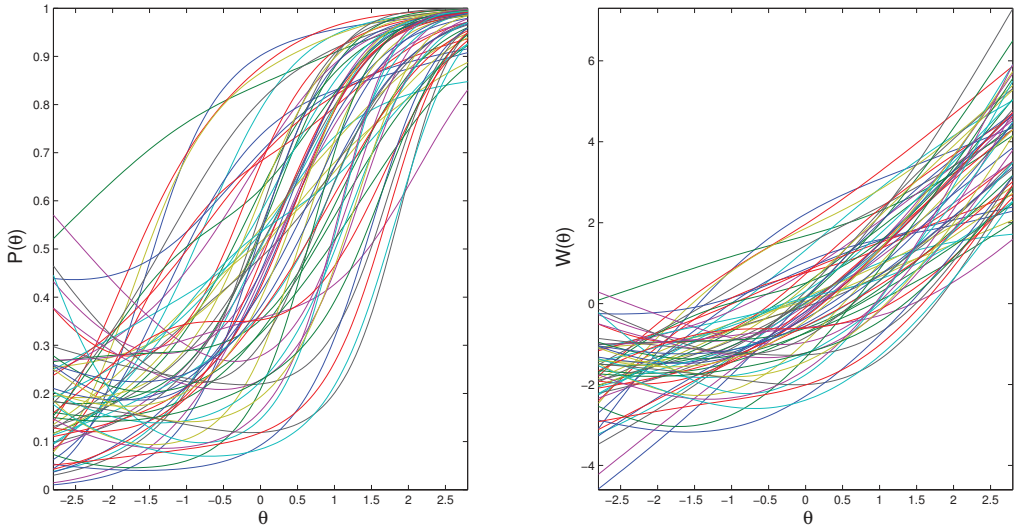

Figure 8: *Estimated IRFs (left-hand panel) and log odds-ratio functions (right-hand panel) for the ACT math exam estimated from the male data.*

curves with extreme PCA scores do not necessarily correspond to archetypal observations. This is discussed in Cutler and Breiman (1994) and shown in Epifanio, Vinué and Alemany (2013) through an example where archetypes could not be restored with PCA, even if all the components had been considered. Not only that, Stone and Cutler (1996) also showed that AA may be more appropriate than PCA when the data do not have elliptical distributions.

In order to show the advantages of ADA over PCA, we compute FPCA and FADA for $W(\theta)$, since they are unconstrained, therefore making them more appropriate for PCA application than the bounded $P_i(\theta)$. This is not a problem with FADA as it works with convex combinations. Figure 9 displays the first four PCs after a varimax rotation having been back-transformed to their probability counterparts, as performed by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002). We base the interpretation of each PC on the detailed description carried out by Ramsay and Silverman (2002, Ch. 9).

The percentage of total variation explained by those four components is nearly 100%, while the percentage explained by each component is reported in Figure 9. The first component concentrates on the middle part of the ability range, in such a way that an item with a high (low) score in that component has a higher (lower) slope than the mean from approximately 0 to 2, i.e. it quantifies a discriminability trade-off between average students and those with rather high abilities. Analogously, the fourth component quantifies a discriminability trade-off between average examinees and those with rather low abilities. On the contrary, the second component concentrates on the upper end of the ability range. As Rossi et al. (2002) explained, the 3PL model is not well suited to modeling this type of variation. An item with a low score on this component is good at sorting out very high ability students from others of moderately high ability, whereas if

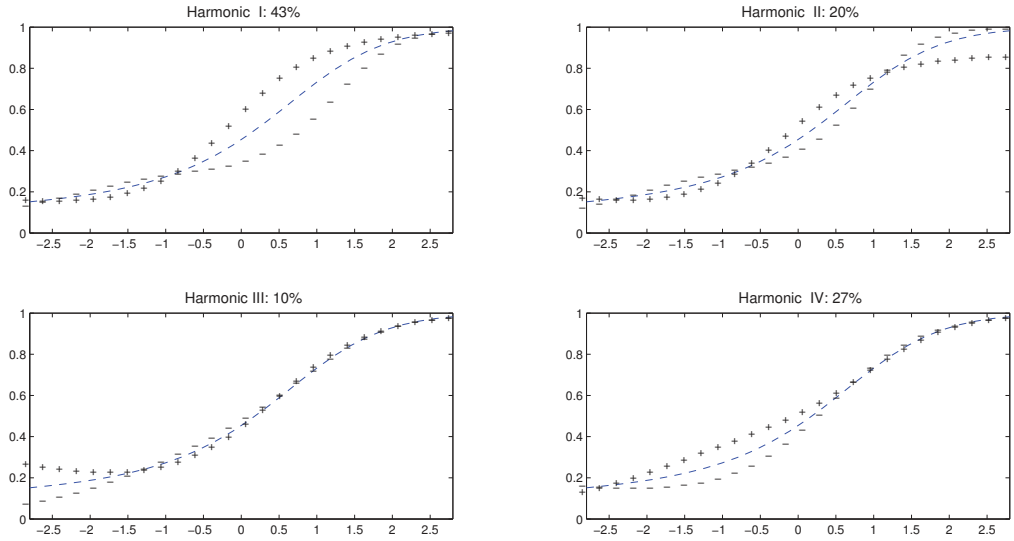

Figure 9: *The first four functional PCs in IRFs after VARIMAX rotation. Plus (negative) signs indicate the effect of adding (subtracting) a multiple of a component to the mean function. The mean function is the dashed line.*

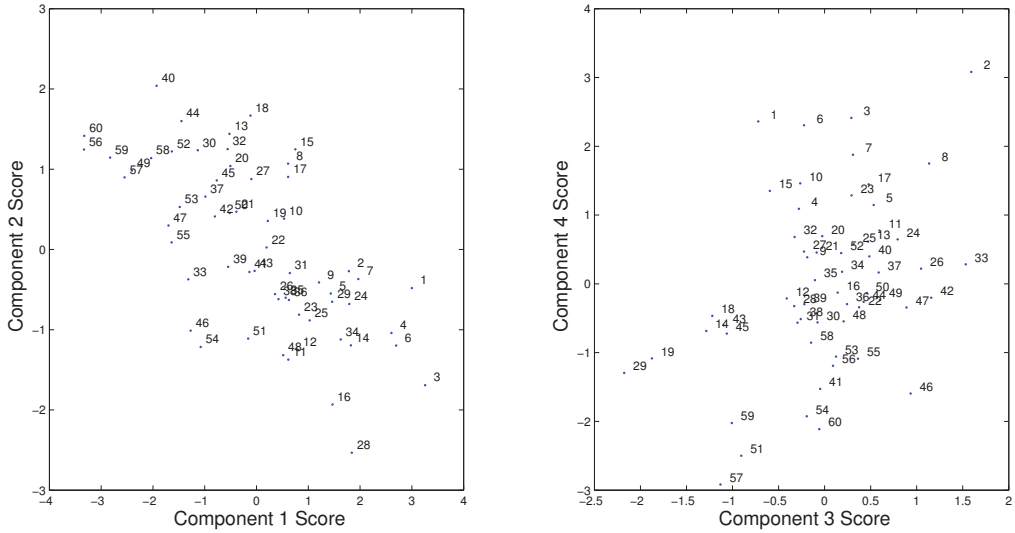

Figure 10: *Bivariate plots of principal component scores of IRFs. PC1 versus PC2 (left-hand panel); PC3 versus PC4 (right-hand panel).*

the score for this item is high, it will discriminate well among most of the population but will be found to be of approximately equal difficulty by all the very good students. Nevertheless, conclusions on the extreme part of the ability range should be made with

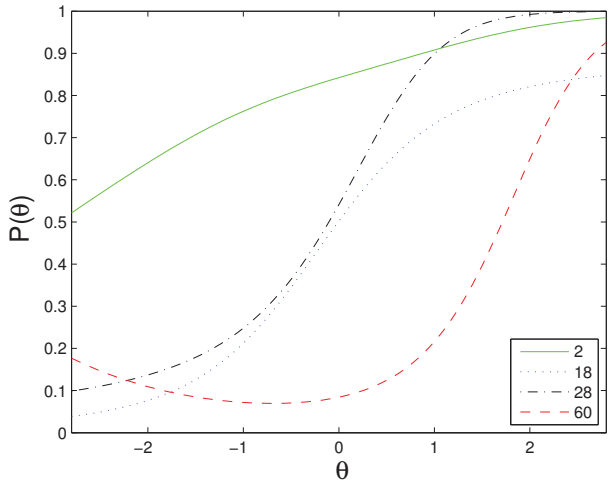

Figure 11: *ACT data: The four IRF archetypoids are items 2, 18, 28 and 60. See the legend inside the plot.*

caution, since the estimation is carried out using a relatively small numbers of students. The third component also accounts for variation in the characteristics of test items in the extreme ability range, but now in low ability ranges. PC scores for these four components can be seen in Figure 10. Note that to evaluate the 4 PC scores simultaneously and combine them to give an idea about each item, it is not easily comprehensible or human-readable.

Figure 11 displays the archetypoids for $k = 4$ explaining 97% of the variability, which is nearly as high as FPCA. The archetypoids are items 2, 18, 28 and 60. These four items describe the extreme patterns found in the sample. Item 2 has very high scores in PC 3 and PC 4, high scores in PC 1 and a score of nearly zero for PC 2. Its IRF is quite flat with a very slight slope, it seems to be a very easy item, with high probabilities of success throughout the ability range. The other archetypoids discriminate better between low and high ability students but in very different ways. Item 18 has a very high score for PC 2 and a negative score for PC 3, but nearly zero for PC 1 and PC 4. It is an item that is quite difficult even for the students in the very high ability range. The IRF of item 28 is quite similar to that of item 18 for the low ability range until θ 0, but its slope for the high ability range is higher, and the probabilities of success are higher than 0.9 for θs higher than 1. On the contrary, the probabilities of success of the IRF of item 60 are quite low as far as 1.5, which means that it is a difficult item, but the probabilities of success for the best students are high. In fact, the probabilities of success for item 60 for θ higher than 2 are higher than those of item 18. Item 28 has high score for PC 1 and low score for PC2, while it has a score of nearly zero for PC 3 and PC 4. However, item 60 has low scores for PC 1 and PC 4, a high score for PC 2 and nearly zero for PC 3. In other words, it would have been very difficult to guess the extreme representatives of the sample returned by ADA from an analysis of the scores in Figure 10.

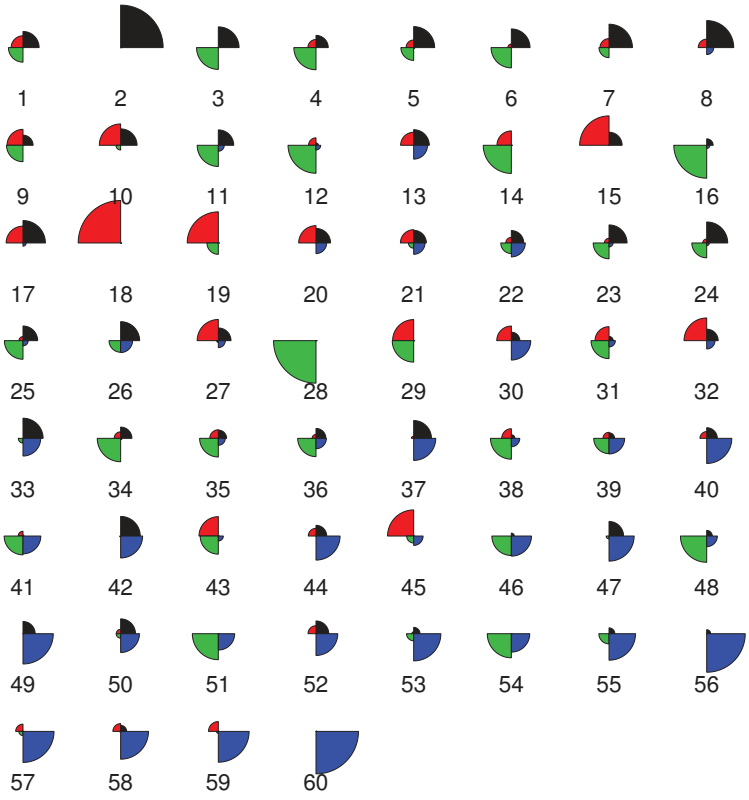

Figure 12: *Star plots of the alphas of each archetypoid for IRFs. The item number appears below each plot. The archetypoids are 2, 18, 28 and 60.*

The alpha values (from 0 to 1) tell us about the contribution of each archetypoid to each item. Remember that they add up to 1. Figure 12 shows star plots of the alpha values for each archetypoid, thus providing a complete human-readable view of the data set. The 4 alpha values in this case are represented starting on the right and going counter-clockwise around the circle. The size of each alpha is shown by the radius of the segment representing it. The items that are similar to the archetypoids can be clearly seen (for example, 7 and 8 are somehow similar to 2; 15 and 19 are somehow similar to 18; 14 and 16 are somehow similar to 28; and 56, 57 and 59 are similar to 60), as can the items that are a mixture of several archetypoids (for example, item 1 is a mixture of mainly item 2, together with items 28 and 18, to a lesser extent). Item 1 was selected by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) as an example of a low difficulty item, although it seems that item 2 would be a better representative of this kind of item. Item 9 was selected by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) as an example of a medium difficulty item, and it is mainly a mixture of items 18 and 28. Finally, item 59 was selected by Ramsay and Silverman (2002, Ch. 9) and Rossi et al. (2002) as an example of a hard item. Item 59 was mainly explained by item 60.

Results of applying FADA with kernel and parametric IRF estimates are discussed in the Supplementary Material.

6 Conclusion

We have proposed to find archetypal patterns in binary data using ADA for a better understanding of a data set. A simulation study and results provided in two applications have highlighted the benefits of ADA for binary questionnaires as an alternative that can be used instead of (or in addition to) other established methodologies.

Although, much of statistics is based on the idea that averaging over many elements of a data set is a good thing to do, in this paper we adopt a different perspective. We have selected a small number of representative observations, archetypal observations, and the data composition is explained through mixtures of those extreme observations. We have shown that this can be highly informative and is a useful tool for making a data set more "human-readable", even to non-experts.

In the first application, we have shown how ADA returns the most complementary profiles, which can be more useful in order to establish groups of students with similar mastery of skills. Furthermore, ADA returns composition information of each observation through alpha values, which is a richer information than the simple assignation to groups returned by CLA. In the second application, FADA has discovered the extreme patterns in the data, which cannot be recovered by FPCA. Furthermore, we have explained each item of the ACT math exam as a percentage of the archetypal items, which is easily understandable even for non-experts.

As regards future work, throughout the paper all variables share the same weight, but for certain situations some variables could have more weight in RSS. Another direction of future work would be to consider ADA for nominal observations, for example, by converting those variables into dummy variables, i.e. with binary codes. Furthermore, this work is limited to binary data, but questionnaires can also have Likert-type scale responses. Therefore, archetypal techniques for ordinal data would be very valuable. Another not so immediate extension, would be to consider the case of mixed data, with real valued and categorical data, together with missing data. Finally, from the computational point of view, in case of working with a very big data set, the ADA algorithm described in Section 2.1.1 could be slow. In that case, a recent alternative implemented in the R package **adamethods** (Vinue and Epifanio, 2019) for computing ADA with large data sets could be used.

Acknowledgments

This work is supported by the following grants: DPI2017-87333-R from the Spanish Ministry of Science, Innovation and Universities (AEI/FEDER, EU) and UJI-B2017-13 from Universitat Jaume I.

References

Alcacer, A., Epifanio, I., Ibáñez, M. V., Simó, A. and Ballester, A. (2020). A data-driven classification of 3D foot types by archetypal shapes based on landmarks. *PLOS ONE*, 15, 1–19.

Cabero, I. and Epifanio, I. (2019). Archetypal analysis: an alternative to clustering for unsupervised texture segmentation. *Image Analysis & Stereology*, 38, 151–160.

Canhasi, E. and Kononenko, I. (2013). Multi-document summarization via archetypal analysis of the content-graph joint model. *Knowledge and Information Systems*, 1–22.

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

Chan, B., Mitchell, D. and Cram, L. (2003). Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, 338, 790–795.

Chiu, C.-Y., Douglas, J. A. and Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74, 633.

Cutler, A. and Breiman, L. (1994). Archetypal analysis. *Technometrics*, 36, 338–347.

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

de Leeuw, J. and Mair, P. (2009). Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, 31, 1–20.

Dean, N. and Nugent, R. (2013). Clustering student skill set profiles in a unit hypercube using mixtures of multivariate betas. *Advances in Data Analysis and Classification*, 7, 339–357.

D'Esposito, M. R., Palumbo, F. and Ragozini, G. (2012). Interval archetypes: a new tool for interval data analysis. *Statistical Analysis and Data Mining*, 5, 322–335.

Epifanio, I. (2013). H-plots for displaying nonmetric dissimilarity matrices. *Statistical Analysis and Data Mining*, 6, 136–143.

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

Epifanio, I., Ibáñez, M. V. and Simó, A. (2018). Archetypal shapes based on landmarks and extension to handle missing data. *Advances in Data Analysis and Classification*, 12, 705–735.

Epifanio, I., Ibáñez, M. V. and Simó, A. (2020). Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. *The American Statistician*, 74, 169–183.

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

Eugster, M. J. and Leisch, F. (2009). From Spider-Man to Hero - Archetypal Analysis in R. *Journal of Statistical Software*, 30, 1–23.

Eugster, M. J. A. (2012). Performance profiles based on archetypal athletes. *International Journal of Performance Analysis in Sport*, 12, 166–187.

Fernandez, M. and Barnard, A. S. (2015). Identification of nanoparticle prototypes and archetypes. *ACS Nano*, 9, 11980–11992.

Fletcher, R. (2000). *Practical Methods of Optimization* (Second ed.). John Wiley & Sons.

Flynt, A. and Dean, N. (2016). A survey of popular R packages for cluster analysis. *Journal of Educational and Behavioral Statistics*, 41, 205–225.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23, 881–890.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857–871.

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

Henry, D., Dymnicki, A. B., Mohatt, N., Allen, J. and Kelly, J. G. (2015). Clustering methods with qualitative data: a mixed-methods approach for prevention research with small samples. *Prevention Science*, 16, 1007–1016.

Hinrich, J. L., Bardenfleth, S. E., Roge, R. E., Churchill, N. W., Madsen, K. H. and Mørup, M. (2016). Archetypal analysis for modeling multisubject fMRI data. *IEEE Journal on Selected Topics in Signal Processing*, 10, 1160–1171.

IBM Support (2016). Clustering binary data with K-Means (should be avoided). http://www-01.ibm.com/support/docview.wss?uid=swg21477401. Accessed: 2018-07-09.

Jones, M. C. and Rice, J. A. (1992). Displaying the important features of large collections of similar curves. *The American Statistician*, 46, 140–145.

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

Lawson, C. L. and Hanson, R. J. (1974). *Solving Least Squares Problems*. Prentice Hall.

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

Linzer, D. A. and Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42, 1–29.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28, 129–137.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K. (2018). *Cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.7-1.

Makowski, D. (2016). *Package 'neuropsychology': An R Toolbox for Psychologists, Neuropsychologists and Neuroscientists*. (0.5.0).

Mazza, A., Punzo, A. and McGuire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in item response theory. *Journal of Statistical Software*, 58, 1–34.

Midgley, D. and 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.

Millán-Roures, L., Epifanio, I. and Martínez, V. (2018). Detection of anomalies in water networks by functional data analysis. *Mathematical Problems in Engineering*, 2018 (Article ID 5129735), 13.

Moliner, J. and Epifanio, I. (2019). Robust multivariate and functional archetypal analysis with application to financial time series analysis. *Physica A: Statistical Mechanics and its Applications*, 519, 195–208.

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

Orús, P. and Gregori, P. (2008). *Fictitious Pupils and Implicative Analysis: a Case Study*, pp. 321–345. Berlin, Heidelberg: Springer.

Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

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

R Development Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

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

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

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*. Springer.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis* (2nd ed.). Springer.

Ramsay, J. O. and Wiberg, M. (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics*, 42, 282–307.

Rossi, N., Wang, X. and Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics*, 27, 291–317.

Seth, S. and Eugster, M. J. A. (2016a). Archetypal analysis for nominal observations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38, 849–861.

Seth, S. and Eugster, M. J. A. (2016b). Probabilistic archetypal analysis. *Machine Learning*, 102, 85–113.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S. and Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, 42, 85–106.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59, 1–34.

Steinschneider, S. and Lall, U. (2015). Daily precipitation and tropical moisture exports across the Eastern United States: An application of archetypal analysis to identify spatiotemporal structure. *Journal of Climate*, 28, 8585–8602.

Stone, E. and Cutler, A. (1996). Introduction to archetypal analysis of spatio-temporal dynamics. *Physica D: Nonlinear Phenomena*, 96, 110–131.

Su, Z., Hao, Z., Yuan, F., Chen, X. and Cao, Q. (2017). Spatiotemporal variability of extreme summer precipitation over the Yangtze river basin and the associations with climate patterns. *Water*, 9.

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

Thøgersen, J. C., Mørup, M., Damkiær, S., Molin, S. and 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. and Bauckhage, C. (2012). Descriptive matrix factorization for sustainability: Adopting the principle of opposites. *Data Mining and Knowledge Discovery*, 24, 325–354.

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

Unwin, A. (2010). Exploratory data analysis. In P. Peterson, E. Baker, and B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)*, pp. 156–161. Oxford: Elsevier.

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

Vinué, G. and Epifanio, I. (2017). Archetypoid analysis for sports analytics. *Data Mining and Knowledge Discovery*, 31, 1643–1677.

Vinue, G. and Epifanio, I. (2019). *Adamethods: Archetypoid Algorithms and Anomaly Detection*. R package version 1.2.

Vinué, G. and Epifanio, I. (2019). Forecasting basketball players' performance using sparse functional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12, 534–547.

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

Wu, C., Kamar, E. and Horvitz, E. (2016). Clustering for set partitioning with a case study in ridesharing. In *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1384–1388.