Biarchetype analysis: simultaneous learning of observations and features based on extremes

Aleix Alcacer, Irene Epifanio, and Ximo Gual-Arnau
E-mail: epifanio@uji.es

Abstract—We introduce a novel exploratory technique, termed biarchetype analysis, which extends archetype analysis to simultaneously identify archetypes of both observations and features. This innovative unsupervised machine learning tool aims to represent observations and features through instances of pure types, or biarchetypes, which are easily interpretable as they embody mixtures of observations and features. Furthermore, the observations and features are expressed as mixtures of the biarchetypes, which makes the structure of the data easier to understand. We propose an algorithm to solve biarchetype analysis. Although clustering is not the primary aim of this technique, biarchetype analysis is demonstrated to offer significant advantages over biclustering methods, particularly in terms of interpretability. This is attributed to biarchetypes being extreme instances, in contrast to the centroids produced by biclustering, which inherently enhances human comprehension. The application of biarchetype analysis across various machine learning challenges underscores its value, and both the source code and examples are readily accessible in R and Python at https://github.com/aleixalcacer/JA-BIAA.

Index Terms—Archetype analysis, biclustering, prototype, unsupervised learning.

1 Introduction

Cluster analysis (CLA) is one of the most widely used tools in exploratory data analysis. The idea of clustering is to make groups of observations in such a way that each group contains similar observations that are different to those of the rest of the groups. If the data consist of well-separated clusters, appropriate clustering techniques can obtain, on the one hand, the representative of each cluster (the mean or centroid of the cluster for the popular k-means technique), and, on the other hand, the assignations of each observation to one cluster, or a degree of belonging to each cluster for fuzzy clustering techniques.

However, CLA is also used as a segmentation technique in the absence of well-separated (clearly differentiated) clusters in data. Many times, data follow a fan-spread pattern, i.e. features vary continuously across observations. The centroids are located in the middle of the data cloud since data points have to be covered in such a way that the distance between them and the assigned centroid is minimized (see [1] about the relationship between CLA and set partitioning). In those cases, where data can be viewed as a superposition of various populations, it is of particular interest to use Archetype Analysis (AA) for segmenting [2].

Instead of segmenting on the centroids, AA segments on the extremes. AA was defined by [3]. The objective of AA is to represent the observations by means of a convex combination of archetypes, which in turn are convex combinations of observations. Archetypes or 'pure types' lie on the boundary of the convex hull of the data and are therefore extreme profiles. Being extreme instances rather than central instances makes human understanding and interpretation of data easier [4] since human cognition prefers extreme opposites [5]. An illustrative example of this was analyzed in [6], where CLA and AA were compared and archetypes were much more informative and understandable than centroids, because archetypes are further apart from each other than centroids.

Biclustering is a data mining technique introduced by [7], although it was popularized by [8], who applied it to gene expression data analysis. In biclustering (also known as block clustering, co-clustering, or two-mode clustering), rows (observations) and columns (features) of a data matrix are simultaneously clustered. An excellent overview of biclustering and fuzzy biclustering is found in [9]. Biclustering is widely used in biological and medical applications [10], especially in gene expression data [11], [12]. However, it is also applied in many other fields, such as marketing [13], psychology [14], recommender systems [15], sports [16], [17], website traffic [18], and many other pattern recognition applications, such as collaborative filtering, text mining, multimedia data processing and retrieval, etc. [10], [19].

In recent years, there has been growing interest in AA. On the one hand, there has been an increasing number of papers proposing efficient computational methods to calculate AA [20], [21], [22], [23], with applications in computer vision. On the other hand, AA has been applied in other very diverse fields, such as, climatology [24], [25], ergonomics [26], [27], genetics [28], [29], [30], image processing [31], [32], [33], [34], machine learning problems [20], [35], [36], [37], market research [38], multi-document summarization [39], nanotechnology [40], neuroscience [41], [42], sports [43], [44], [45] and sustainability [4]. Finally, other papers have proposed extensions and new methodologies derived from AA with applications in a broad spectrum of fields: kernel AA [20], AA with missing data [20], [46], robust AA [47], [48], interval archetypes [49], archetypoid analysis (ADA) [50], functional AA [51], data-driven prototype

• *A. Alcacer, I. Epifanio and X. Gual-Arnau are with the Department of Mathematics, Jaume I University, Castelló, 12071, Spain.*

Manuscript received April 19, 2005; revised August 26, 2015.

identification [52], archetypal networks [35], probabilistic AA [53], AA for nominal [6], [54] and ordinal observations [55], directional AA [56], AA for shapes [57], deep AA [37], [58], and outlier detection [59], [60], [61]. Nevertheless, no previous work has developed archetypal analysis for both rows and columns simultaneously, which we refer to as biarchetype analysis (biAA), co-archetype analysis or two-mode archetype analysis.

[12] reviews biclustering in biological and biomedical fields. They point out the need to improve the interpretability of biclustering results, and they also highlight that possible overlapping homogeneous submatrices have to be identified. This clashes with the idea of CLA, whose origin was to find separate (not overlapping) groups, but it is in the line with the basis of AA. Moreover, biclustering of human gene expression data has been used to identify phenotype–genotype associations in studies of common or rare diseases. Note that archetypes themselves are phenotypes [37]; in fact, archetypes have been used also to explain the evolutionary development of biological systems [62]. Therefore, putting all this together, it seems that biAA could be a reasonable alternative to biclustering in biology, as biAA could improve the interpretability of results. Nevertheless, the fields of application of biAA are not just restricted to biology; they would be the same as for biclustering, i.e. biAA can be applied to many pattern recognition problems.

Our contributions consist of defining biAA for the first time, proposing a computational method to calculate it, whose implementation is available in the R package biaa https://github.com/aleixalcacer/biaa and the Python package https://github.com/aleixalcacer/archetypes, showing how it works and the advantages of using archetypes (extremes) rather than the centroids of biclustering in an illustrative example, and finally, applying it to several real data sets in different fields to demonstrate the usefulness of biAA in various problems.

The outline of the paper is as follows: previous methodologies (CLA, biclustering, fuzzy biclustering, AA) are reviewed in a common framework in Sec. 2. In Sec. 3, biAA is defined and a computational procedure is proposed. An illustrative example is used to exemplify biAA and compare it to biclustering. In Sec. 4, our proposal is applied to three real data sets. Some conclusions and ideas for future work are provided in Sec. 5.

2 Background

Matrix factorization is our common framework for describing the established methods (as used in [63] for clustering) and our proposal. Let $\mathbf{X}_{n\times m}$ be a data matrix with n observations and m continuous features (they should be standardized in order to avoid problems if they measure different dimensions). Let $\alpha_{n\times k}$ and $\gamma_{c\times m}$ be matrices with values in $[0,1]$. α is the membership matrix of the observations, while γ is the membership matrix of the features. $\mathbf{Z}$ is the matrix of representative instances that approximates $\mathbf{X}$. The objective is to minimize: $\|\mathbf{X}-\alpha\mathbf{Z}\gamma\|^2$, with different constraints, where $\|.\|$ stands for the Frobenius norm.

2.1 Clustering

For clustering, $\mathbf{Z}$ is the matrix of centroids, which is computed by $\mathbf{Z} = (\alpha'\alpha)^{-1}\alpha'\mathbf{X}\gamma'(\gamma\gamma')^{-1}$, where $'$ denotes transpose.

k-means clustering: The constraints are: $\sum_{g=1}^{k}\alpha_{ig}=1$ with $\alpha_{ig}\in\{0,1\}$ for $i=1,\dots,n$ and $\gamma = \mathbf{I}_{m\times m}$ is the identity matrix of order m. The matrix $\mathbf{Z}_{k\times m} = (\alpha'\alpha)^{-1}\alpha'\mathbf{X}$ has the centroids of each one of k groups that partition the data set.

Fuzzy clustering: In soft clustering, each observation is assigned membership to each group. The restrictions are: $\sum_{g=1}^{k}\alpha_{ig}=1$ with $\alpha_{ig}\geq 0$ for $i=1,\dots,n$ and $\gamma = \mathbf{I}_{m\times m}$. Again, the matrix $\mathbf{Z}_{k\times m}$ has the centroids of each one of k groups.

Biclustering: This is also called double k-means with hard partitions by [63], where algorithms to solve it are proposed. The constraints are: $\sum_{g=1}^{k}\alpha_{ig}=1$ with $\alpha_{ig}\in\{0,1\}$ for $i=1,\dots,n$ and $\sum_{h=1}^{c}\gamma_{hj}=1$ with $\gamma_{hj}\in\{0,1\}$ for $j=1,\dots,m$. Now, the dimension of $\mathbf{Z}$ is $k\times c$, since there are k groups for observations and c groups of variables.

Fuzzy biclustering: This is also called fuzzy double k-means by [63]. The constraints are now continuous: $\sum_{g=1}^{k}\alpha_{ig}=1$ with $\alpha_{ig}\geq 0$ for $i=1,\dots,n$ and $\sum_{h=1}^{c}\gamma_{hj}=1$ with $\gamma_{hj}\geq 0$ for $j=1,\dots,m$. [9] proposed several algorithms for solving fuzzy double k-means with continuous data, called FDkM and FDkMpf (Fuzzy Double k-Means with polynomial fuzzifiers), whose Matlab implementations are available in [9].

Besides the previous framework, there are some proposals of model-based biclustering. In that case, it is supposed that data are generated by a mixture distribution, as in [64], referred to as BMM (Block Mixture Model). Instead of memberships, it returns the final posterior probabilities for rows and columns, in addition to the mean and variance of each co-cluster. This is implemented in the R package blockcluster [65].

2.2 Archetype analysis

In AA, $\mathbf{Z}_{k\times m}=\beta_{k\times n}\mathbf{X}_{n\times m}$, where $\sum_{l=1}^{n}\beta_{gl}=1$ with $\beta_{gl}\geq 0$ for $g=1,\dots,k$, i.e. the archetypes are mixture of the data. The other restrictions are: $\sum_{g=1}^{k}\alpha_{ig}=1$ with $\alpha_{ig}\geq 0$ for $i=1,\dots,n$ and $\gamma = \mathbf{I}_{m\times m}$. Therefore, the objective function to minimize subject to the previous constraints is:

$$\begin{aligned} RSS = \|\mathbf{X}-\alpha\mathbf{Z}\|^2 = \|\mathbf{X}-\alpha\beta\mathbf{X}\|^2 = \\ \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij}-\sum_{g=1}^{k}\alpha_{ig}z_{gj}\right)^2 = \\ \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij}-\sum_{g=1}^{k}\alpha_{ig}\left(\sum_{l=1}^{n}\beta_{gl}x_{lj}\right)\right)^2. \end{aligned} \tag{1}$$

The α coefficients determine how much each archetype contributes to the approximation of each observation, i.e. α_{ig} is the weight of the archetype g for the i-th observation. Archetypes are built as mixtures of observations weighted by β coefficients.

If $k = 1$, the archetype coincides with the mean, but with $k > 1$, the archetypes are located on the boundary of the convex hull of the data [3]. Archetypes are not necessarily nested, so different ks may reveal distinct structures of the data. Therefore, as happens in other unsupervised statistical learning procedures, the selection of the number k of prototypes has to be determined. If we have prior knowledge of the arrangement of the data, k can be selected based on this. Otherwise, we can use a simple but effective heuristic method, the elbow criterion, which has been used elsewhere [3], [66]. The elbow criterion consists of displaying the RSS for different k values and choosing the value k as the position where the elbow is located. This method is also used in clustering.

[3] proposed an alternating minimizing algorithm to find the matrices α and β that minimizes RSS. This consists of alternating between estimating the best α for given archetypes $\mathbf{Z}$, and the optimum archetypes $\mathbf{Z}$ for given α. In each phase, convex least squares problems have to be solved. They used a penalized version of the non-negative least squares algorithm [67].

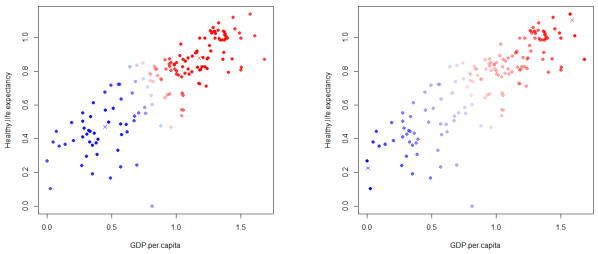
2.3 Illustrative example of fuzzy k-means versus AA

Fig. 1 illustrates the outcomes of applying fuzzy k-means and AA with $k = 2$ to variables associated with World Happiness, to demonstrate the concept of archetypes and their distinction from CLA. Contrary to CLA, AA models distinct aspects in the data rather than focus on the most central dynamics. With AA, two archetypal countries are identified as prototypes. One is associated with Utopia, an imaginary nation characterized by the world's happiest populace, and the other with Dystopia, an imaginary nation marked by the world's least happy populace, commonly used in sociology as a benchmark. Countries are described as mixtures (encapsulated in α coefficients) of these idealized nations. This approach aligns with the human tendency to represent a group of objects by its extreme elements [5]. However, with fuzzy k-means, the prototypes are situated in the middle of the data cloud; they are not the purest, hence their profiles are not as distinct as those of the archetypes. AA qualitatively offers a better explanation of the data structure. For instance, a country with values 0.657 and 0.672 for GDP.per.capita and Healthy.life.expectancy, respectively, is explained by AA as 43% Utopia and 57% Dystopia. Conversely, with CLA, this country falls into the blue cluster with a membership degree of 79% and a centroid distance of 0.08. These results are paralleled by a country with values 0.268 and 0.242 for GDP.per.capita and Healthy.life.expectancy, which also has a centroid distance of 0.08 and a membership degree of 94% in CLA. Yet, AA explains this country as 13% Utopia and 87% Dystopia. Therefore, AA better distinguishes the difference in the profile's of these two countries.

3 Biarchetype Analysis

3.1 Definition

In biAA, biarchetypes are $\mathbf{Z}_{k\times c} = \beta_{k\times n}\mathbf{X}_{n\times m}\theta_{m\times c}$, where $\sum_{l=1}^{n}\beta_{gl} = 1$ with $\beta_{gl} \geq 0$ for $g = 1,\ldots,k$ and $\sum_{r=1}^{m}\theta_{rh} = 1$ with $\theta_{rh} \geq 0$ for $h = 1,\ldots,c$, i.e. the archetypes are mixture of the data points and variables.



(a) Fuzzy k-means clustering assignments.

(b) AA assignments by the maximum α.

Fig. 1. Plot of world happiness example. The crosses represent the prototypes.

There are k archetypes for rows and c for columns. The other restrictions are: $\sum_{g=1}^{k}\alpha_{ig} = 1$ with $\alpha_{ig} \geq 0$ for $i = 1,\ldots,n$ and $\sum_{h=1}^{c}\gamma_{hj} = 1$ with $\gamma_{hj} \geq 0$ for $j = 1,\ldots,m$. Therefore, the objective function to minimize subject to the previous constraints is:

$$RSS = \|\mathbf{X} - \alpha\mathbf{Z}\gamma\|^2 = \|\mathbf{X} - \alpha\beta\mathbf{X}\theta\gamma\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \sum_{g=1}^{k}\sum_{h=1}^{c}\alpha_{ig}z_{gh}\gamma_{hj}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \sum_{g=1}^{k}\sum_{h=1}^{c}\alpha_{ig}\left(\sum_{l=1}^{n}\sum_{r=1}^{m}\beta_{gl}x_{lr}\theta_{rh}\right)\gamma_{hj}\right)^2. \quad (2)$$

As before, the α coefficients determine how much each archetype contributes to the approximation of each observation, i.e. α_{ig} is the weight of the archetype g for the i-th observation. Analogously, the γ coefficients determine how much each archetype contributes to the approximation of each variable, i.e. γ_{hj} is the weight of the archetype h for the j-th variable. Biarchetypes are built as mixtures of observations and variables weighted by β and θ coefficients, respectively.

3.2 Relationship with clustering

Although the main objective of biarchetype analysis is to identify extreme values that define the dataset, it is worth noting that biAA can also be applied to clustering tasks, despite this not being its primary focus.

Fig. 2 displays a scheme showing the relationship between biAA and other unsupervised methods, where $\sum_{g=1}^{k}\alpha_{ig} = 1$ with $\alpha_{ig} \geq 0$ for $i = 1,\ldots,n$. biAA is to fuzzy biclustering as AA is to fuzzy clustering; and biAA is to AA as fuzzy biclustering is to fuzzy clustering. In simple words, in clustering methods, $\mathbf{Z} = (\alpha'\alpha)^{-1}\alpha'\mathbf{X}\gamma'(\gamma\gamma')^{-1}$ are centroids, but in archetypal methods, $\mathbf{Z} = \beta\mathbf{X}\theta$ are archetypes (extremes) ($\theta = \mathbf{I}_{m\times m}$ for AA); only observations are considered in AA and fuzzy clustering; therefore, $\gamma = \mathbf{I}_{m\times m}$, unlike biAA and fuzzy biclustering, where observations and variables are considered simultaneously.

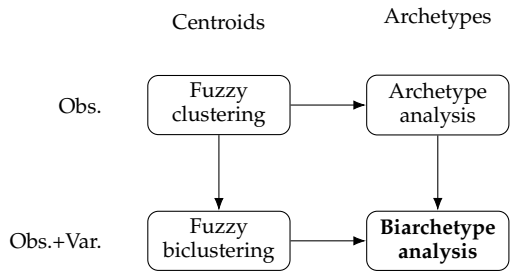


Fig. 2. Diagram of the relationship between biAA and other unsupervised methods.

3.2.1 Location of biarchetypes

In this section, we state some results that help in understanding the behavior of the row and column vectors of the biarchetype matrix $\mathbf{Z}_{k\times c}$.

Let $\mathbf{X}_{n\times m}$ be a data matrix with n observations and m continuous features. We denote by $\mathbf{x}_i^d$, $i = 1,\ldots,n$ the row vectors of the matrix $\mathbf{X}_{n\times m}$ (in this case observations) and by $\mathbf{x}_i^f$, $i = 1,\ldots,m$ the column vectors of the matrix $\mathbf{X}_{n\times m}$ (in this case features). This notation will be the same for all the matrices used.

The problem is to find a matrix $\mathbf{Z}_{k\times c}$ with $1 \le k \le n$, $1 \le c \le m$, which is expressed as $\mathbf{Z}_{k\times c} = \beta_{k\times n}\mathbf{X}_{n\times m}\theta_{m\times c}$ and the matrices $\alpha_{n\times k}$, $\mathbf{Z}_{k\times c}$ and $\gamma_{c\times m}$ minimize

$$RSS = \left\|\mathbf{X}_{n\times m} - \alpha_{n\times k}\mathbf{Z}_{k\times c}\gamma_{c\times m}\right\|^2.$$

Now we distinguish several cases depending on the values of k and c.

Case I: $k = 1$ **and** $c = 1$. In this case, $\alpha_{n\times 1} = (1,\ldots,1)'$ and $\gamma_{1\times m} = (1,\ldots,1)$; then, the real value $\mathbf{Z}_{1\times 1}$ that minimizes RSS is the mean of all entries in matrix $\mathbf{X}_{n\times m}$, that is,

$$\mathbf{Z}_{1\times 1} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}.$$

Case II: $k = n$ **and** $c = m$. In this case we consider $\alpha_{n\times n} = \mathbf{I}_{n\times n}$ and $\gamma_{m\times m} = \mathbf{I}_{m\times m}$. Then, by choosing $\mathbf{Z}_{n\times m} = \mathbf{X}_{n\times m}$ we obtain RSS = 0.

Case III: $1 \le k < n$ **and** $c = m$. We consider $\gamma_{m\times m} = \mathbf{I}_{m\times m}$ and $\theta_{m\times m} = \mathbf{I}_{m\times m}$, then, the problem now consists of minimizing

$$RSS = \left\|\mathbf{X}_{n\times m} - \alpha_{n\times k}\mathbf{Z}_{k\times m}\right\|^2,$$

which is a typical problem in AA and the location of the archetypes is explained in [3] and reviewed in Sec. 2.2.

Case IV: $k = n$ **and** $1 \le c < m$. This is the case of finding only the archetypes of features, i.e. we consider $\alpha = \mathbf{I}_{n\times n}$ and $\beta = \mathbf{I}_{n\times n}$. As the Frobenius norm of a matrix is the same as the Frobenius norm of its transpose, $\left\|\mathbf{X} - \mathbf{Z}\gamma\right\|^2 = \left\|\mathbf{X}' - \gamma'\mathbf{Z}'\right\|^2$, features can adopt the role of observations when $\mathbf{X}$ is transposed. Then, the same reasoning as in the preceding case entails a problem of AA.

Case V: $k = 1$ **and** $1 < c < m$. In this case $\alpha = \mathbf{I}_{n\times n}$ and the problem of minimizing

$$RSS = \left\|\mathbf{X}_{n\times m} - (1,\ldots,1)'\left(\mathbf{Z}_{1\times c}\gamma_{c\times m}\right)\right\|^2,$$

is satisfied if the mean value $\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^d = (\bar{x}_1^d,\ldots,\bar{x}_m^d) = (\mathbf{Z}_{1\times c}\gamma_{c\times m})$ where $\bar{x}_j^d = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. Note that each real number $\bar{x}_i^d$ belongs to the convex hull of the components of the vector $\mathbf{Z}_{1\times c}$.

Case VI: $c = 1$ **and** $1 < k < n$. In this case RSS is minimized if $\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^f = (\bar{x}_1^f,\ldots,\bar{x}_n^f) = \left(\mathbf{Z}'_{k\times 1}\alpha'_{n\times k}\right)$ where $\bar{x}_i^f = \frac{1}{m}\sum_{j=1}^{m} x_{ij}$. Note that each real number $\bar{x}_i^f$ belongs to the convex hull of the components of the vector $\mathbf{Z}_{k\times 1}$.

Case VII: $1 < k < n$ **and** $1 < c < m$.

Let us call $\mathbf{V}_{n\times c} = \mathbf{X}_{n\times m}\theta_{m\times c}$; then, each $\mathbf{v}_j^f$ $(j = 1,\ldots,c)$ belongs to the convex hull C_X^f of the data $\mathbf{x}_i^f$ $(i = 1,\ldots,m)$. Moreover, since $\mathbf{Z}_{k\times c} = \beta_{k\times n}\mathbf{V}_{n\times c}$, each vector $\mathbf{z}_j^d$ $(j = 1,\ldots,k)$ belongs to the convex hull C_V^d of the vectors $\mathbf{v}_i^d$ $(i = 1,\ldots,n)$.

Proposition 1. Having fixed the matrix $\theta_{m\times c}$, there is a matrix of biarchetypes $\mathbf{Z}_{k\times c}$ such that each row vector $\mathbf{z}_j^d$ $(j = 1,\ldots,k)$ belongs to the boundary of the convex hull C_V^d.

Now, let us call $\mathbf{Y}_{k\times m} = \beta_{k\times n}\mathbf{X}_{n\times m}$; then, each $\mathbf{y}_j^d$ $(j = 1,\ldots,k)$ belongs to the convex hull C_X^d of the data $\mathbf{x}_i^d$ $(i = 1,\ldots,n)$. Moreover, since $\mathbf{Z}_{k\times c} = \mathbf{Y}_{k\times m}\theta_{m\times c}$, each vector $\mathbf{z}_j^f$ $(j = 1,\ldots,c)$ belongs to the convex hull C_Y^f of the vectors $\mathbf{y}_i^f$ $(i = 1,\ldots,m)$.

Proposition 2. Having fixed the matrix $\beta_{k\times n}$, there is a matrix of biarchetypes $\mathbf{Z}_{k\times c}$ such that each column vector $\mathbf{z}_j^f$ $(j = 1,\ldots,c)$ belongs to the boundary of the convex hull C_Y^f.

Proof of Propositions 1 and 2 is detailed in Appendix A.

Example 1. This toy example illustrates the location of the biarchetypes for different values of k and c, for the following matrix $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \\ 16 & 17 & 18 & 19 & 20 \\ 21 & 22 & 23 & 24 & 25 \end{pmatrix}$.

For k = 1 and c = 1, $\mathbf{z}$ = 13 (mean of all entries of the matrix according to Case I). For k = 1 and c =2, $\mathbf{z}$ = $\begin{pmatrix} 11 & 15 \end{pmatrix}$; here $(\bar{x}_1^d,\ldots,\bar{x}_5^d) = (11,12,13,14,15)$ and each real number $\bar{x}_i^d$ belongs to the convex hull of the components of $\mathbf{z}$ according to Case V. For k = 2 and c = 1, $\mathbf{z}$ = $\begin{pmatrix} 3 \\ 23 \end{pmatrix}$ and, according to Case VI, each real number $\bar{x}_i^f$ belongs to the convex hull of the components of the vector $\mathbf{z}$. For k = 2 and c = 2, $\mathbf{z}$ = $\begin{pmatrix} 1 & 5 \\ 21 & 25 \end{pmatrix}$. For this last case, RSS = 0, and, according to Case VII, the vectors $\mathbf{z}_j^d$ and $\mathbf{z}_j^f$, $j = 1, 2$ are located at the boundary of convex sets.

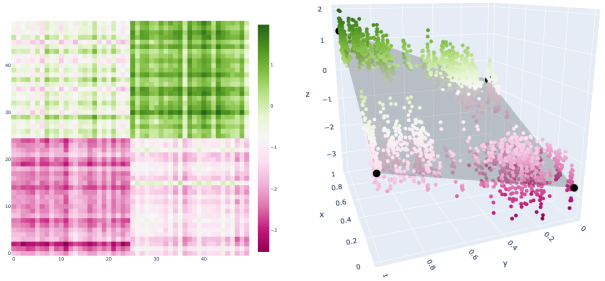
Example 2. In this example, we will generate the data from a multivariate random distribution.

The covariance matrix for both rows and columns is:

$$\Sigma = \begin{pmatrix} 1 & 0.8 & \dots & 0.8 & 0.8 & 0 & 0 & \dots & 0 & 0 \\ 0.8 & 1 & \dots & 0.8 & 0.8 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0.8 & 0.8 & \dots & 1 & 0.8 & 0 & 0 & \cdots & 0 & 0 \\ 0.8 & 0.8 & \dots & 0.8 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 0.8 & \cdots & 0.8 & 0.8 \\ 0 & 0 & \dots & 0 & 0 & 0.8 & 1 & \cdots & 0.8 & 0.8 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0.8 & 0.8 & \cdots & 1 & 0.8 \\ 0 & 0 & \dots & 0 & 0 & 0.8 & 0.8 & \cdots & 0.8 & 1 \end{pmatrix}$$

And the mean also for both rows and columns is:

$$\mu = (0 \cdots 0)$$



(a) Simulated data as a heatmap.

(b) Simulated data in the biarchetype space.

Fig. 3. Representations of data and the biarchetypal space for example 2, i.e. representation of the coefficients α and γ.

The data generated can be seen in Fig. 3a. In Fig. 3b, the data are represented using the coefficients α and γ of biAA with k = 2 and c = 2. These coefficients are mapped to the x and y axes of the figure and the z axis represents the value of each observation. The biarchetypes are represented in black and, as seen, they are at the extremes of the data.

3.2.2 Selecting the number of biarchetypes

As in the case of AA, if there is no information available a priori, we can use the elbow criterion, but in this case, we look for the elbow of a surface instead of a curve. In biAA, we run biAA for different values of k and c and display their RSS values in a 3D plot. We select the point (k,c) where the surface "flattens", i.e. (k, c) is the point at which the RSS of the following points $(k + 1, c)$, $(k, c + 1)$, and $(k + 1, c + 1)$ stops decreasing drastically with respect to the RSS of the point (k, c).

3.3 Algorithm

The following iterative method is proposed to solve biAA. It is based on alternating minimization as in the AA algorithm by [3].

1) Data preparation: To randomly initialize the matrices α, γ, β and θ, fulfilling the constraints in Sect. 3.1.
2) Repeat until RSS is sufficiently small or the number of maximum iterations is reached:
   a) Find the best α (fixed γ): solve n convex least squares problems using $\mathbf{X}' = (\mathbf{Z}\gamma)'\alpha'$.
   b) Find the best γ (fixed α): solve m convex least squares problems using $\mathbf{X} = (\alpha\mathbf{Z})\gamma$.
   c) Recalculate the biarchetypes, where $^+$ stands for the Moore-Penrose pseudoinverse: $\mathbf{Z} = \alpha^+\mathbf{X}\gamma^+$.
   d) Find the best β (fixed θ): solve k convex least squares problems using $\mathbf{Z}' = (\mathbf{X}\theta)'\beta'$.
   e) Find the best θ (fixed β): solve c convex least squares problems using $\mathbf{Z} = (\beta\mathbf{X})\theta$.
   f) Recalculate the biarchetypes: $\mathbf{Z} = \beta\mathbf{X}\theta$.
   g) Calculate the new RSS.

Note that the convex least squares problems can be solved as proposed by [3], i.e. using a penalized least squares problem [67]. The idea is, given a least squares problem $\mathbf{A}_{n\times k}\mathbf{X}_{k\times m} = \mathbf{B}_{n\times m}$, to add a row with constant elements C to $\mathbf{A}$ and $\mathbf{B}$, in order to obtain a new problem $\mathbf{A}_{(n+1)\times k}\mathbf{X}_{k\times m} = \mathbf{B}_{(n+1)\times m}$, in such a way that RSS would be:

$$\begin{aligned} RSS = \sum_{i=1}^{n+1}\sum_{j=1}^{m}\left(b_{ij} - \sum_{h=1}^{k} a_{ih}x_{hj}\right)^2 = \\ \sum_{j=1}^{m}\left(\sum_{i=1}^{n}\left(b_{ij} - \sum_{h=1}^{k} a_{ih}x_{hj}\right)^2 + \left(b_{n+1,j} - \sum_{h=1}^{k} a_{n+1,h}x_{hj}\right)^2\right) = \\ \sum_{j=1}^{m}\left(\sum_{i=1}^{n}\left(b_{ij} - \sum_{h=1}^{k} a_{ih}x_{hj}\right)^2 + \sum_{j=1}^{m}\left(C - \sum_{h=1}^{k} Cx_{hj}\right)^2\right) = \\ \sum_{j=1}^{m}\left(\sum_{i=1}^{n}\left(b_{ij} - \sum_{h=1}^{k} a_{ih}x_{hj}\right)^2 + \sum_{j=1}^{m} C^2\left(1 - \sum_{h=1}^{k} x_{hj}\right)^2\right). \end{aligned} \tag{3}$$

Therefore, if value C is high, the term $C^2\left(1 - \sum_{h=1}^{k} x_{hj}\right)^2$ forces the convexity of the elements of $\mathbf{X}$ in eq. 3.

As regards computational complexity, the biAA algorithm can be considered as complex as computing the AA algorithm twice. Like AA, the speed of biAA depends on the efficiency of the convex least squares method. The computational complexity for the AA algorithm was analyzed by [66], based on this analysis, the computation time increases linearly as the number of observations increases, while it remains approximately constant as the number of archetypes increases. In practical terms, the biAA algorithm is a computer-intensive algorithm and its convergence speed depends on the data structure, so if convergence is not attained in a few steps for specific numbers k and c, those numbers probably do not explain the data well.

3.4 Illustrative example and comparison with biclustering

The following example illustrates the use of biAA and its advantages in comparison with biclustering, especially when working with non-clustered data. We consider the data of 45 students from Universitat Jaume I, who reported the number of hours per week spent working on a subject at home over 17 weeks. The complete description of the data can be found in [68]. Missing data are imputed by *mice* [69].

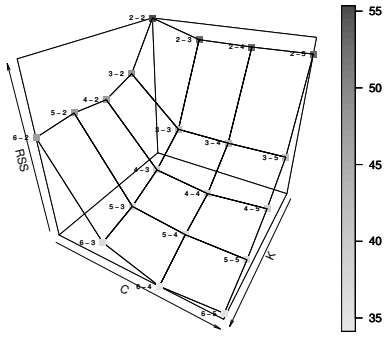


Fig. 4. RSS for k from 2 to 6 and c from 2 to 5.

TABLE 1
The γ coefficients for biAA, BMM, and FDkMpf of the illustrative example.

1	0	0	1	0	0	1	0	0
0.71	0.12	0.17	1	0	0	1	0	0
0.57	0.42	0.01	1	0	0	0.85	0.08	0.08
0	0.80	0.20	0	0.01	0.99	0	0.5	0.5
0.13	0.76	0.11	0	1	0	0	0.5	0.5
0.14	0.86	0	0	1	0	0.29	0.35	0.35
0	0.80	0.20	0	1	0	0	0.5	0.5
0.01	0.99	0	0	0	1	0.2	0.4	0.4
0	0.90	0.11	0	1	0	0	0.5	0.5
0	0.89	0.11	0	1	0	0	0.5	0.5
0	0.42	0.58	0	1	0	0	0.5	0.5
0.05	0.45	0.5	0	0.96	0.04	0	0.5	0.5
0	0.62	0.38	0	1	0	0	0.5	0.5
0.05	0	0.95	0	0	1	0	0.5	0.5
0	0.04	0.96	0	0	1	0	0.5	0.5
0	0.10	0.91	0	0	1	0	0.5	0.5
0.13	0	0.87	0	0	1	0	0.5	0.5

The data range from 0 to 10, with mean 4.94 and standard deviation 2.46. We apply biAA with k = 4 and c = 3, since the elbow is found at those values (see Fig. 4).

The γ coefficients for biAA, BMM, and FDkMpf are shown in Table 1 (we sort them in order to make the comments easier). FDkM was also applied, but the solution is not valid since the same prototypes are obtained for all the groups.

(From now on, we will use archetype or archetypal instead of biarchetype or biarchetypal to simplify the language).

The feature similar to the first archetypal variable corresponds to the first week. The second and third weeks are also similar, but with a temporal gradation (0.71 and 0.57). Week 8, an intermediate week of the semester, is similar to the second archetypal variable. Other intermediate weeks (4, 5, 6, 7, 9, and 10) are also similar. Week 15, a week at the end of the semester, is similar to the third archetypal variable, as well as weeks 14, 16, and 17. Weeks 11, 12, and 13 are explained as mixtures (nearly 50% -50%) of the second and third archetypal variables. Note that the third week was also explained as a mixture close to 50% -50% of the first and second archetypal variables. In summary, the archetypal variables correspond to the profile of the beginning, middle and end of the semester, respectively. The weeks in the transitions between these temporal points are reflected as mixtures.

As regards the prototypical variables for biclustering methodologies, the first three weeks have probabilities of 1 (or nearly 1 for FDkMpf) for the first prototypical variable of BMM. Unlike biAA where gradation was found, the probabilities (memberships) are nearly crisp classifications for BMM, not only for the first prototypical variable, but for the rest as well. The intermediate weeks 5, 6, 7, 9, 10, 11, 12, and 13 have probabilities of 1 for the second prototypical variable, while the final weeks (14, 15, 16, and 17), and the intermediate weeks 4 and 8 have probabilities of nearly 1 or 1 for the third prototypical variable.

Note the difference with biAA. On the one hand, in BMM there is a lack of gradation over time in the memberships (no mixture is found, but the memberships are extremely high, nearly all ones), as if changes between adjoining weeks were radical (as breaking jumps) rather than smooth. Therefore, the information provided by biAA is richer. On the other hand, there are two intermediate weeks (4 and 8) belonging to the third prototypical variable corresponding to the end of semester weeks, which is not very coherent. Therefore, the information provided by biAA is more reasonable. Finally, the second and third prototypical variables are identical for FDkMpf, with 50%-50% or close degrees of membership for weeks 4 to 17. Therefore, the information returned by FDkMpf is poorer than that of BMM and biAA.

Table 2 displays the representative points $\mathbf{Z}$, archetypes or centroids for biAA and biclustering, respectively (we sort them in order to make the comments easier). The first archetype describes a student who works very few hours per week throughout the semester. The second archetype represents a student who studies very few hours throughout the semester, except at the end of the semester, when they work for 9 hours per week. The third archetype describes a student who works very few hours at the beginning of the semester (1h per week), many hours during the semester (10h per week), and intermediate hours (4h per week) at the end of the semester. The fourth archetype represents a student who studies many hours throughout the whole semester.

For BMM, the prototypes are not as pure as the archetypes. For example, there is no great difference between centroids 2 and 3: centroid 3 studies only one or one and a half hours more than centroid 2 per week. The centroids are not as intuitively interpretable as archetypes. Centroid 1 corresponds to a student who studies 2 or 3 hours throughout the semester; centroid 2 studies 2h per week at the beginning and 4 or 5h per week for the rest of the semester; centroid 3 works 4 hours at the beginning and 6 hours per week for the rest of the semester; while centroid 4 studies 5h per week at the beginning of the semester and 7 or 8 hours throughout the rest of the semester. It seems that centroids are limited to following a gradation according to the total number of hours studied throughout the semester rather than by differences in behavior throughout the semester. For FDkMpf, the comments are similar, but in addition there is no difference between the intermediate and final weeks. For example, the profiles of students 32 and 33, who are similar to archetype 2 (with αs of 0.84 and

TABLE 2
The archetypes and centroids for biAA, BMM, and FDkMpf of the toy example.

1.54	2.36	0.36	1.67	2.07	3.49	2.03	3.28	3.28
1	2	9	2.32	4.15	5.02	2.74	4.89	4.89
1	10	4	3.79	5.70	6.05	3.42	5.97	5.97
8	6.32	9.36	5.01	7.69	6.60	5.55	6.97	6.97

0.88, respectively), would not be reflected by the centroids of BMM or FDkMpf. They belong to cluster 2 of BMM, with probabilities of 0.97 and 1, respectively. But this does not say anything about how far (or in which direction) from centroid 2 those students are. This happens because the goal of clustering is to assign the data to groups, not to explain the structure of the data more qualitatively.

3.5 Ablation study

Another important aspect to consider is the ablation study. In this analysis, we study the implications of calculating archetypes separately (solely by rows and solely by columns) and then combining them, as opposed to computing the archetypes simultaneously.

In particular, the biarchetypes in biAA are reconstructed from the βs and θs obtained by applying $\mathbf{X} \simeq \alpha\beta\mathbf{X}\theta\gamma$ to the dataset ($arch_{biaa} = \beta\mathbf{X}\theta$), whereas those in the ensemble are derived from applying AA across the rows of $\mathbf{X}$, (i.e. $\mathbf{X} \simeq \alpha_r\beta_r\mathbf{X}$) and also applying it across the columns (i.e. $X \simeq \mathbf{X}\beta_c\alpha_c$). In the latter case, the biarchetypes are obtained by combining the βs from both methods ($arch_{ensemble} = \beta_r\mathbf{X}\beta_c$).

To conduct this study, we initiated an experimental procedure. We generated a synthetic dataset 50 times whose shape is 100×100 and with 3×3 biarchetypes. The values for these datasets were constrained to fall within the range of 0 to 1, ensuring a standardized scale for comparison.

Specifically, each dataset is constructed to conform a mixture of biarchetypes, where the $\mathbf{Z}$ matrix encapsulates the biarchetype values. Moreover, the entries of matrix α's rows and matrix γ's columns are determined by sampling from the $U(0, \phi)$ distribution, with ϕ being a parameter within the interval [0, 1]. After that, the procedure assigns a 1 to the entry that denotes the group to which the observation belongs, thereby ensuring that this maximum coefficient signifies the group assignment. Subsequently, this vector is normalized to achieve a unit norm, thus ensuring its convexity.

The parameter ϕ significantly influences the membership characteristics within the archetypal model. At a setting of $\phi = 0$, each observation is completely represented by one archetype, exhibiting pure membership. As ϕ approaches 1, the model shifts towards a mixed-membership framework, allowing observations to have more evenly distributed memberships across the different archetypes. In this instance, the parameter ϕ has been determined to be 0.05.

The histograms, as depicted in Figure 5, reveals a notable distinction in the distribution of values obtained using our methodology compared to the separate calculation of archetypes. Specifically, the values derived from the biAA approach manages to recover the true archetypes, which does not happen with the ensemble archetype approach.

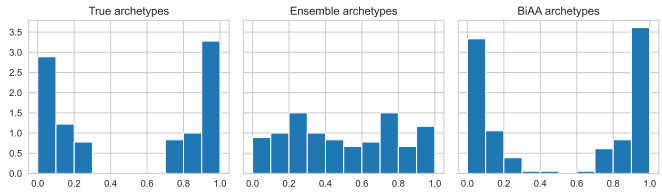


Fig. 5. Distribution of the archetype values.

4 Results and Discussion

Like biclustering, biAA can be applied to a wide range of fields. In this section, we will apply it to biology, document analysis and community detection.

The code and data sets for reproducing the results including those in Sec. 3.4 are available at https://github.com/aleixalcacer/JA-BIAA.

4.1 Gene expression data

To show how biAA can be applied, we examine data from gene expression of cutaneous melanoma used in [70], [71], [72]. Instead of meticulously re-analyzing this data set, we use it to highlight the salient features of biAA.

The aim of this study was to test the idea that molecular profiles generated by cDNA microarrays could be used to differentiate between several subtypes of cutaneous melanoma, a kind of skin cancer. mRNA was collected from the 31 cutaneous melanoma samples, and Cy5-labeled cDNA was created. All samples were examined with the same reference probe, identified as Cy3. For each sample, Cy5 and Cy3-labeled cDNA were combined and hybridized to a different melanoma microarray. Red and green lasers were used to scan the hybridization array, and the resulting image was then analyzed.

The same pre-processing was carried out as in [70], [71]. Only 3613 cDNAs of the 8150 observations were classified as well measured. Cy5/Cy3 expression ratios were computed for the accurately measured genes. Ratios that were more than or equal to 50 and less than or equal to 0.02 were reduced to 50 and 0.02, respectively. A logarithmic scale was applied to the derived ratios (base 2). The log ratios were adjusted so that the median log-ratio for each experiment was equal to zero by subtracting the median log-ratio within an experiment from all log-ratios for that experiment. Since a single reference probe was utilized in all experiments, there was no standardization between trials.

Using one minus the Pearson correlation coefficient of log-ratios as a measure of dissimilarity between two experiments, [70] applied the average linkage hierarchical clustering on the 31 cutaneous melanoma samples and obtained two clusters of 12 and 19 samples.

Regarding [71], they used double k-means to analyze the same data set. In this case, the columns were centered and scaled to unit variance, finding that the separation between two columns is proportional to one minus the Pearson correlation coefficient. In particular, their analysis indicates that samples 4 and 7 are members of the '19-samples' group obtained by [70], i.e. the main cluster group. The membership of these two samples in [71] differs from that obtained by [70].

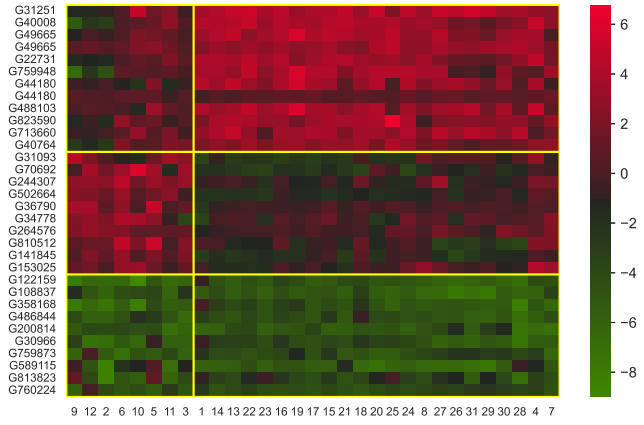
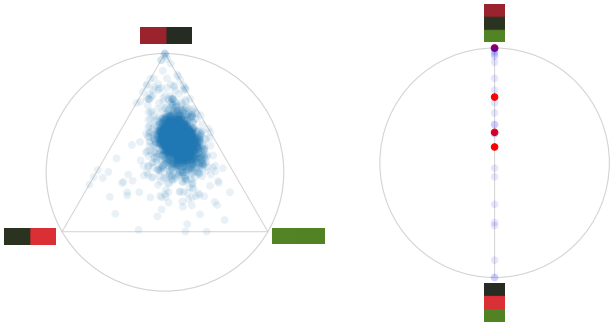


Fig. 6. Representation of the most similar observations to each archetype. The color represents the expression ratio of each gene for each sample.



(a) Representation of the observations (α matrix) in their archetypal space. Each corner represents an archetype.

(b) Representation of the variables (γ matrix) in their archetypal space. Each corner represents an archetype. Samples 1, 8, 4 and 7 are colored in red.

Fig. 7. Representations of the archetypal space for Gene expression data.

In our case, we applied biarchetype analysis to the same data set as [71], extracting three archetypes for the genes (rows) and two archetypes for the melanoma samples (columns).

As can be seen in Fig. 6, regarding the melanoma samples, if we cluster the data using the location of the maximum archetypal coefficient γ (i.e. the archetype that is most similar to the sample), we obtain two clusters of 8 and 23 samples. Regarding the archetypes of the genes, the first two discriminate the two groups quite well, while the third archetype (or group of genes) is not expressed for any melanoma group.

If we compare our results to those obtained by [70], four samples are classified in different clusters. Samples 1, 4, 7 and 8 belong to the main cluster group with biAA (see Fig. 7), unlike results provided by [70]. If we compare our results to those obtained by [71], two samples are classified in different clusters, samples 1 and 8. Classification group of samples 4 and 7 is shared with biAA and results by [71]. According to the γ coefficients of biAA, samples 4 and 7 are a nearly equal mixture between both archetypes, with the values of the coefficient being 0.4 and 0.6 corresponding to the first and second archetype for sample 4, and 0.45 and 0.55 corresponding to the first and second archetype for sample 7. Therefore, samples 4 and 7 could be in the border between both groups, which could explain the difference in classification by different methods. However, the γ coefficients for sample 8 and 1 are 0.75 (0.25) and 1 (0) for the second (first) archetype, respectively. In other words, sample 1 (and to lesser extent sample 8) should definitely be in the main group according with biAA.

4.2 Text documents

Another common use for biclustering is for clustering documents and words. In this case, we have applied biAA to a subset of the 20 Newsgroups collection, set up by [73]. Specifically, we have analyzed three topics: *rec.autos*, *rec.sport.hockey* and *talk.politics.guns*.

Additionally, although our algorithm is designed to identify extreme prototypes rather than clusters, due to the absence of comparable benchmarks, we have conducted comparisons with the following popular biclustering algorithms, which are also considered as the baseline elsewhere [74]: Louvain Clustering [75], Spectral Co-clustering [76] and Spectral Biclustering [77]. For the three methods we have left all the default values and, in those that allow it, we have determined the number of clusters to search.

Specific, for each document, the number of times each word is repeated in the document has been stored in a count matrix, where each row represents a document, each column a word, and the values indicate how many times each word is repeated in each document.

In addition, a Tfidf transformer [73] was used to convert a count matrix into a normalized tf or tf-idf representation. Tf stands for term frequency, and tf-idf stands for term frequency multiplied by inverse document frequency. This is a standard term weighting method used in information retrieval, and it is also effective for classifying documents.

We have applied biarchetype analysis to this normalized matrix, obtaining three archetypal profiles for documents and three archetypal profiles for words.

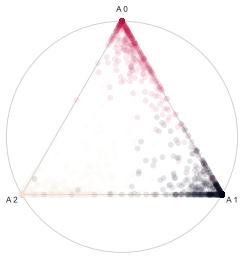


Fig. 8. Representation of the documents in their archetypal space (α values). The color represents the category of the document.

In Fig. 8, it is clear that the three archetypes discriminate the three groups of documents perfectly.

Regarding the words, in Fig. 9 the words are represented as a graph. The weight of each edge represents the similarity of the words in terms of the archetypes (the gamma coefficients). The graph weights (i.e. the gamma coefficients) split

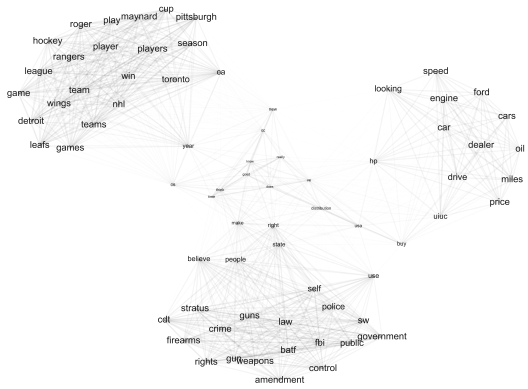
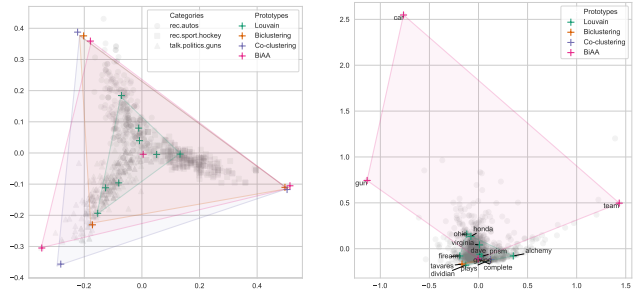


Fig. 9. Representation of the most similar words to each archetype (filtered using a threshold over the coefficient matrix). The weight of each edge is the cosine similarity between the two words in their archetypal space. This plot was created using the *networkx* Python package.

the words into three groups, where the words within each group are related to one of the selected topics.



(a) PCA and prototypes obtained in the *documents* part.

(b) PCA and prototypes obtained in the *words* part.

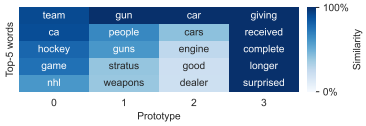
Fig. 10. First 2 components of PCA along with the prototypes of the dataset discovered by multiple clustering methods. The colored areas represent the convex hulls of the prototypes for each method.

community detection

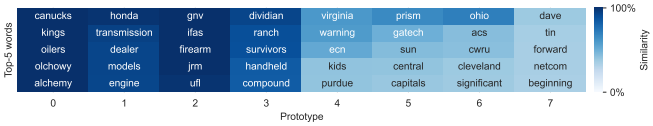
TABLE 3
RSS for text documents and community detection example.

Method	RSS (text)	RSS (community)	Fuzzy
biAA	1605.58	1064.4	True
Louvain	1620.85	1268.24	True
Biclustering	1664.42	1452.03	False
Co-clustering	1666.97	1663.77	False

Upon examining the results of biAA, Figure 10 displays the prototypes identified by each method. To facilitate their representation, Principal Component Analysis (PCA) was applied to the data for both rows and columns (the transposed matrix), and the first two components were plotted. It is observable that when analyzing the matrix by rows, which in this example correspond to documents, both biAA and Spectral Clustering methods yield quite extreme prototypes. However, when the dataset is analyzed from the perspective of words, only biAA identifies extreme prototypes. Specifically, it identifies the words *car*, *team* and *gun* which correspond to highly archetypal words for groups *rec.autos*, *rec.sport.hockey* and *talk.politics.guns* respectively.



(a) Biarchetype Analysis.



(b) Louvain method.

Fig. 11. Top-5 most similar words to each prototype for different algorithms.

Finally, building on the previous observations, Figure 11 includes the five most similar words to each prototype for methods that allow for mixed membership, i.e. biAA and the Louvain method. It is evident that the biAA identifies words that are more archetypical compared to those identified by the Louvain method. Specifically, the first three prototypes discovered by biAA can be clearly associated with the three groups of documents present in the dataset.

Table 3 compiles RSS for all the methods. biAA provides the lowest RSS.

4.3 Community detection

Finally, we have also applied biAA to detect communities within the company Enron. For that, we have studied the data set described in [78], which contains a collection of emails between the company's employees.

We have created an adjacency matrix between employees, containing 1 if one employee has emailed another or 0 if the first one has never sent an email to the second one.

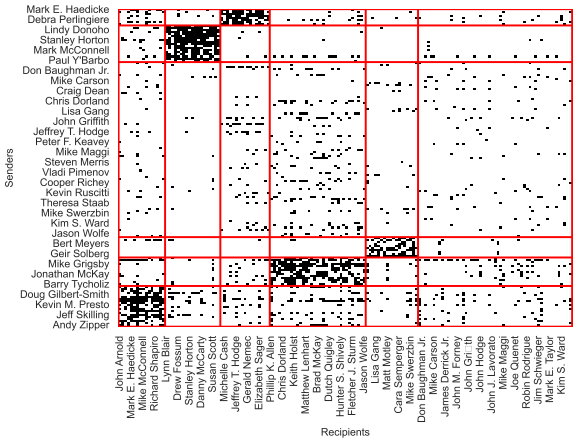


Fig. 12. The adjacency matrix ordered according to the archetypes obtained with biAA.

After applying biAA with $k = 6$ and $c = 6$ to this adjacency matrix, we obtained the results in Fig. 12. In the 'senders' part, the group Z_3 could be omitted (represents employees who haven't sent emails to a specific group). The same occurs with the group Z_6 in the 'recipients' part of the adjacency matrix.

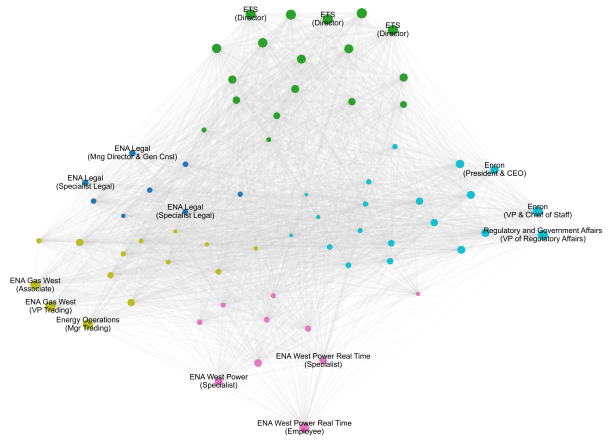
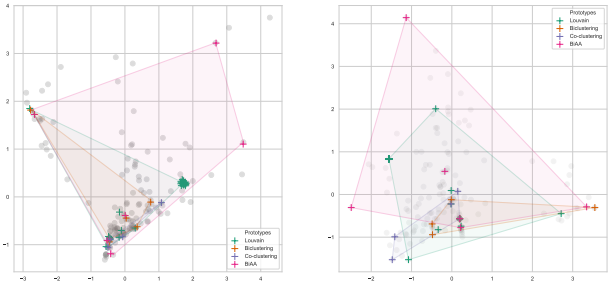


Fig. 13. Representation of employees from the point of view of who they send emails to. The weight of each node is computed as in Fig. 9. The size of each employee is proportional to how similar it is to its closest archetype and the color of each one is determined by the closest archetype.

If we analyze the employees from the point of view of who they send emails to, we obtain the results shown in Fig. 13, in which we have removed Z_3-like employees as they do not exchange emails with a specific group. As can be seen, the dark blue cluster represents the Legal department, the green one, the ETS department; the olive green one, the Gas/Energy department; the pink one, the West Power departments; and the blue one, Enron's top management.



(a) PCA and prototypes obtained in the *senders* part.

(b) PCA and prototypes obtained in the *recipients* part.

Fig. 14. First 2 components of PCA along with the prototypes of the dataset discovered by multiple clustering methods. The colored areas represent the convex hulls generated by the prototypes of each method.

Here, in Figure 14 the same procedure as in the previous problem has been applied. It is clear that in both cases, for both the senders and the recipients of the emails, the convex hull of the prototypes identified by biAA covers the largest area. This could serve as a measure of how extreme (or distant from each other) the prototypes are. Therefore, based on this metric, it is evident that biAA reveals the most extreme prototypes.

As before, biAA also provides the lowest RSS for this example.

5 Conclusion

In this work, we have proposed a new unsupervised machine learning technique: biarchetype analysis. We have compared the results of biAA and biclustering in an illustrative example, showing not only the greater interpretability provided by biAA, but also the greater coherence of the results. We have also seen its usefulness in several problems in different fields, where more distinct aspects are extracted with biAA than using several biclustering methods.

biAA has been defined for continuous data. In future work, it could be extended to other kinds of data, such as functional data, to which AA was also extended [51]. Note that biclustering analysis of time series is used in many fields such as neuroscience [79] and engineering [80]; therefore, biAA could also be used for the same problems. Biarchetypoid analysis could also be introduced in the same way that archetypoid analysis was defined [50], where biarchetypes are not determined by mixtures of observations and features, but by concrete elements of the data set. Just as archetype analysis is sensitive to outliers, biAA is too. Robust biAA could be defined in the same way as robust AA was [48]. Likewise, biAA for missing data could be defined as it was for AA [46], and it could be used in recommender systems to find profiles of users and products, for instance. Another line of future work would be to apply biAA to different fields where biclustering analysis is applied, and to study more computational methods to calculate biAA, especially for big data. Furthermore, biAA could also be easily extended to high dimensions in a similar way to the decomposition proposed in [81]. Finally, non-linear biAA could be proposed by using deep learning, based on the works on deep AA by [58] and [37].

Appendix A
Proof of Propositions 1 and 2

See Supplementary material.

Acknowledgments

The authors would like to thank Francesca Martella for providing them with gene expression data.

This research was partially supported by the Spanish Ministry of Universities (FPU grant FPU20/01825), Spanish Ministry of Science and Innovation (PID2022-141699NB-I00, PID2020-118763GA-I00 and PID2020-115930GA-I00) and UJI-B2020-22 and TRANSUJI/2023/6 from Universitat Jaume I, Spain.

References

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

[2] S. M. Keller, M. Samarin, M. Wieser, and V. Roth, "Deep archetypal analysis," in *Pattern Recognition*, G. A. Fink, S. Frintrop, and X. Jiang, Eds. Cham: Springer International Publishing, 2019, pp. 171–185.

[3] A. Cutler and L. Breiman, "Archetypal Analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[4] C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage, "Descriptive matrix factorization for sustainability: Adopting the principle of opposites," *Data Mining and Knowledge Discovery*, vol. 24, no. 2, pp. 325–354, 2012.

[5] T. Davis and B. Love, "Memory for category information is idealized through contrast with competing options," *Psychological Science*, vol. 21, no. 2, pp. 234–242, 2010.

[6] I. Cabero and I. Epifanio, "Finding archetypal patterns for binary questionnaires," *SORT*, vol. 44, no. 1, pp. 39–66, 2020.

[7] J. A. Hartigan, "Direct clustering of a data matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, 1972.

[8] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, vol. 8, no. 2000, 2000, pp. 93–103.

[9] M. B. Ferraro, P. Giordani, and M. Vichi, "A class of two-mode clustering algorithms in a fuzzy setting," *Econometrics and Statistics*, vol. 18, pp. 63–78, 2021.

[10] H. Zhao, A. Wee-Chung Liew, D. Z Wang, and H. Yan, "Biclustering analysis for pattern discovery: current techniques, comparative studies and applications," *Current Bioinformatics*, vol. 7, no. 1, pp. 43–55, 2012.

[11] G. Kerr, H. J. Ruskin, M. Crane, and P. Doolan, "Techniques for clustering gene expression data," *Computers in Biology and Medicine*, vol. 38, no. 3, pp. 283–293, 2008.

[12] J. Xie, A. Ma, A. Fennell, Q. Ma, and J. Zhao, "It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data," *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1450–1465, 2019.

[13] S. Dolnicar, S. Kaiser, K. Lazarevski, and F. Leisch, "Biclustering: Overcoming data dimensionality problems in market segmentation," *Journal of Travel Research*, vol. 51, no. 1, pp. 41–49, 2012.

[14] I. Van Mechelen, H.-H. Bock, and P. De Boeck, "Two-mode clustering methods: a structured overview," *Statistical Methods in Medical Research*, vol. 13, no. 5, pp. 363–394, 2004.

[15] R. Forsati, H. M. Doustdar, M. Shamsfard, A. Keikha, and M. R. Meybodi, "A fuzzy co-clustering approach for hybrid recommender systems," *International Journal of Hybrid Intelligent Systems*, vol. 10, no. 2, pp. 71–81, 2013.

[16] S. Kaiser, "Biclustering: methods, software and application," Ph.D. dissertation, Ludwig-Maximilians-Universität München, 2011.

[17] Z. Shkedy, R. Sengupta, and N. J. Perualila, "Identification of local patterns in the NBA performance indicators," in *Applied Biclustering Methods for Big and High-Dimensional Data Using R*. Chapman and Hall/CRC, 2016, pp. 323–344.

[18] V. A. Koutsonikola and A. Vakali, "A fuzzy bi-clustering approach to correlate web users and pages." *IJ Knowledge and Web Intelligence*, vol. 1, no. 1/2, pp. 3–23, 2009.

[19] R. Henriques, C. Antunes, and S. C. Madeira, "A structured view on pattern mining-based biclustering," *Pattern Recognition*, vol. 48, no. 12, pp. 3941–3958, 2015.

[20] M. Mørup and L. K. Hansen, "Archetypal analysis for machine learning and data mining," *Neurocomputing*, vol. 80, pp. 54–63, 2012.

[21] Y. Chen, J. Mairal, and Z. Harchaoui, "Fast and Robust Archetypal Analysis for Representation Learning," in *CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition*, 2014, pp. 1478–1485.

[22] C. Bauckhage, K. Kersting, F. Hoppe, and C. Thurau, "Archetypal analysis as an autoencoder," in *Workshop New Challenges in Neural Computation*, 2015, pp. 8–15.

[23] S. Mair, A. Boubekki, and U. Brefeld, "Frame-based data factorizations," in *International Conference on Machine Learning*, 2017, pp. 2305–2313.

[24] S. Steinschneider and U. Lall, "Daily precipitation and tropical moisture exports across the Eastern United States: An application of archetypal analysis to identify spatiotemporal structure," *Journal of Climate*, vol. 28, no. 21, pp. 8585–8602, 2015.

[25] Z. Su, Z. Hao, F. Yuan, X. Chen, and Q. Cao, "Spatiotemporal variability of extreme summer precipitation over the Yangtze river basin and the associations with climate patterns," *Water*, vol. 9, no. 11, 2017.

[26] I. Epifanio, G. Vinué, and S. Alemany, "Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem," *Computers & Industrial Engineering*, vol. 64, no. 3, pp. 757–765, 2013.

[27] A. Alcacer, I. Epifanio, M. V. Ibáñez, A. Simó, and A. Ballester, "A data-driven classification of 3D foot types by archetypal shapes based on landmarks," *PLOS ONE*, vol. 15, no. 1, p. e0228016, 2020.

[28] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon, "Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space," *Science*, vol. 336, no. 6085, pp. 1157–1160, 2012.

[29] J. C. Thøgersen, M. Mørup, S. Damkiær, S. Molin, and L. Jelsbak, "Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways," *BMC Bioinformatics*, vol. 14, p. 279, 2013.

[30] Y. Wang and H. Zhao, "Non-linear archetypal analysis of single-cell RNA-seq data by deep autoencoders," *PLoS computational biology*, vol. 18, no. 4, p. e1010025, 2022.

[31] E. N. Zois, I. Theodorakopoulos, and G. Economou, "Offline handwritten signature modeling and verification based on archetypal analysis," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5515–5524.

[32] W. Sun, G. Yang, K. Wu, W. Li, and D. Zhang, "Pure endmember extraction using robust kernel archetypoid analysis for hyperspectral imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 131, pp. 147 – 159, 2017.

[33] W. Sun, D. Zhang, Y. Xu, L. Tian, G. Yang, and W. Li, "A probabilistic weighted archetypal analysis method with Earth mover's distance for endmember extraction from hyperspectral imagery," *Remote Sensing*, vol. 9, no. 8, p. 841, 2017.

[34] I. Cabero and I. Epifanio, "Archetypal analysis: an alternative to clustering for unsupervised texture segmentation," *Image Analysis & Stereology*, vol. 38, pp. 151–160, 2019.

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

[36] A. Alcacer, I. Epifanio, J. Valero, and A. Ballester, "Combining classification and user-based collaborative filtering for matching footwear size," *Mathematics*, vol. 9, no. 7, 2021.

[37] S. M. Keller, M. Samarin, F. Arend Torres, M. Wieser, and V. Roth, "Learning extremal representations with deep archetypal analysis," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 805–820, 2021.

[38] G. C. Porzio, G. Ragozini, and D. Vistocco, "On the use of archetypes as benchmarks," *Applied Stochastic Models in Business and Industry*, vol. 24, pp. 419–437, 2008.

[39] E. Canhasi and I. Kononenko, "Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization," *Expert Systems with Applications*, vol. 41, no. 2, pp. 535 – 543, 2014.

[40] M. Fernandez and A. S. Barnard, "Identification of nanoparticle prototypes and archetypes," *ACS Nano*, vol. 9, no. 12, pp. 11 980–11 992, 2015.

[41] A. Tsanousa, N. Laskaris, and L. Angelis, "A novel single-trial methodology for studying brain response variability based on archetypal analysis," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8454 – 8462, 2015.

[42] J. L. Hinrich, S. E. Bardenfleth, R. E. Roge, N. W. Churchill, K. H. Madsen, and M. Mørup, "Archetypal analysis for modeling multisubject fMRI data," *IEEE Journal on Selected Topics in Signal Processing*, vol. 10, no. 7, pp. 1160–1171, 2016.

[43] M. J. A. Eugster, "Performance profiles based on archetypal athletes," *International Journal of Performance Analysis in Sport*, vol. 12, no. 1, pp. 166–187, 2012.

[44] G. Vinué and I. Epifanio, "Archetypoid analysis for sports analytics," *Data Mining and Knowledge Discovery*, vol. 31, no. 6, pp. 1643–1677, 2017.

[45] ——, "Forecasting basketball players' performance using sparse functional data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 12, pp. 534–547, 2019.

[46] I. Epifanio, M. V. Ibáñez, and A. Simó, "Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles," *The American Statistician*, vol. 74, no. 2, pp. 169–183, 2020.

[47] M. J. A. Eugster and F. Leisch, "Weighted and robust archetypal analysis," *Computational Statistics & Data Analysis*, vol. 55, no. 3, pp. 1215–1225, 2011.

[48] J. Moliner and I. Epifanio, "Robust multivariate and functional archetypal analysis with application to financial time series analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 519, pp. 195 – 208, 2019.

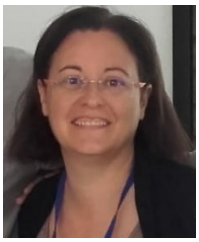
[49] M. R. D'Esposito, F. Palumbo, and G. Ragozini, "Interval Archetypes: A New Tool for Interval Data Analysis," *Statistical Analysis and Data Mining*, vol. 5, no. 4, pp. 322–335, 2012.

[50] G. Vinué, I. Epifanio, and S. Alemany, "Archetypoids: A new approach to define representative archetypal data," *Computational Statistics & Data Analysis*, vol. 87, pp. 102 – 115, 2015.

[51] I. Epifanio, "Functional archetype and archetypoid analysis," *Computational Statistics & Data Analysis*, vol. 104, pp. 24 – 34, 2016.

[52] G. Ragozini, F. Palumbo, and M. R. D'Esposito, "Archetypal analysis for data-driven prototype identification," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 10, no. 1, pp. 6–20, 2017.
[53] S. Seth and M. J. A. Eugster, "Probabilistic archetypal analysis," *Machine Learning*, vol. 102, no. 1, pp. 85–113, 2016.
[54] ——, "Archetypal analysis for nominal observations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 849–861, 2016.
[55] D. Fernández, I. Epifanio, and L. F. McMillan, "Archetypal analysis for ordinal data," *Information Sciences*, vol. 579, pp. 281–292, 2021.
[56] A. S. Olsen, R. M. T. Høegh, J. L. Hinrich, K. H. Madsen, and M. Mørup, "Combining electro- and magnetoencephalography data using directional archetypal analysis," *Frontiers in Neuroscience*, vol. 16, 2022.
[57] I. Epifanio, M. V. Ibáñez, and A. Simó, "Archetypal shapes based on landmarks and extension to handle missing data," *Advances in Data Analysis and Classification*, vol. 12, no. 3, pp. 705–735, 2018.
[58] D. van Dijk, D. B. Burkhardt, M. Amodio, A. Tong, G. Wolf, and S. Krishnaswamy, "Finding archetypal spaces using neural networks," in *2019 IEEE International Conference on Big Data*. IEEE, 2019, pp. 2634–2643.
[59] L. Millán-Roures, I. Epifanio, and V. Martínez, "Detection of anomalies in water networks by functional data analysis," *Mathematical Problems in Engineering*, vol. 2018, no. Article ID 5129735, p. 13, 2018.
[60] G. Vinué and I. Epifanio, "Robust archetypoids for anomaly detection in big functional data," *Advances in Data Analysis and Classification*, pp. 1–26, 2020.
[61] I. Cabero, I. Epifanio, A. Piérola, and A. Ballester, "Archetype analysis: A new subspace outlier detection approach," *Knowledge-Based Systems*, vol. 217, p. 106830, 2021.
[62] A. Tendler, A. Mayo, and U. Alon, "Evolutionary tradeoffs, Pareto optimality and the morphology of ammonite shells," *BMC systems biology*, vol. 9, no. 1, pp. 1–12, 2015.
[63] V. Maurizio, "Double k-means clustering for simultaneous classification of objects and variables," in *Advances in Classification and Data Analysis*, S. Borra, R. Rocci, M. Vichi, and M. Schader, Eds. Berlin, Heidelberg: Springer, 2001, pp. 43–52.
[64] G. Govaert and M. Nadif, "Clustering with block mixture models," *Pattern Recognition*, vol. 36, no. 2, pp. 463–473, 2003.
[65] P. S. Bhatia, S. Iovleff, and G. Govaert, "blockcluster: An R package for model-based co-clustering," *Journal of Statistical Software*, vol. 76, no. 9, p. 1–24, 2017.
[66] M. J. Eugster and F. Leisch, "From Spider-Man to Hero - Archetypal Analysis in R," *Journal of Statistical Software*, vol. 30, no. 8, pp. 1–23, 2009.
[67] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, 1974.
[68] I. Epifanio, "Cargas de trabajo no presencial ECTS arquetípicas del estudiantado:¿cómo se reparten el trabajo semanalmente?" in *Actas del Congreso Virtual: Avances en Tecnologías, Innovación y Desafíos de la Educación Superior ATIDES 2016*, 2016, pp. 367–376.
[69] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 3, p. 1–67, 2011.
[70] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor *et al.*, "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, no. 6795, pp. 536–540, 2000.
[71] R. Rocci and M. Vichi, "Two-mode multi-partitioning," *Computational Statistics and Data Analysis*, vol. 52, pp. 1984–2003, 1 2008.
[72] F. Martella, M. Alfò, and M. Vichi, "Biclustering of gene expression data by an extension of mixtures of factor analyzers," *The International Journal of Biostatistics*, vol. 4, no. 1, 2008. [Online]. Available: https://doi.org/10.2202/1557-4679.1078
[73] K. Lang, "Newsweeder: Learning to filter netnews," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 331–339.
[74] J. M. Cotelo, F. J. Ortega, J. A. Troyano, F. Enríquez, and F. L. Cruz, "Known by who we follow: A biclustering application to community detection," *IEEE Access*, vol. 8, pp. 192 218–192 228, 2020.
[75] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.
[76] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 269–274.
[77] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: coclustering genes and conditions," *Genome research*, vol. 13, no. 4, pp. 703–716, 2003.
[78] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *Proceedings of the 15th European Conference on Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2004, p. 217–226.
[79] E. N. Castanho, H. Aidos, and S. C. Madeira, "Biclustering fMRI time series: a comparative study," *BMC bioinformatics*, vol. 23, no. 1, pp. 1–30, 2022.
[80] M. G. Silva, S. C. Madeira, and R. Henriques, "Water consumption pattern analysis using biclustering: When, why and how," *Water*, vol. 14, no. 12, 2022.
[81] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

Aleix Alcacer is currently a Ph.D. candidate in Mathematics at Universitat Jaume I (Spain) thanks to a FPU grant (FPU20/01825) from the Spanish Ministry of Universities. Aleix obtained his M.Sc. in Computational Mathematics from Universitat Jaume I in 2020 and his B.Sc. in Computational Mathematics from the same institution in 2018. He was the recipient of multiple awards from the university during his undergraduate studies. His research interests include Archetypal Analysis, Matrix Factorization, and Data Visualisation. He is passionate about applying mathematical techniques to real-world problems and exploring new methods for data analysis and visualization.



Irene Epifanio received the M.S. degree in mathematics and the Ph.D. degree in statistics from the Universitat de València (Spain), in 1997 and 2002. In 1999, she joined the Computer Science Department, Universitat de València. In 2002, she was a postdoctoral researcher with Joint Research Centre (JRC) of the European Commission, Ispra, Italy. Since October 2000 she has been with the Department of Mathematics, Universitat Jaume I, Castelló, Spain, where she is currently a Full Professor since March 2021. Her current research interests include statistical learning, functional data analysis, computer vision and equality. She is the recipient of various awards in research, teaching, scientific dissemination and social commitment, including the Margarita Salas Prize of Talent Woman.

Ximo Gual-Arnau graduated in mathematics and got his Ph.D. in Integral Geometry and Stereology at the University of Valencia, Spain, in 1990 and 1995, respectively. On March 1, 2009, he received full professorship recognition and became a full professor at the University Jaume I of Castelló in Spain. He has been invited as a keynote speaker in many different events worldwide to describe and discuss the current state of the art on integral geometry, stereology, and applications to morphological image analysis. He was one of the founders and is a member of the Institute of New Imaging Technologies at the Universitat Jaume I