Archetypal Analysis of Interval Data*

Stefania Corsaro
University of Naples Parthenope
corsaro@uniparthenope.it

Marina Marino
University of Naples Federico II

Abstract

In this paper we present a mathematical model for archetypal analysis of data represented by means of intervals of real numbers. We extend the model for single-valued data proposed in the pioneering work of Cutler and Breiman on this topic. The core problem is a non-convex optimization one, which we solve by means of a sequential quadratic programming method. We show numerical experiments performed on both single-valued and interval data in order to validate the model.

Keywords: interval analysis, archetypal analysis, sequential quadratic programming, non-convex optimization
AMS subject classifications: 62-07, 65G20, 65K05

1 Introduction

Archetypal Analysis (AA) is a statistical technique, first introduced in [3], which aims, given a data set (set of *individuals*), at defining a set of *archetypes* such that each individual can be uniquely represented as a combination of the archetypes. Early applications of AA concerned the design of face masks for the army [3], in the following, Archetypal Analysis has been applied in several fields, among which marketing [4] and fluid dynamics [10].
AA presents conceptual similarities with Principal Component Analysis (PCA). However, while the central idea of PCA is to reduce the dimensionality of a data set, retaining as much as possible of the variation present in it, archetypes characterize extreme data values on the convex hull of the data set. Thus, PCA is basically a dimensionality reduction technique, AA is not necessarily aimed at this: AA does not produce an orthogonal basis, indeed, one can fit more archetypes than the number of dimensions. Comparisons between PCA and AA have been discussed in [3, 10], from which it emerged that the performance of PCA versus AA actually depends on

*Submitted: January 20, 2009; First revision: July 9, 2009; Second revision: July 28, 2009 Accepted: September 17, 2009.

the structure of the data set. Moreover, in [11] PCA and AA have been applied in combination for taking advantage from the specific properties of the two methods.
Here we recall the original formulation of the problem. In section 3 we give an equivalent matrix formulation in order to extend the mathematical model to interval data. Let $\mathbf{x}_i \in \mathbb{R}^n$, $i = 1, ..., m$ represent the individuals and p the dimension of the space spanned by the archetypes. Archetypal Analysis aims at finding p vectors $\mathbf{z}_k \in \mathbb{R}^n$, $k = 1, ..., p$ that characterize the archetypal pattern in the data, that is, they minimize the quantity:

$$\sum_{i=1}^{m} \|\mathbf{x}_i - \sum_{k=1}^{p} \alpha_{ik}\mathbf{z}_k\|_2^2. \tag{1}$$

The core of Archetypal Analysis is, thus, the minimization of a non-convex function. The following conditions are imposed to the coefficients in (1):

$$\begin{array}{l} \alpha_{ik} \geq 0 \\ \sum_{k=1}^{p} \alpha_{ik} = 1, \quad i = 1, \ldots, m. \end{array} \tag{2}$$

Relations (2) state that the individuals belong to the convex hull of the archetypes. Furthermore, the archetypes are supposed to be linear combinations of the individuals, thus:

$$\mathbf{z}_k = \sum_{i=1}^{m} \beta_{ki}\mathbf{x}_i, \quad k = 1, \ldots, p. \tag{3}$$

More precisely, archetypes are supposed to belong to the convex hull of the individuals, so the coefficients of the linear combinations in (3) are subject to the following constraints:

$$\begin{array}{l} \beta_{ki} \geq 0 \\ \sum_{i=1}^{m} \beta_{ki} = 1, \quad k = 1, \ldots, p. \end{array} \tag{4}$$

The AA problem has been approached by means of alternating minimization algorithms [3]. Due to the non-convexity of the objective function, convergence to a global optimal solution cannot be guaranteed. In that paper, authors performed an empirical analysis of the convergence of the algorithm and they observed that convergence to local minima or other stationary points becomes a crucial matter when the number of archetypes increases.
A meaningful upper bound for the number p of archetypes is the cardinality N of the set of points defining the convex hull of the individuals. In that case, indeed, choosing as archetypes the N boundary data points of the convex hull, then the value of (1) is zero [3]. This number is actually unknown in practice, thus, the "best" number of archetypes is, generally, empirically computed.
The use of interval arithmetic techniques in Archetypal Analysis allows one to obtain models which take into account valuable information. For instance, when dealing with marketing applications, often product features and consumer preferences are more adequately expressed by a range of values. Therefore, interval arithmetic techniques provide more reliable statistical analysis methods. In this paper, we propose a mathematical model for the Archetypal Analysis of interval data; we do not deal here with numerical issues in the solution process. In the following of the paper we show the results of numerical experiments we performed both on single-valued and interval data, with the purpose of validating our model.
The paper is organized as follows: in section 2 we briefly recall basic elements of interval arithmetic, which are essential in our discussion; in section 3 we present our mathematical model for Interval Archetypal Analysis (IAA); in section 4 we show the

results of the numerical experiments we performed both on single-valued data and interval data to validate the model. Finally, in section 5 we give some conclusions and outline our future activity on this topic.

2 Interval Data

In this section we briefly recall basic elements of interval arithmetics which are essential in our discussion. For a deep insight into the matter we refer to [2, 8].
Let

$$\mathbf{x} = [\underline{x}, \overline{x}] = [x_c - \Delta x, x_c + \Delta x]$$

be an interval of real numbers. The set of such intervals is denoted with $\mathbb{IR}$. Given two intervals $\mathbf{x}$, $\mathbf{y}$ we define the arithmetic operations in the following way:

$$\begin{array}{l} \mathbf{x} + \mathbf{y} := [\underline{x} + \underline{y}, \overline{x} + \overline{y}] \\ \mathbf{x} - \mathbf{y} := [\underline{x} - \overline{y}, \overline{x} - \underline{y}] \\ \mathbf{x} \cdot \mathbf{y} := [\min(\underline{x} \cdot \underline{y}, \underline{x} \cdot \overline{y}, \overline{x} \cdot \underline{y}, \overline{x} \cdot \overline{y}), \max(\underline{x} \cdot \underline{y}, \underline{x} \cdot \overline{y}, \overline{x} \cdot \underline{y}, \overline{x} \cdot \overline{y})] \\ \mathbf{x}/\mathbf{y} := [\underline{x}, \overline{x}] \cdot [1/\overline{y}, 1/\underline{y}] \quad \text{if } 0 \notin \mathbf{y} \end{array} \tag{5}$$

Given n intervals:

$$[\underline{x}_i, \overline{x}_i], \quad i = 1, \ldots, n$$

we define their mean as the interval:

$$\left[\frac{1}{n} \sum_{i=1}^{n} \underline{x}_i, \frac{1}{n} \sum_{i=1}^{n} \overline{x}_i \right]. \tag{6}$$

We consider the *distance* between two given intervals $\mathbf{x}$ and $\mathbf{y}$; the distance function

$$q : (\mathbf{x}, \mathbf{y}) \in \mathbb{IR} \times \mathbb{IR} \mapsto q(\mathbf{x}, \mathbf{y}) \in \mathbb{R}_0^+$$

is defined by

$$q(\mathbf{x}, \mathbf{y}) = \sup\{|\underline{x} - \underline{y}|, |\overline{x} - \overline{y}|\} = |x_c - y_c| + |\Delta x - \Delta y|. \tag{7}$$

$(\mathbb{IR}, q)$ is a complete metric space.
Let $m,\ n \in \mathcal{N}$, and $\underline{A}, \overline{A} \in \mathbb{R}^{m \times n}$ be two rectangular matrices with $\underline{A} \leq \overline{A}$, where comparison operators are to be understood componentwise. The set of matrices

$$\mathbf{A} = [\underline{A}, \overline{A}] = \{A \in \mathbb{R}^{m \times n} :\ \underline{A} \leq A \leq \overline{A}\}$$

is called *interval matrix*, $\mathbb{IR}^{m \times n}$ is the set of interval matrices of dimension $m \times n$. According with standard notations, we denote by $\mathbf{A}_{ij}$ the generic element in the interval matrix $\mathbf{A}$. Furthermore, we denote with:

$$A_c = \frac{1}{2}(\underline{A} + \overline{A}), \quad \Delta A = \frac{1}{2}(\overline{A} - \underline{A})$$

the *center* and the *radius matrices* respectively, so

$$\mathbf{A} = [\underline{A}, \overline{A}] = [A_c - \Delta A, A_c + \Delta A].$$

When $\Delta A = 0$, the interval matrix is said *thin*. A special case of interval matrices are interval vectors

$$\mathbf{b} = [\underline{b}, \overline{b}] = \{b \in \mathbb{R}^{m \times 1} :\ \underline{b} \leq b \leq \overline{b}\}.$$

Operations between interval matrices are formally defined as the corresponding ones between single valued matrices, the pointwise algebraic operations are to be meant according to (5).
Let $\mathbf{A}$, $\mathbf{B}$ be two interval matrices. We define *distance matrix* between $\mathbf{A}$ and $\mathbf{B}$ the following non-negative matrix:

$$(q(\mathbf{A},\mathbf{B})_{i,j}) := (q(\mathbf{A}_{ij},\mathbf{B}_{ij})), \tag{8}$$

that is, the pointwise distance, to be meant in the sense of (7), between the elements of $\mathbf{A}$ and $\mathbf{B}$. It can be shown that if $||\cdot||$ denotes a real matrix norm, then $||q(\mathbf{A},\mathbf{B})||$ defines a metric on the set of interval matrices [2].

3 A mathematical model for Interval Archetypal Analysis

In this section we present our mathematical model for Archetypal Analysis of interval data. We start by giving an equivalent matrix formulation of the problem, since it facilitates the extension of the mathematical model discussed in section 1 to interval data.
Let us organize the data into a matrix $X = (x_{ij}) \in \mathbb{R}^{m\times n}$, in which the rows refer to the individuals and the columns to the variables. Then, it can be easily seen that the core problem of Archetypal Analysis can be stated in the following way:

Problem 1. *given a matrix $X \in \mathbb{R}^{m\times n}$, and an integer p, find matrices $A = (\alpha_{ik}) \in \mathbb{R}^{m\times p}$, $B = (\beta_{ki}) \in \mathbb{R}^{p\times m}$ which solve the non-convex minimization problem:*

$$\min_{A,B} f(A,B) = \min_{A,B} \|X - A\cdot B\cdot X\|_F, \tag{9}$$

where $\|\cdot\|_F$ denotes, as usual, the Frobenius norm, under the constraints (2), (4)*. The archetypes are then the rows of the matrix $Z \in \mathbb{R}^{p\times n}$ given by the product:*

$$Z = B\cdot X. \tag{10}$$

□

Let $\mathbf{X}, \mathbf{Y} \in \mathbb{IR}^{m\times n}$ be two interval matrices. According to (8), the following norm:

$$\|q(\mathbf{X},\mathbf{Y})\|_F = \||X_c - Y_c| + |\Delta X - \Delta Y|\|_F$$

is a metric on the set $\mathbb{IR}^{m\times n}$. We refer to this metric for defining the objective function in Interval Archetypal Analysis. Our choice is motivated by the fact that this metric allows us to keep under control both the distance between the centers, for the sake of localization, and the width of the involved intervals, for the sake of accuracy. Moreover, the Frobenius norm is the natural choice for it is the one employed in the original definition of Archetypal Analysis given by Cutler and Breiman, and it can be efficiently computed.
We state the IAA problem:

Problem 2. *Let $\mathbf{X} \in \mathbb{IR}^{m\times n}$ be an interval matrix, in which the rows represent the individuals and the columns the variables. Given an integer p, find matrices $A = (\alpha_{ik}) \in \mathbb{R}^{m\times p}$, $B = (\beta_{ki}) \in \mathbb{R}^{p\times m}$ which solve the non-convex minimization problem:*

$$\min_{A,B} \|q(\mathbf{X}, A\cdot B\cdot \mathbf{X})\|_F = \||X_c - (A\cdot B\cdot \mathbf{X})_c| + |\Delta X - \Delta(A\cdot B\cdot \mathbf{X})|\|_F \tag{11}$$

under the constraints (2), (4). *The archetypes are then given by the row vectors of the matrix* $\mathbf{Z} \in \mathbb{IR}^{p\times n}$ *defined by the product:*

$$\mathbf{Z} = B \cdot \mathbf{X}. \tag{12}$$

□

The IAA problem is again a non-convex, single-valued optimization problem.
Some preliminary experiments concerning Interval Archetypal Analysis of marketing data were shown in [4]. In that paper, relation (12) is violated, that is, the archetypes do not belong to the convex hull of the data, coherently with the original formulation of the single-valued problem, expressed by equation (10). In [4] authors compute two real matrices, namely B_c and B_r, such that:

$$Z_c = B_c \cdot X_c, \quad \Delta Z = B_r \cdot \Delta X,$$

where the elements of B_c and B_r separately satisfy the constraints (4). Thus, centers and radii of the archetypes respectively belong to convex hull of the centers and the radii of the data.

4 Numerical experiments

In this section we discuss the results of the numerical experiments we performed to validate our model for IAA. Our purpose in this framework is to analyze the coherence between the results obtained from Archetypal Analysis of single-valued data and the ones obtained working with interval data. A relevant issue in interval data computations is the capability of accurately approximating the centers, since it allows one to preserve the location of the intervals on the real axis. For this reason, we start from an interval data set and we work on the centers of the intervals when dealing with single-valued Archetypal Analysis. We consider the data set presented in [7]; data refer to a set of sixteen different fruit juices that were submitted to a group of judges called to assign a score to six features, namely, appearance, smell, taste, naturalness, sweetness and density. Since the interindividual differences in judges were unknown, data were organized into an interval matrix. More precisely, for the rating of each juice with respect to each feature, the lower bound and the upper bound were observed.
Here we do not focus on numerical issues in the solution process, therefore we developed the algorithms in MatLab environment, using the routine `fmincon` of MatLab Optimization Toolbox [1] to solve the core minimization problem both in the single-valued case and in the interval data case. The method implemented in the mentioned optimization routine is a sequential quadratic programming one [5, 6]. This is an iterative method in which the basic idea is to solve, at each iteration, a quadratic programming problem that is formulated considering a quadratic approximation of the Lagrangian function associated to the optimization problem. The algorithm implemented is based on a line search strategy, that is, at each iteration the algorithm chooses a direction and searches along this direction for a new iterate, starting from the current one. In particular, the algorithm chooses Quasi-Newton search directions. An approximation of the Hessian is used, which is updated after each step to take account of the additional knowledge gained during the step [9].

	Appearance	*Smell*	*Taste*	*Naturalness*	*Sweetness*	*Density*
Pineapple1	7.14	6.14	6.66	6.02	6.27	4.58
Pineapple2	7.04	6.64	6.30	6.00	6.12	4.18
Orange1	7.04	7.56	6.84	6.22	6.16	4.28
Orange2	7.24	6.48	7.24	6.22	6.41	4.52
Grapefruit1	6.78	7.12	6.02	6.68	3.02	4.28
Grapefruit2	6.82	6.27	6.88	6.54	4.58	3.86
Peer1	7.28	7.66	7.58	6.90	7.98	7.70
Peer2	7.86	6.84	8.01	7.10	8.08	7.16
Apricot1	7.16	8.16	7.98	7.88	7.79	7.43
Apricot2	7.84	7.49	5.92	5.32	5.65	6.30
Peach1	7.51	7.32	6.98	6.38	7.16	5.48
Peach2	7.38	6.70	7.77	7.22	7.35	5.41
Apple1	7.14	6.04	7.79	6.32	7.63	6.26
Apple2	7.22	6.79	6.83	6.39	7.12	6.04
Banana1	5.73	4.67	4.67	5.16	5.45	4.26
Banana2	5.96	4.53	4.14	4.67	5.10	4.30

Table 1: Single-valued data.

4.1 Single-valued data

In this section, we summarize the results of some numerical experiments we performed on single-valued data. The matrix X reported in table 1 contains the data we analyze. In all the numerical experiments we are going to discuss, we confine the precision to two decimal digits, coherently with the available information on the data.
As already pointed out, it should be $p \leq N$, where N denotes the number of boundary points of the data convex hull, a typically unknown value. For this reason, in the following we analyze the behaviour of the objective function as p increases. For the non-convexity of the objective function, local minima are obtained, therefore, we performed one thousand runs with randomly chosen starting points and referred to the minimum value. In order to determine the number of archetypes which realizes the best trade-off between accuracy and efficiency, we considered a wide range of p values.
In figure 1 we represent, on the left, the percentage of decrease of the value of the objective function at the optimum as the number p of archetypes increases, ranging from 1 to 9, that is, the ratio:

$$\frac{f_p^*}{f_1^*} \cdot 100, \quad p = 1, ..., 9 \tag{13}$$

where f_p^* is the minimum of the objective function with p archetypes. On the right, we report the values:

$$\frac{f_{p-1}^* - f_p^*}{f_{p-1}^*} \cdot 100, \quad p = 2, ..., 9 \tag{14}$$

versus the number of archetypes, that is, we represent the relative percentage of decrease of the objective function. Looking at the left-side graphic, we note that for $p = 6$ we have reduced the value of the objective function by 80% with respect of its optimal value for $p = 1$. On the other hand, for $p > 6$ the gain in terms of decrease of the minimum reduces. Indeed, we have that the percentage of absolute decrease (13) is reduced of four points only when passing from $p = 6$ to $p = 7$. In terms of relative percentage decrease (14), shown in the right-side graphic, as p increases, ranging from six to seven, we gain less than the 20%, thus, $p = 6$ is a good choice for the dimension

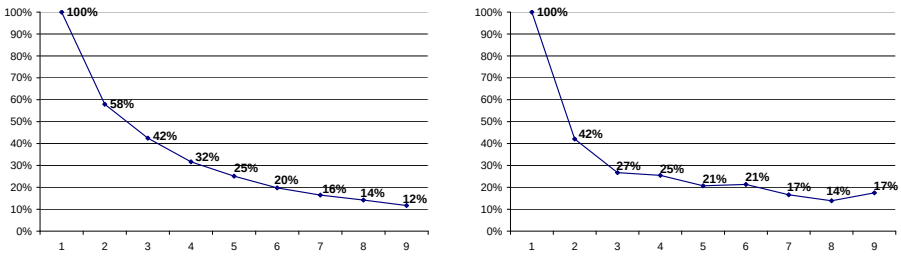

Figure 1: Left: absolute percentage of decrease of the objective function; right: relative percentage of decrease of the objective function.

spanned by the archetypes. Moreover, in table 2 we report the estimated values of (9), with the relative frequency of the detected local minima, for different values of archetypes. More precisely, we sort the values in descending order with respect to their relative frequencies and then report the values occurred with a frequency not smaller than 5%. In our experiments, the archetypes corresponding to a fixed stationary point of the objective function were the same for each occurrence of the mentioned point. We note that, up to $p = 4$, we just detect one minimum, for $p = 5, 6, 7$ the algorithms finds three stationary points. When $p = 7$ the two most frequent stationary points (namely, 1.66 and 1.70) are quite close, finally, when $p = 8, 9$ the number of local minima significantly increases. This analysis lead us to the observation that, especially for $p > 6$, numerical problems related to the conditioning of the involved matrices most probably occur. As already pointed out, we do not deal with this aspect in the present work, we plan to focus on this in the next future.
We then consider $p = 1$; in this case, we obtain that, within two decimal digits, the archetype is given by the sample mean of the features, which confirms the theoretical result shown in [3].

4.2 Interval data

In this section, we summarize the results of some numerical experiments we performed on interval data. In our experiments, we refer again to the juices data set, now in their original interval structure, reported in table 3. We performed one hundred runs with randomly chosen starting points, because of the higher computational complexity of the interval-based minimization algorithm with respect to the single-valued version, where we performed one thousand runs. As done for single-value Archetypal Analysis, in figure 2 we again represent, on the left, the percentage of decrease of the value of the objective function at the optimum as the number of archetypes increases, on the right, the relative percentage of decrease of the objective function, up to $p = 6$, which is the "best" value of p computed in the single-valued data framework. Moreover, in table 4 we report the estimated values of (11), with the relative frequency of the detected local minima, for different values of archetypes.

Comparing figures 1 and 2, we note that the slopes of the curves which give the rate of decrease of the detected minima with respect to p in single-valued Archety-

number of archetypes	*objective function*	*frequency*
1	10.08	100%
2	5.84	100%
3	4.28	100%
4	3.19	97.1%
5	2.53	53.5%
	2.76	36.4%
	3.04	7.2%
6	1.99	45.9%
	2.53	37%
	2.22	11.7%
7	1.66	66%
	1.70	20%
	2.36	8%
8	1.43	51%
	1.57	7%
	1.47	6%
	2.26	6%
	1.49	5%
	1.59	5%
	1.63	5%
	2.21	5%
9	1.35	18%
	1.26	16%
	1.18	10%
	1.40	9%
	1.21	8%
	1.59	6%
	1.38	5%

Table 2: local minima, with the respective relative frequency, for different values of archetypes.

	Appearance	*Smell*	*Taste*	*Naturalness*	*Sweetness*	*Density*
Pineapple1	[6.61,7.66]	[5.74,6.66]	[6.18,7.31]	[5.45,6.85]	[5.63,6.75]	[3.92,5.00]
Pineapple2	[6.66,7.59]	[5.90,7.30]	[5.65,6.98]	[5.23,6.56]	[5.52,6.92]	[3.28,4.69]
Orange1	[6.64,7.59]	[7.12,8.24]	[6.39,7.44]	[5.67,6.72]	[5.75,6.67]	[3.64,4.97]
Orange2	[6.89,7.55]	[6.06,6.90]	[6.82,7.94]	[5.60,6.72]	[5.93,7.13]	[3.88,4.98]
Grapefruit1	[6.28,7.40]	[6.52,7.65]	[5.17,6.85]	[6.00,7.33]	[2.45,3.39]	[3.64,4.76]
Grapefruit2	[6.31,7.43]	[5.63,6.75]	[6.35,7.47]	[6.11,7.23]	[4.14,5.19]	[3.06,4.46]
Peer1	[6.89,7.76]	[7.19,8.24]	[7.14,8.19]	[6.44,7.49]	[7.59,8.54]	[7.22,8.27]
Peer2	[7.52,8.20]	[6.32,7.44]	[7.69,8.57]	[6.72,7.63]	[7.71,8.62]	[6.72,7.67]
Apricot1	[6.82,7.68]	[7.87,8.68]	[7.60,8.54]	[7.35,8.47]	[7.42,8.40]	[7.03,8.15]
Apricot2	[7.32,8.16]	[7.09,8.19]	[5.17,6.71]	[4.66,6.06]	[4.90,6.31]	[5.79,6.77]
Peach1	[7.09,7.93]	[6.94,7.78]	[6.42,7.54]	[5.70,7.10]	[6.69,7.68]	[5.03,5.92]
Peach2	[6.98,7.82]	[6.22,7.11]	[7.38,8.38]	[6.83,7.72]	[6.83,7.81]	[4.99,5.85]
Apple1	[6.78,7.52]	[5.47,6.59]	[7.40,8.40]	[5.66,7.20]	[7.27,8.29]	[5.81,6.74]
Apple2	[6.60,7.72]	[6.28,7.40]	[6.31,7.43]	[5.72,7.12]	[6.67,7.65]	[5.47,6.59]
Banana1	[4.96,6.37]	[3.92,5.60]	[3.64,5.32]	[4.27,5.95]	[4.76,6.16]	[3.62,4.74]
Banana2	[5.27,6.67]	[3.68,5.36]	[3.26,4.94]	[3.92,5.46]	[4.23,5.91]	[3.65,4.77]

Table 3: Interval data.

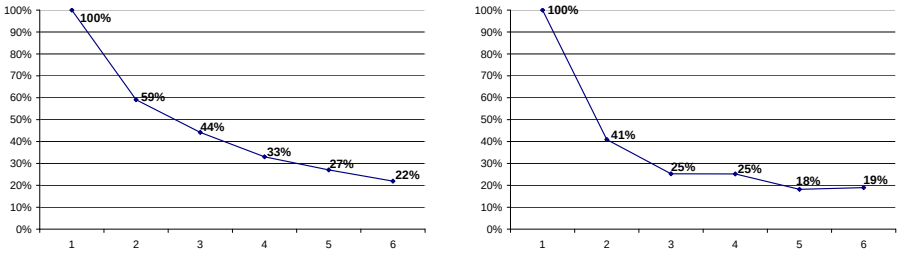

Figure 2: Left: absolute percentage of decrease of the objective function; right: relative percentage of decrease of the objective function.

number of archetypes	*objective function*	*frequency*
1	11.32	100%
2	6.69	92%
	6.70	7%
3	5.00	85%
	5.01	6%
4	3.74	76%
	4.33	9%
5	3.06	26%
	3.67	19%
	3.21	13%
	3.39	7%
	3.22	6%
6	3.00	17%
	2.48	16%
	2.69	10%
	3.14	8%

Table 4: values of the objective function, with the respective relative frequency, for different values of archetypes.

pal Analysis and IAA are almost the same, as the reported percentage values reveal. This means that the convergence rate is almost preserved when passing in interval arithmetics framework. Looking at the graphics of figure 2, we note that both the percentage of absolute (13) and relative decrease (14) are almost the same as p ranges from four to five and from five to six. On the other hand, from the table it is clear that it is not worth considering $p > 5$, since we have two stationary points with very close relative frequencies, a situation which reveals that the numerical problems observed in the single-valued case are, not surprisingly, more severe in this case. We therefore fix $p = 5$ as "best" value for the number of archetypes for intervals. We now consider the case $p = 1$. In this case, we again expect the archetype to be the sample mean of the features, now represented by intervals, according to (6). In table 5 we report the results we obtained; the maximum distance between the "mean juice" and the archetype is 0.08, the maximum error in centers approximation is 0.04.
In figure 3 we represent the coordinates of the archetypes in the variables space for

	Appearance	*Smell*	*Taste*	*Naturalness*	*Sweetness*	*Density*
data	[6.60,7.57]	[6.12,7.24]	[6.16,7.38]	[5.71,6.98]	[5.84,6.97]	[4.80,5.90]
archetype	[6.58,7.58]	[6.11,7.23]	[6.19,7.39]	[5.76,7.01]	[5.85,6.86]	[4.78,5.88]

Table 5: Mean value of data and the estimated archetype for $p = 1$.

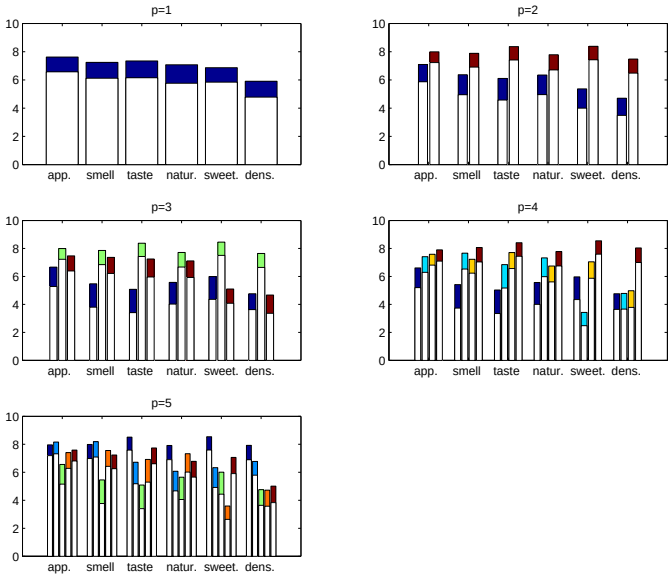

Figure 3: Coordinates of the archetypes in the variables space.

p ranging from one to five. More precisely, the filled area in each bar represents the interval which measures the value of an archetype corresponding to a feature, each archetype being identified by a filling. The intervals reported in the top-left graphic represent the mean values of the juice features; in the others, the first bar in each group refers to the first archetype and so on.
When approximating interval data, it is often important to accurately reconstruct the centers, since it allows one to preserve the location of the intervals on the real axis. Moreover, the width of the intervals must be kept under control, since wide intervals lead to inaccurate estimates. Here we separately analyze the results in centers approximation and radii estimation.
In order to analyze the accuracy in the centers reconstruction, we estimate the maximum absolute error:

$$\max_{i,j} |X_c(i,j) - (A \cdot B \cdot \mathbf{X})_c(i,j)|$$

and the median absolute error over all the reconstructed intervals, for p ranging from one to five. The errors and the median values versus the number of archetypes are

p	1	2	3	4	5
	single-valued data				
maximum error	3.45	1.85	1.64	1.10	0.84
median	0.74	0.39	0.27	0.17	0.15
	interval data				
maximum error	3.44	1.90	1.67	1.21	0.84
median	0.76	0.37	0.27	0.14	0.13

Table 6: mean error and median error in single valued data and interval centers reconstruction versus the number of archetypes.

p	1	2	3	4	5
maximum error	52%	51%	50%	31%	29%
median	13%	11%	9%	8%	7%

Table 7: maximum relative error and median relative error in interval radii reconstruction versus the number of archetypes.

reported in table 6, together with the same quantities estimated in single-valued data approximation, since the single-valued data are actually the centers of our interval data. From table 6, we observe that the same level of accuracy obtained in the single-valued case is preserved, in some cases we obtain a slightly higher accuracy in interval centers reconstruction than in single-valued data approximation. Moreover, we computed some quantiles of the errors; from this analysis, we observe that for $p = 5$ in the 40% of the estimated centers, we preserve at least one decimal digit.
Finally, in table 7 we report the maximum relative error in radii reconstruction:

$$\max_{i,j} \frac{|X_r(i,j) - (A \cdot B \cdot \mathbf{X})_r(i,j)|}{X_r(i,j)}$$

together with the median relative error. We note that the median error is about 0.1 even for $p = 1$ and it is smaller than this value for $p \geq 3$. We computed the mean relative error too; its values are very close to the median error ones.

5 Conclusions and future work

In this work, we presented an extension of the model for Archetypal Analysis of real numbers proposed by Cutler and Breiman to handle real interval numbers. We focused on theoretical issues in the definition of the mathematical model and showed the results of some numerical experiments to validate it. We did not investigate numerical issues in the solution algorithm, developing our procedures in MatLab environment for testing purposes. Results seem promising, since the interval-based model produces results which are as accurate as those obtained in the single-valued data case; nevertheless, numerical results reveal a probable ill-conditioning of the involved matrices, thus, future work will mainly concern numerical analysis of the model.

References

[1] *MatLab Optimization Toolbox 4 Users Guide*. The MathWorks, Inc., 2008.

[2] G. Alefeld and J. Herzberger. *Introduction to Interval Computations*. Academic Press, NY, 1990.

[3] A. Cutler and L. Breiman. Archetypal analysis. *Technometrics*, 36:338–347, 1994.

[4] M.R. D'Esposito, F. Palumbo, and G. Ragozini. Archetypal analysis for interval data in marketing research. *Statistica Applicata*, 18:343–358, 2006.

[5] R. Fletcher. *Practical Methods of Optimization*. John Wiley and Sons, 1987.

[6] P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization*. Academic Press, London, 1981.

[7] P. Giordani and H.A.L. Kiers. A comparison of three methods for principal component analysis of fuzzy interval data. *Computational Statatistics and Data Analysis*, 51:379–397, 2006.

[8] R.E. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, 1979.

[9] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer-Verlag, NY, 1999.

[10] E. Stone and A. Cutler. Introduction to archetypal analysis of spatio-temporal dynamics. *Physica D*, 96:110–131, 1996.

[11] E. Stone and B. Olson. Archetypal Analysis of Cellular Flame Data. Tech. Rep. of Dept Mathematics and Statistics, Utah State University, 1999.