Prototyping through Archetypal Analysis: looking at data from a different perspective

Francesco Palumbo

Università degli Studi di Napoli Federico II
Outline

1. Framework: Prototypes
   - Notion
   - Definition
   - Identification

2. Our proposal

3. A Two-Step Procedure

4. Methodological Advances on AA

5. The study of uncertainty

6. Interval Valued Variables and Statistics

7. Archetypal Analysis for Interval Data

8. Archetypes and Prototypes

9. Final remarks

10. Main references
Outline

1 Framework: Prototypes
   • Notion
   • Definition
   • Identification

2 Our proposal

3 A Two-Step Procedure

4 Methodological Advances on AA

5 The study of uncertainty

6 Interval Valued Variables and Statistics

7 Archetypal Analysis for Interval Data

8 Archetypes and Prototypes

9 Final remarks

10 Main references
A **Prototype**, in the Rosch definition (1978), is an “ideal exemplar” that summarize and represent a group of data, or a category, in terms of their most relevant features and their specificity in contrast to other groups or categories.

For data analysis purposes

- Prototypes can serve as **distillation or a condensed view of a data set**
- Prototypes are usually used as means to build **efficient classifiers** or **prototype-based clustering algorithms**
- However, there is an **inherent value of having a set of prototypical elements** (Bien and Tibshirani, 2011)
A possible operational definition

Prototypes have been defined as elements of the data set that maximize a specified **typicality** or **prototypicality degree** (Rifqi, 1996)

- The prototypicality degree combines two complementary components (Lesot and Kruze, 2006):
  - **internal resemblance**: a more typical point resembles the other members of its categories
  - **external dissimilarity**: a more typical point of one category differs from members of other category
Formally

Given

- a set $\Omega$ of $n$ objects $x_i = (x_{i1}, \ldots, x_{ip})$, a partition $(C_1, \ldots, C_K)$ of $\Omega$ in $K$ groups
- $\rho$ and $\delta$ a resemblance and a dissimilarity measures
- the internal resemblance $R(x, C_k) = P(\rho(x, x_i), x_i \in C_k)$ measures the similarity of $x$ wrt the $x_i$'s belonging to $C_k$
- the external dissimilarity $D(x, C_k) = \Delta(\delta(x, x_i), x_i \notin C_k)$ measures the dissimilarity of $x$ wrt the $x_i$'s not belonging to $C_k$

The prototipicality index is a function $\phi$ combing this two measures:

$$T(x, C_k) = \phi(R(x, C_k), D(x, C_k))$$

A prototype for a group $C_k$ is the point that maximizes $T(x, C_k)$
On the use of prototypes

- The concept of prototypes has found application in supervised and unsupervised learning framework to perform classification and clustering, both crisp and fuzzy.
  - internal resemblance and external dissimilarity match the classification and clustering objectives (homogeneity and separability).

- $\Delta$ and $P$ are often the average (Lesot and Kruse, 2006)
  - then prototypes usually coincide with some *cluster centroids or medoids*
Prototypes for classification and clustering

Drawbacks of Prototypes as centroids or medoids

Prototypes being some average of a group

- could be not well separated among each other and with not clear profiles (Hastie et al., 2009)
- may not adequately describe clusters of arbitrary shape and size (Liu et al., 2009)
- could lead to not informative results in specific fields (Riedesel, 2008)
Outline

1 Framework: Prototypes
   - Notion
   - Definition
   - Identification

2 Our proposal

3 A Two-Step Procedure

4 Methodological Advances on AA

5 The study of uncertainty

6 Interval Valued Variables and Statistics

7 Archetypal Analysis for Interval Data

8 Archetypes and Prototypes

9 Final remarks

10 Main references
Our proposal

Alternative Prototype Identification Procedures

Possible solutions

To overcome these drawbacks it has been proposed to:

- use multiple prototypes to describe clusters of arbitrary shape and size (Liu et al., 2009; Bien and Tibshirani, 2011)
- identify prototypes through the archetypes to have well separated and informative prototypes (Hastie et al., 2009).

but archetypes used as prototypes could produce too extreme representative objects (lack of resemblance)

Our aim

- to keep the external dissimilarity proper of archetypes
- to cope with their possible lack of internal resemblance
A Two-Step Procedure

Outline

1. Framework: Prototypes
   - Notion
   - Definition
   - Identification

2. Our proposal

3. A Two-Step Procedure

4. Methodological Advances on AA

5. The study of uncertainty

6. Interval Valued Variables and Statistics

7. Archetypal Analysis for Interval Data

8. Archetypes and Prototypes

9. Final remarks

10. Main references
A Two-Step Procedure

Prototypes Identification

Our proposal

We propose a two-step procedure

1. **Maximize the external dissimilarity**
   - find the archetypes to exploit their properties (pureness, separability, strong characterization)

2. **Improve their internal resemblance**
   - find clusters around the archetypes in the space spanned by the archetypes
   - maximize the internal resemblance in the space spanned by the archetypes

Additional advantage

- The method can be used for any data type: punctual data, interval data, functional data,....
- overcoming some possible computational problems
Archetypal Analysis (AA) was introduced by Cutler and Breiman (1994) as a new dimensionality reduction approach for multivariate data. The basic idea is to approximate each point in a data set as a convex combination of a set of archetypes.

Archetypal Analysis... are few pure types, pure individual points such that:

- they are a mixture of the observed data
- the observed data can be well represented through a convex mixture of archetypes

enable the researcher:

- to synthesize data through few data points
- to better understand heterogeneity
- to compare data each others

have found several application field:

- Marketing researches
- Physical science
- Medicine
- Multivariate Ordering
- Performance analysis
Archetypal Analysis (AA) was introduced by Cutler and Breiman (1994) as a new dimensionality reduction approach for multivariate data. The basic idea is to approximate each point in a data set as a convex combination of a set of archetypes.

**Archetypes...**

- ...are few *pure* types, pure individual points such that:
  - they are a mixture of the observed data
  - the observed data can be well represented through a convex mixture of archetypes

- ...enable the researcher:
  - to synthesize data through few data points
  - to better understand heterogeneity
  - to compare data each others

- ...have found several application field:
  - Marketing researches
  - Physical science
  - Medicine
  - Multivariate Ordering
  - Performance analysis
Archetypal Analysis (AA) was introduced by Cutler and Breiman (1994) as a new dimensionality reduction approach for multivariate data. The basic idea is to approximate each point in a data set as a convex combination of a set of archetypes.

**Archetypes...**

- ...are few *pure* types, pure individual points such that:
  - they are a mixture of the observed data
  - the observed data can be well represented through a convex mixture of archetypes

- ...enable the researcher:
  - to synthesize data through few data points
  - to better understand heterogeneity
  - to compare data each others

- ...have found several application field:
  - Marketing researches
  - Physical science
  - Medicine
  - Multivariate Ordering
  - Performance analysis
Archetypal Analysis (Cutler and Breiman, 1994)

Formally:

- the archetypes $a_j$, $j = 1, \ldots, m$, are those points such that:

  $$x'_i = \gamma'_i A$$
  \[ (1) \]

  with
  \[ \gamma_{ij} \geq 0 \quad \forall i, j; \quad \gamma'_i 1 = 1 \quad \forall i, \]

- archetypes must be also a mixture of the observed data:

  $$a'_j = \beta'_j X$$
  \[ (2) \]

  with
  \[ \beta_{ji} \geq 0 \quad \forall j, i; \quad \beta'_j 1 = 1 \quad \forall j, \]

- where
  - $X$ is the observed data matrix having generic row $x_i$;
  - coefficients $\beta_{ji}$ are the elements of the $\beta'_j$ vectors, i.e., the weights of the $n$ units;
  - $A$ is the archetype matrix with $a'_j$ its $j$-th row;
  - $\gamma'_i$ is the vector of the convex combination coefficients of the $m$ archetypes for the $i$-th data point, with generic elements $\gamma_{ij}$, $j = 1, \ldots, m$,.
A Two-Step Procedure

Archetypal Analysis (Cutler and Breiman, 1994)

Formally:

- the archetypes $a_j$, $j = 1, \ldots, m$, are those points such that:

  $$x'_i = \gamma'_i A$$  \hspace{1cm} (1)

  where  \hspace{1cm} $\gamma'_{ij} \geq 0$ \hspace{0.5cm} $\forall i, j$; \hspace{0.5cm} $\gamma'_i 1 = 1$ \hspace{0.5cm} $\forall i$;

- archetypes must be also a mixture of the observed data:

  $$a'_j = \beta'_j X$$  \hspace{1cm} (2)

  where  \hspace{1cm} $\beta_{ji} \geq 0$ \hspace{0.5cm} $\forall j, i$; \hspace{0.5cm} $\beta'_j 1 = 1$ \hspace{0.5cm} $\forall j$;

- where

  - $X$ is the observed data matrix having generic row $x_i$;
  - coefficients $\beta_{ji}$ are the elements of the $\beta'_j$ vectors, i.e., the weights of the $n$ units;
  - $A$ is the archetype matrix with $a'_j$ its $j$-th row;
  - $\gamma'_i$ is the vector of the convex combination coefficients of the $m$ archetypes for the $i$-th data point, with generic elements $\gamma_{ij}$, $j = 1, \ldots, m$,.
Given $m$, the $m$ archetypes $\mathbf{A}(m) = \mathbf{a}_1, \ldots, \mathbf{a}_m$ are:

$$\mathbf{A}(m) = \arg \min_{\Gamma(m), \mathbf{B}(m)} \| \mathbf{X} - \Gamma(m) \mathbf{B}'(m) \mathbf{X} \|_F$$

(3)

$s, t$, all the conditions on coefficients $\beta_{ji}$ and $\gamma_{ij}$, and

$$\Gamma(m) = (\gamma_1, \ldots, \gamma_n)', \quad \Gamma(m) \in \mathbb{R}^{n \times m},$$

$$\mathbf{B}(m) = (\beta_1, \ldots, \beta_m), \quad \mathbf{B}(m) \in \mathbb{R}^{n \times m},$$

where $\| \mathbf{Y} \|_F = \sqrt{\text{Tr}(\mathbf{Y} \mathbf{Y}')} \,$ is the Frobenius norm for a generic matrix $\mathbf{Y}$,
A Two-Step Procedure

Archetypal Analysis (Cutler and Breiman, 1994)

Given $m$, the $m$ archetypes $A(m) = a_1, \ldots, a_m$ are:

$$A(m) = \arg \min_{\Gamma(m),B(m)} \|X - \Gamma(m)B'(m)X\|_F$$

(3)

$s,t,$ all the conditions on coefficients $\beta_{ji}$ and $\gamma_{ij}$, and

$$\Gamma(m) = (\gamma_1, \ldots, \gamma_n)', \quad \Gamma(m) \in \mathbb{R}^{n \times m},$$

$$B(m) = (\beta_1, \ldots, \beta_m), \quad B(m) \in \mathbb{R}^{n \times m},$$

where $\|Y\|_F = \sqrt{Tr(YY')}$ is the Frobenius norm for a generic matrix $Y$,
A Two-Step Procedure

Example: Wheat varieties

<table>
<thead>
<tr>
<th>Wheat variety</th>
<th>Humidity</th>
<th>Weight</th>
<th>Protein</th>
<th>Ash</th>
<th>Glutine</th>
<th>iGlut</th>
<th>iYell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadrato</td>
<td>11.60</td>
<td>82.00</td>
<td>12.00</td>
<td>1.96</td>
<td>10.15</td>
<td>93.00</td>
<td>24.55</td>
</tr>
<tr>
<td>Saragolla</td>
<td>12.00</td>
<td>80.00</td>
<td>13.35</td>
<td>2.15</td>
<td>9.95</td>
<td>86.50</td>
<td>24.15</td>
</tr>
<tr>
<td>Anco Marzio</td>
<td>12.30</td>
<td>83.50</td>
<td>12.25</td>
<td>1.76</td>
<td>10.20</td>
<td>82.00</td>
<td>22.35</td>
</tr>
<tr>
<td>Levante</td>
<td>12.15</td>
<td>82.00</td>
<td>12.55</td>
<td>2.08</td>
<td>10.70</td>
<td>82.00</td>
<td>25.55</td>
</tr>
<tr>
<td>Ciccio</td>
<td>11.65</td>
<td>84.50</td>
<td>11.95</td>
<td>1.93</td>
<td>8.40</td>
<td>81.00</td>
<td>22.45</td>
</tr>
<tr>
<td>Iride</td>
<td>12.25</td>
<td>80.50</td>
<td>12.35</td>
<td>2.00</td>
<td>9.35</td>
<td>79.00</td>
<td>23.60</td>
</tr>
<tr>
<td>Rusticano</td>
<td>12.75</td>
<td>83.00</td>
<td>13.05</td>
<td>1.97</td>
<td>10.95</td>
<td>79.00</td>
<td>22.75</td>
</tr>
<tr>
<td>San Carlo</td>
<td>12.85</td>
<td>80.50</td>
<td>13.15</td>
<td>1.99</td>
<td>10.70</td>
<td>78.50</td>
<td>24.30</td>
</tr>
<tr>
<td>Claudio</td>
<td>11.95</td>
<td>83.00</td>
<td>12.50</td>
<td>1.99</td>
<td>10.70</td>
<td>76.50</td>
<td>22.80</td>
</tr>
<tr>
<td>Simeto</td>
<td>11.80</td>
<td>83.50</td>
<td>10.75</td>
<td>1.84</td>
<td>8.85</td>
<td>73.50</td>
<td>23.60</td>
</tr>
<tr>
<td>Duilio</td>
<td>12.10</td>
<td>81.50</td>
<td>12.50</td>
<td>1.93</td>
<td>9.75</td>
<td>72.00</td>
<td>21.95</td>
</tr>
<tr>
<td>Svevo</td>
<td>11.90</td>
<td>80.50</td>
<td>13.55</td>
<td>2.00</td>
<td>10.95</td>
<td>71.50</td>
<td>26.05</td>
</tr>
<tr>
<td>Creso</td>
<td>12.20</td>
<td>83.00</td>
<td>12.40</td>
<td>1.86</td>
<td>11.05</td>
<td>69.00</td>
<td>21.35</td>
</tr>
<tr>
<td>Orobel</td>
<td>11.80</td>
<td>81.50</td>
<td>12.70</td>
<td>2.13</td>
<td>10.75</td>
<td>65.50</td>
<td>23.70</td>
</tr>
</tbody>
</table>
A Two-Step Procedure

Let us see AA at work

EXCEL
A Two-Step Procedure

Archetypal analysis: where and when

**Application of Archetypal Analysis for:**

- **detecting clusters**
  - cellular flames (Stone and Cutler, 1996; Stone, 2002)
  - galaxy spectra (Chan et al., 2003)
  - pattern recognition and image analysis (Marinetti, Finesso, Marsilio, 2006; 2007; Mørup and Hansen, 2010)

- **segmentation and fuzzy clustering**
  - marketing researches (Elder and Pinnel, 2003; Li et al., 2003; Riedesel, 2003)
  - sensory data analysis (D'Esposito, Palumbo, Ragozini; 2006, 2011)

- **performance analysis**
  - performing portfolio style analysis (Vistocco and Conversano 2008)
  - CPU performance analysis (Heavlin, 2008)
Outline

1 Framework: Prototypes
   - Notion
   - Definition
   - Identification

2 Our proposal

3 A Two-Step Procedure

4 Methodological Advances on AA

5 The study of uncertainty

6 Interval Valued Variables and Statistics

7 Archetypal Analysis for Interval Data

8 Archetypes and Prototypes

9 Final remarks

10 Main references
Methodological Advances on AA

AA extensions

Recently AA has been extended to other type of data:

Interval data: D’Esposito et al. (2006) and Corsaro and Marino (2010) have proposed a generalization of the AA algorithm to interval valued variables. More recently D’Esposito et al. have formalized the properties of the interval data archetypes (2012) and they have shown an application in sensorial analysis (2011).

Functional data: Cutler and Breiman (1994) have already introduced archetypal functions, using the Euclidean distance as metric to estimate the distance between curves, Costantini et al. (2012) propose to use basis functions to model the shape of smooth curve observed over time.
Recently AA has been extended to other type of data:

**Interval data:** D’Esposito et al. (2006) and Corsaro and Marino (2010) have proposed a generalization of the AA algorithm to interval valued variables. More recently D’Esposito et al. have formalized the properties of the interval data archetypes (2012) and they have shown an application in sensorial analysis (2011).

**Functional data:** Cutler and Breiman (1994) have already introduced archetypal functions, using the Euclidean distance as metric to estimate the distance between curves, Costantini et al. (2012) propose to use basis functions to model the shape of smooth curve observed over time.
Methodological Advances on AA

AA extensions

Recently AA has been extended to other type of data:

**Interval data:** D’Esposito et al. (2006) and Corsaro and Marino (2010) have proposed a generalization of the AA algorithm to interval valued variables. More recently D’Esposito et al.\(^1\) have formalized the properties of the interval data archetypes (2012) and they have shown an application in sensorial analysis (2011).

**Functional data:** Cutler and Breiman (1994) have already introduced archetypal functions, using the Euclidean distance as metric to estimate the distance between curves, Costantini et al. (2012) propose to use basis functions to model the shape of smooth curve observed over time.

---

The study of uncertainty

Outline

1 Framework: Prototypes
   - Notion
   - Definition
   - Identification

2 Our proposal

3 A Two-Step Procedure

4 Methodological Advances on AA

5 The study of uncertainty

6 Interval Valued Variables and Statistics

7 Archetypal Analysis for Interval Data

8 Archetypes and Prototypes

9 Final remarks

10 Main references
Are there other sources of uncertainty beyond randomness?

- The question divides scholars in two opposite formations. It involves the whole human knowledge process and goes further than Statistics itself.

  The debate is very long (and fascinating). It appears very unlikely that one formation could succeed in persuading the other one.

- Dennis V, Lindely (The philosophy of statistics, *The Statistician*, 2000) wrote: “... statistical issues concern uncertainty [...] , uncertainty can only be measured by probability”

- Twelve eminent Statisticians have discussed the Lindely’s paper. Some of them have touched the relationship between uncertainty and randomness,

  Among them, D,J, Hand replied “This is (*the probability*) certainly one of the most important aspects of statistics, perhaps the largest part, but it does not define it.” *(Discussion to Lindely’s paper, idem).*
The development of computer technology and information sciences have revealed the weakness of a coincidence between uncertainty and randomness.

When became clear that quantitative approaches could have been profitable for the soft sciences too, difficulties in coding uncertain information (the “real world”) in crisp data became evident at the same time.
The development of computer technology and information sciences have revealed the weakness of a coincidence between uncertainty and randomness.

When became clear that quantitative approaches could have been profitable for the soft sciences too, difficulties in coding uncertain information (the “real world”) in crisp data became evident at the same time.
Imprecision and Vagueness clearly represent two distinct and opposite facets of uncertainty (the lack of knowledge):

- **precision/accuracy**: to have exact measures of precise objects;
- **vagueness**: the need of having exact measures of vague concepts.
Imprecision and Vagueness clearly represent two distinct and opposite facets of uncertainty (the lack of knowledge):

**precision/accuracy:** to have exact measures of precise objects;

**vagueness:** the need of having exact measures of vague concepts.

Imprecision and inaccuracy are intrinsic to any measure, in most of cases they can be neglected and data are assumed to represent the exact variable value. Becoming the imprecision and inaccuracy a more significant part of the whole measure, they cannot be traced back to the phenomenon under investigation, they rather pertain to our measuring ability or capability and must be separately considered.
Imprecision and Vagueness clearly represent two distinct and opposite facets of uncertainty (the lack of knowledge):

**precision/accuracy:** to have exact measures of precise objects;

**vagueness:** the need of having exact measures of vague concepts.

Vagueness refers to the willing to measure (and then analyze) concepts representing a higher level of knowledge with respect to the classical statistical unit. Let us think to the study of the wheat varieties or to the species of dogs, just to take two simple examples.
The study of uncertainty

Imprecision and Vagueness

Imprecision and Vagueness clearly represent two distinct and opposite facets of uncertainty (the lack of knowledge):

- **precision/accuracy**: to have exact measures of precise objects;
- **vagueness**: the need of having exact measures of vague concepts.

Coding of Imprecision and Vagueness

*Interval data* [Moore, 1966] represent a suitable data coding to take into account imprecision and inaccuracy, *fuzzy sets* [Zadeh, 1965] can take into account the part of uncertainty due to the vagueness.
Precision

Precision is an indication of the uniformity or reproducibility of a result.

Same precision but different accuracy
Accuracy

Accuracy is the degree of conformity with a standard ("the truth").

Same accuracy but different precision
Interval Analysis has been developed to face the lack of accuracy problem in the numerical calculus with fixed-point CPU’s. Interval Analysis is based on the Interval Arithmetic (IA) and on the notation of interval value:

\[ x = [x, \overline{x}], \]
\[ x = \{ x \in \mathbb{R} | x \leq x \leq \overline{x} \}, \]

where \( x \) and \( \overline{x} \) are respectively called lower and upper bound of the closed (compact) and bounded set \( x \) [Kearfott, 1996]. Interval data are set-valued data defined by ordered couples of values: \([x, \overline{x}]\).

The textbook of [Neumaier, 1990] is good reference to approach the IA.
Interval Analysis was developed to cope with the *round-off* problem of digital computing, when the CPU’s were based on the *fixed-point* architecture,

Interval Analysis is founded on the Interval Arithmetics (IA) that generalizes the classical arithmetics to the *interval data*,

IA did not furnish a more precise result; however, *interval-valued* results expressed the precision order.

Many methods were developed to solve systems of linear equations [Alefeld and Herzerberger, 1983, Hickey et al., 2001]

These methods are extremely favorable to treat intervals having spreads relatively “small” with respect to the central value
Given $x = [x, \bar{x}]$ and $y = [y, \bar{y}]$, the result of $x \diamond y$ is again an interval $z$ with property:

$$x \diamond y = z = \{ z = x \diamond y | x \in x, y \in y \}$$

(4)

where $\diamond$ belongs to the set $\{+, -, \times, \div\}$.

Arithmetic operations on intervals are expressed in terms of ordinary arithmetics on their bounds as follows:

$$x + y = [x + y, \bar{x} + \bar{y}]$$

$$x - y = [x - \bar{y}, \bar{x} - y]$$

$$x \times y = [\min([x \times y, x \times \bar{y}, \bar{x} \times y, \bar{x} \times \bar{y}]), \max([x \times y, x \times \bar{y}, \bar{x} \times y, \bar{x} \times \bar{y}])]$$

$$x \div y = [\min([x \div y, x \div \bar{y}, \bar{x} \div y, \bar{x} \div \bar{y}]), \max([x \div y, x \div \bar{y}, \bar{x} \div y, \bar{x} \div \bar{y}])$$

with the extra condition $0 \notin y$ for the division,
Interval Valued Variables and Statistics

Midpoint/Range notation

The set of all intervals is usually indicated as $\mathbb{IR}$: $x \in \mathbb{IR}$,
Intervals can also be expressed in the midpoint and range notation,
Given a generic interval $x \in \mathbb{IR}$, the quantities midpoint and range are respectively defined as:

$$\hat{x} = \frac{1}{2}(x + \bar{x}),$$
$$\Delta x = \frac{1}{2}(\bar{x} - x),$$

The interval $x$ can also be written as:

$$x = [x, \bar{x}] = [\hat{x} - \Delta x, \hat{x} + \Delta x] = \{\hat{x}, \Delta x\},$$
Dealing with intervals

- The Hausdorff distance [Braun et al., 2003] has a central role in the Statistical Analysis of interval data;
- However, many definitions that are considered obvious in the case of dimensionless points are no longer valid dealing with intervals,
Dealing with intervals

- The Hausdorff distance [Braun et al., 2003] has a central role in the Statistical Analysis of interval data;
- However, many definitions that are considered obvious in the case of dimensionless points are no longer valid dealing with intervals,

A point is a point in the $\mathbb{R}^p$, $\forall p \geq 1$, 
Dealing with intervals

The Hausdorff distance [Braun et al., 2003] has a central role in the Statistical Analysis of interval data;

However, many definitions that are considered obvious in the case of dimensionless points are no longer valid dealing with intervals,

An interval in \( \mathbb{R}^p, \forall p \geq 1 \), varies according to \( p \): line for \( p = 1 \), rectangle for \( p = 2 \), parallelootope for \( p \geq 3 \),
The Hausdorff Distance in \( \mathbb{R} \)

In the special case of \( \mathbb{R} \), the Hausdorff distance between two generic intervals is given by:

\[
H(A, B) = \max\{|\bar{a} - \bar{b}|, |a - b|\} = |\bar{a} - \bar{b}| + |\Delta a - \Delta b|,
\]

It is easy to verify that:

- \( H(A, B) \geq 0 \),
- \( H(A, B) = H(B, A) \),
- \( H(A, C) \leq H(A, B) + H(B, C) \), where \( C \) is a generic compact subset in \( \mathbb{R} \),
Interval Valued Variables and Statistics

The Hausdorff Distance in $\mathbb{R}$

In the special case of $\mathbb{R}$, the Hausdorff distance between two generic intervals is given by:

$$H(A, B) = \max\{ |\bar{a} - \bar{b}|, |a - b| \} = |\bar{a} - \bar{b}| + |\Delta a - \Delta b|,$$

It is easy to verify that:

- $H(A, B) \geq 0$,
- $H(A, B) = H(B, A)$,
- $H(A, C) \leq H(A, B) + H(B, C)$, where $C$ is a generic compact subset in $\mathbb{R}$,

Example:

$A = [0, 3], B = [7, 11]$

$$H(A, B) = \max\{(7 - 0); (11 - 3)\} = |9 - 1, 5| + |2 - 1, 5| = 8$$
The generalisation of the Hausdorff distance in $\mathbb{R}^p$ is NP-hard complete problem. Under special restrictions there are some satisfactory approximations.

Given two parallelotopes $\{A, B\}$ in $\mathbb{R}^p$, the quantity:

$$H(A, B) = \left\{ \sum_{j=1}^{p} |H(A_j, B_j)|^{\alpha} \right\}^{\frac{1}{\alpha}} \geq 0,$$

for any $\alpha \geq 1$, is a metric.

The following properties hold in $\mathbb{R}^p$, $\forall p \geq 1$:

i) $H(A, A) = 0 \iff A = A$, $\forall A$, being $H(A_j, A_j) = 0, \forall j = 1, \ldots, p$,

ii) $H(A, B) = H(B, A)$ (Symmetry)

iii) $H(A, B) + H(A, C) \geq H(B, C)$ (Triangular inequality)
The Hausdorff Distance in $\mathbb{R}^p$

The generalisation of the Hausdorff distance in $\mathbb{R}^p$ is $NP$-hard complete problem, Under special restrictions there are some satisfactory approximations,

Given two parallelotopes $\{A, B\}$ in $\mathbb{R}^p$, the quantity:

$$H(A, B) = \left\{ \sum_{j=1}^{p} |H(A_j, B_j)|^\alpha \right\}^{\frac{1}{\alpha}} \geq 0,$$

for any $\alpha \geq 1$, is a metric,

The following properties hold in $\mathbb{R}^p$, $\forall p \geq 1$:

i) $H(A, A) = 0 \iff A = A$, $\forall A$, being $H(A_j, A_j) = 0$, $\forall j = 1, \ldots, p$,

ii) $H(A, B) = H(B, A)$ (Symmetry)

iii) $H(A, B) + H(A, C) \geq H(B, C)$ (Triangular inequality)
The Hausdorff Distance in $\mathbb{R}^p$

The distance $H(A, B)$ in $\mathbb{R}^p$ introduced, for $\alpha = 2$ can also be expressed in terms of midpoints and ranges:

$H(A, B) = \sqrt{\sum_{j=1}^{p} [(\bar{a}_j - \bar{b}_j)^2 + (\Delta a_j - \Delta b_j)^2 + 2 |\bar{a}_j - \bar{b}_j| |\Delta a_j - \Delta b_j|]},$
Interval Valued Variables and Statistics

Digression: Hausdorff distance between hyperspheres

On this slide capital letters \{A, B, \ldots\} indicate spheres in the \(\mathbb{R}^p\) space; the general sphere \(A\) in \(\mathbb{R}^p\) has center in \(\tilde{A} = [\tilde{a}_j] (j = 1, \ldots, p)\) and radius \(r_A \geq 0\).

**Theorem (Palumbo and Irpino 2005)**

Let \{\(A, B\)\} be two spheres in the \(\mathbb{R}^p\) space, the Hausdorff distance between \(A\) and \(B\) is given by:

\[
H(A, B) = \sqrt{\sum_{j=1}^{p} (\tilde{a}_j - \tilde{b}_j)^2} + |r_A - r_B|
\]

(5)

Where \(\tilde{a}_j\) and \(\tilde{b}_j\) indicate the generic centers coordinates of \(A\) and \(B\), and \(r_A\) and \(r_B\) are the respective radii,
Interval Valued Variables and Statistics

Hausdorff distance as measure of variability

The Hausdorff Distance for multivariate intervals

It can be also proved that the sum of the Hausdorff distances between two parallelotopes in $\mathbb{R}^p$ is a distance [Chavent, 2004, de Carvalho et al., 2006],

\[ H(x_i, x'_i) = \sum_{k=1}^{p} \max \{ |\bar{x}_{ik} - \bar{x}_{i'k}|, |\bar{x}_{ik} - \bar{x}_{i'k}| \} \]

\[ = \sum_{k=1}^{p} (|\bar{x}_{ik} - \bar{x}_{i'k}| + |\Delta x_{ik} - \Delta x_{i'k}|), \]

[Chavent, 2004]
Interval Valued Variables and Statistics

Variability

Let \( \{ \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i \} \) be a set of \( N \) of \( p \) dimensional interval-valued statistical units, the distance between \( \mathbf{x}_i \) and \( \mathbf{x}_i' \) is defined by:

\[
d(\mathbf{x}_i, \mathbf{x}_i') = \sqrt{\sum_{j=1}^{p} d^2(x_{ij}, x_{ij}')},
\]

The centered interval-valued data matrix \( \mathbf{Y} = \mathbf{X} - \mathbf{u}\bar{x}' \), has \( N \) rows and \( p \) columns, and \( \bar{x}' \) indicates the mean vector (\( \mathbf{u} \) is the unitary vector of \( N \) terms), and with \( j \in (1, \ldots, p) \), The matrix \( \mathbf{Y} \) can be also written in midpoints and ranges notation: \( \mathbf{Y} \equiv \{ \tilde{\mathbf{Y}}, (\Delta \mathbf{Y}) \} \).

The index of variability \( v^2 \) ([Palumbo and Lauro, 2003])

The generic diagonal term \( v_{jj} \) (\( 1 \leq j \leq p \)) of the \( p \times p \) square symmetric matrix \( \mathbf{V} \) is an absolute index of variability for interval valued variables, where \( \mathbf{V} \) is given by:

\[
\mathbf{V}_\mathbf{Y} = \frac{1}{N} \left\{ \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} + (\Delta \mathbf{Y})'(\Delta \mathbf{Y}) + [ | \tilde{\mathbf{Y}}'(\Delta \mathbf{Y}) | + | (\Delta \mathbf{Y})'\tilde{\mathbf{Y}} | ] \right\},
\]
Outline

1. Framework: Prototypes
   - Notion
   - Definition
   - Identification

2. Our proposal

3. A Two-Step Procedure

4. Methodological Advances on AA

5. The study of uncertainty

6. Interval Valued Variables and Statistics

7. Archetypal Analysis for Interval Data

8. Archetypes and Prototypes

9. Final remarks

10. Main references
In analogy with the single value case, we define the archetypes $\mathbf{A}$ for interval valued symbolic data [D’Esposito et al., 2006, D’Esposito et al., 2011].

- Considering the midpoint and range spaces two sets of archetypes, $\mathbf{A}^c$ and $\mathbf{A}^r$ are defined.
- Each data should be expressed as a unique convex combination of the interval data archetype in terms of midpoints and ranges.
- Therefore the mixture coefficients $\gamma'_i$ are imposed to be the same in the two spaces.
- Hence the $\gamma'_i$ coefficients represent the algebraic linkage of the two optimizations, and hence the linkage between the two spaces.
- Optimisation procedure to derive interval archetypes is based on the use Hausdorff distance and Frobenius norm [Corsaro and Marino, 2010, D’Esposito et al., 2006].
Archetypal Analysis for Interval Data

Archetypes for Interval Data

Given the metric space provided by the Frobenius norm and the distance between interval matrices, for each $m$, the $m$ interval valued archetypes $A(m)$ can be determined by minimizing

$$X - \tilde{X}(m)$$

$$\tilde{X}(m) = \Gamma(m)A(m), \tilde{X}(m) \in \mathbb{IR}^{n \times p}, \text{i.e. the data matrix reconstructed by } m \text{ archetypal hyper-rectangle.}$$

Thus, given $m$ and the quantity:

$$RSS(m) = \|d(X, \tilde{X}(m))\|_F = \|X - (\Gamma(m)B(m)X)\|_F.$$ 

the $m$ archetypes solve the minimization problem:

$$\min_{\Gamma(m),B(m)} RSS(m) = \min_{\Gamma(m),B(m)} \|X - (\Gamma(m)B(m)X)\|_F.$$
## Example: Wheat varieties

Interval valued data

<table>
<thead>
<tr>
<th>Wheat variety</th>
<th>Humidity</th>
<th>Weight</th>
<th>Protein</th>
<th>Ash</th>
<th>Glutine</th>
<th>iGlut</th>
<th>iYell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anco Marzio</td>
<td>11.2 - 13.4</td>
<td>82 - 85</td>
<td>11.0 - 13.5</td>
<td>1.51 - 2.01</td>
<td>9.0 - 11.4</td>
<td>72 - 92</td>
<td>20.5 - 24.2</td>
</tr>
<tr>
<td>Ciccio</td>
<td>11.2 - 12.1</td>
<td>83 - 86</td>
<td>9.7 - 14.2</td>
<td>1.80 - 2.06</td>
<td>6.3 - 10.5</td>
<td>67 - 95</td>
<td>20.6 - 24.3</td>
</tr>
<tr>
<td>Claudio</td>
<td>11.1 - 12.8</td>
<td>81 - 85</td>
<td>10.8 - 14.2</td>
<td>1.82 - 2.16</td>
<td>8.6 - 12.8</td>
<td>62 - 91</td>
<td>21.5 - 24.1</td>
</tr>
<tr>
<td>Creso</td>
<td>11.0 - 13.4</td>
<td>81 - 85</td>
<td>10.8 - 14.0</td>
<td>1.72 - 2.00</td>
<td>8.9 - 13.2</td>
<td>51 - 87</td>
<td>18.5 - 24.2</td>
</tr>
<tr>
<td>Duilio</td>
<td>11.2 - 13.0</td>
<td>79 - 84</td>
<td>9.8 - 15.2</td>
<td>1.77 - 2.09</td>
<td>8.1 - 11.4</td>
<td>60 - 84</td>
<td>21.2 - 22.7</td>
</tr>
<tr>
<td>Iride</td>
<td>11.3 - 13.2</td>
<td>77 - 84</td>
<td>10.5 - 14.2</td>
<td>1.79 - 2.20</td>
<td>7.3 - 11.4</td>
<td>65 - 93</td>
<td>22.0 - 25.2</td>
</tr>
<tr>
<td>Levante</td>
<td>11.3 - 13.0</td>
<td>79 - 85</td>
<td>11.4 - 13.7</td>
<td>1.87 - 2.28</td>
<td>9.4 - 12.0</td>
<td>71 - 93</td>
<td>23.9 - 27.2</td>
</tr>
<tr>
<td>Orobel</td>
<td>10.9 - 12.7</td>
<td>79 - 84</td>
<td>11.7 - 13.7</td>
<td>2.05 - 2.20</td>
<td>9.2 - 12.3</td>
<td>42 - 89</td>
<td>20.7 - 26.7</td>
</tr>
<tr>
<td>Quadrato</td>
<td>11.5 - 11.7</td>
<td>81 - 83</td>
<td>11.5 - 12.5</td>
<td>1.80 - 2.12</td>
<td>9.5 - 10.8</td>
<td>90 - 96</td>
<td>23.8 - 25.3</td>
</tr>
<tr>
<td>Rusticano</td>
<td>11.9 - 13.6</td>
<td>81 - 85</td>
<td>11.6 - 14.5</td>
<td>1.86 - 2.07</td>
<td>10.0 - 11.9</td>
<td>62 - 96</td>
<td>21.5 - 24.0</td>
</tr>
<tr>
<td>San Carlo</td>
<td>11.7 - 14.0</td>
<td>77 - 84</td>
<td>11.9 - 14.4</td>
<td>1.77 - 2.20</td>
<td>8.4 - 13.0</td>
<td>62 - 95</td>
<td>22.8 - 25.8</td>
</tr>
<tr>
<td>Saragolla</td>
<td>11.2 - 12.8</td>
<td>78 - 82</td>
<td>11.1 - 15.6</td>
<td>2.08 - 2.21</td>
<td>7.5 - 12.4</td>
<td>75 - 98</td>
<td>21.7 - 26.6</td>
</tr>
<tr>
<td>Simeto</td>
<td>10.2 - 13.4</td>
<td>81 - 86</td>
<td>8.7 - 12.8</td>
<td>1.67 - 2.00</td>
<td>6.4 - 11.3</td>
<td>52 - 95</td>
<td>21.5 - 25.7</td>
</tr>
</tbody>
</table>
Data represent the average RSS value reached over 20 random start solutions and the average CPU time (mill-seconds on this notebook), for any given $m$. Bubble radii refer the RSS Standard Deviation.
How many archetypes?

Data represent the best RSS value reached over 20 random start solutions and the corresponding CPU time (mill-seconds on this notebook), for any given $m$. 
## Weighting coefficients

Weighting coefficient matrix $\Gamma(m)$ for $m = 4$,

<table>
<thead>
<tr>
<th>Wheat</th>
<th>$\gamma_1(4)$</th>
<th>$\gamma_2(4)$</th>
<th>$\gamma_3(4)$</th>
<th>$\gamma_4(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anco Marzio</td>
<td>0.000</td>
<td>0.528</td>
<td>0.213</td>
<td>0.259</td>
</tr>
<tr>
<td>Ciccio</td>
<td>0.000</td>
<td>0.000</td>
<td>0.828</td>
<td>0.172</td>
</tr>
<tr>
<td>Claudio</td>
<td>0.167</td>
<td>0.599</td>
<td>0.018</td>
<td>0.217</td>
</tr>
<tr>
<td>Creso</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duilio</td>
<td>0.062</td>
<td>0.841</td>
<td>0.000</td>
<td>0.097</td>
</tr>
<tr>
<td>Iride</td>
<td>0.670</td>
<td>0.000</td>
<td>0.272</td>
<td>0.058</td>
</tr>
<tr>
<td>Levante</td>
<td>0.687</td>
<td>0.000</td>
<td>0.000</td>
<td>0.313</td>
</tr>
<tr>
<td>Orobel</td>
<td>0.659</td>
<td>0.341</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Quadrato</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rusticano</td>
<td>0.144</td>
<td>0.622</td>
<td>0.000</td>
<td>0.234</td>
</tr>
<tr>
<td>San Carlo</td>
<td>0.849</td>
<td>0.151</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Saragolla</td>
<td>0.910</td>
<td>0.000</td>
<td>0.000</td>
<td>0.090</td>
</tr>
<tr>
<td>Simeto</td>
<td>0.013</td>
<td>0.072</td>
<td>0.916</td>
<td>0.000</td>
</tr>
<tr>
<td>Svevo</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Weighting coefficients
Archetypal Analysis for Interval Data

Clusters in the data

Weighting coefficient matrix $\Gamma(m)$ for $m = 4$,

<table>
<thead>
<tr>
<th>Wheat</th>
<th>$\gamma_1(4)$</th>
<th>$\gamma_2(4)$</th>
<th>$\gamma_3(4)$</th>
<th>$\gamma_4(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anco Marzio</td>
<td>0.000</td>
<td>0.528</td>
<td>0.213</td>
<td>0.259</td>
</tr>
<tr>
<td>Ciccio</td>
<td>0.000</td>
<td>0.000</td>
<td>0.828</td>
<td>0.172</td>
</tr>
<tr>
<td>Claudio</td>
<td>0.167</td>
<td>0.599</td>
<td>0.018</td>
<td>0.217</td>
</tr>
<tr>
<td>Creso</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Duilio</td>
<td>0.062</td>
<td>0.841</td>
<td>0.000</td>
<td>0.097</td>
</tr>
<tr>
<td>Iride</td>
<td>0.670</td>
<td>0.000</td>
<td>0.272</td>
<td>0.058</td>
</tr>
<tr>
<td>Levante</td>
<td>0.687</td>
<td>0.000</td>
<td>0.000</td>
<td>0.313</td>
</tr>
<tr>
<td>Orobel</td>
<td>0.659</td>
<td>0.341</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Quadrato</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Rusticano</td>
<td>0.144</td>
<td>0.622</td>
<td>0.000</td>
<td>0.234</td>
</tr>
<tr>
<td>San Carlo</td>
<td>0.849</td>
<td>0.151</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Saragolla</td>
<td>0.910</td>
<td>0.000</td>
<td>0.000</td>
<td>0.090</td>
</tr>
<tr>
<td>Simeto</td>
<td>0.013</td>
<td>0.072</td>
<td>0.916</td>
<td>0.000</td>
</tr>
<tr>
<td>Svevo</td>
<td><strong>1.000</strong></td>
<td><strong>0.000</strong></td>
<td><strong>0.000</strong></td>
<td><strong>0.000</strong></td>
</tr>
</tbody>
</table>
Archetypal Analysis for Interval Data

Clusters in the data

Weighting coefficient matrix $\Gamma(m)$ for $m = 4$,

<table>
<thead>
<tr>
<th>Wheat</th>
<th>$\gamma_1(4)$</th>
<th>$\gamma_2(4)$</th>
<th>$\gamma_3(4)$</th>
<th>$\gamma_4(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anco Marzio</td>
<td>0.000</td>
<td>0.528</td>
<td>0.213</td>
<td>0.259</td>
</tr>
<tr>
<td>Ciccio</td>
<td>0.000</td>
<td>0.000</td>
<td>0.828</td>
<td>0.172</td>
</tr>
<tr>
<td>Claudio</td>
<td>0.167</td>
<td>0.599</td>
<td>0.018</td>
<td>0.217</td>
</tr>
<tr>
<td>Creso</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Duilio</td>
<td>0.062</td>
<td>0.841</td>
<td>0.000</td>
<td>0.097</td>
</tr>
<tr>
<td>Iride</td>
<td>0.670</td>
<td>0.000</td>
<td>0.272</td>
<td>0.058</td>
</tr>
<tr>
<td>Levante</td>
<td>0.687</td>
<td>0.000</td>
<td>0.000</td>
<td>0.313</td>
</tr>
<tr>
<td>Orobel</td>
<td>0.659</td>
<td>0.341</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Quadrato</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Rusticano</td>
<td>0.144</td>
<td>0.622</td>
<td>0.000</td>
<td>0.234</td>
</tr>
<tr>
<td>San Carlo</td>
<td>0.849</td>
<td>0.151</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Saragolla</td>
<td>0.910</td>
<td>0.000</td>
<td>0.000</td>
<td>0.090</td>
</tr>
<tr>
<td>Simeto</td>
<td>0.013</td>
<td>0.072</td>
<td>0.916</td>
<td>0.000</td>
</tr>
<tr>
<td>Svevo</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Clusters in the data

Weighting coefficient matrix $\Gamma(m)$ for $m = 4$,

<table>
<thead>
<tr>
<th>Wheat</th>
<th>$\gamma_1(4)$</th>
<th>$\gamma_2(4)$</th>
<th>$\gamma_3(4)$</th>
<th>$\gamma_4(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anco Marzio</td>
<td>0.000</td>
<td>0.528</td>
<td>0.213</td>
<td>0.259</td>
</tr>
<tr>
<td>Ciccio</td>
<td>0.000</td>
<td>0.000</td>
<td>0.828</td>
<td>0.172</td>
</tr>
<tr>
<td>Claudio</td>
<td>0.167</td>
<td>0.599</td>
<td>0.018</td>
<td>0.217</td>
</tr>
<tr>
<td>Creso</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Duilio</td>
<td>0.062</td>
<td>0.841</td>
<td>0.000</td>
<td>0.097</td>
</tr>
<tr>
<td>Iride</td>
<td>0.670</td>
<td>0.000</td>
<td>0.272</td>
<td>0.058</td>
</tr>
<tr>
<td>Levante</td>
<td>0.687</td>
<td>0.000</td>
<td>0.000</td>
<td>0.313</td>
</tr>
<tr>
<td>Orobel</td>
<td>0.659</td>
<td>0.341</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Quadrato</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Rusticano</td>
<td>0.144</td>
<td>0.622</td>
<td>0.000</td>
<td>0.234</td>
</tr>
<tr>
<td>San Carlo</td>
<td>0.849</td>
<td>0.151</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Saragolla</td>
<td>0.910</td>
<td>0.000</td>
<td>0.000</td>
<td>0.090</td>
</tr>
<tr>
<td>Simeto</td>
<td>0.013</td>
<td>0.072</td>
<td>0.916</td>
<td>0.000</td>
</tr>
<tr>
<td>Svevo</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Archetypal Analysis for Interval Data

Clusters in the data

Weighting coefficient matrix $\Gamma(m)$ for $m = 4$,

<table>
<thead>
<tr>
<th>Wheat</th>
<th>$\gamma_1(4)$</th>
<th>$\gamma_2(4)$</th>
<th>$\gamma_3(4)$</th>
<th>$\gamma_4(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anco Marzio</td>
<td>0.000</td>
<td>0.528</td>
<td>0.213</td>
<td>0.259</td>
</tr>
<tr>
<td>Ciccio</td>
<td>0.000</td>
<td>0.000</td>
<td>0.828</td>
<td>0.172</td>
</tr>
<tr>
<td>Claudio</td>
<td>0.167</td>
<td>0.599</td>
<td>0.018</td>
<td>0.217</td>
</tr>
<tr>
<td>Creso</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Duilio</td>
<td>0.062</td>
<td>0.841</td>
<td>0.000</td>
<td>0.097</td>
</tr>
<tr>
<td>Iride</td>
<td>0.670</td>
<td>0.000</td>
<td>0.272</td>
<td>0.058</td>
</tr>
<tr>
<td>Levante</td>
<td>0.687</td>
<td>0.000</td>
<td>0.000</td>
<td>0.313</td>
</tr>
<tr>
<td>Orobel</td>
<td>0.659</td>
<td>0.341</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Quadrato</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Rusticano</td>
<td>0.144</td>
<td>0.622</td>
<td>0.000</td>
<td>0.234</td>
</tr>
<tr>
<td>San Carlo</td>
<td>0.849</td>
<td>0.151</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Saragolla</td>
<td>0.910</td>
<td>0.000</td>
<td>0.000</td>
<td>0.090</td>
</tr>
<tr>
<td>Simeto</td>
<td>0.013</td>
<td>0.072</td>
<td>0.916</td>
<td>0.000</td>
</tr>
<tr>
<td>Svevo</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Archetypal Analysis for Interval Data

**MR-ACP for interval valued data** (Palumbo and Lauro, 2003)

Four archetypes (in yellow) are represented as supplementary rectangles.
In general, necessary archetypes are a small number: between 3 and 6, according to our experience. The number of variables can be high. Chernoff’s faces in the display represent our interval archetypes, where ranges are related to the eyes and centers to the face.
Archetypes and Prototypes

Outline

1. Framework: Prototypes
   - Notion
   - Definition
   - Identification

2. Our proposal

3. A Two-Step Procedure

4. Methodological Advances on AA

5. The study of uncertainty

6. Interval Valued Variables and Statistics

7. Archetypal Analysis for Interval Data

8. Archetypes and Prototypes

9. Final remarks

10. Main references
The proposed procedure

The archetypes can be interpreted as prototypes, however they could be too extreme with respect to the clusters they represent. The following procedure helps to find more 'internal' prototypes.

- Evaluate $K$ the archetypes ($A(K)$ or $A(K')$);
- Represent data in the space spanned by the archetypes through the coefficient $\gamma$;
- Detect clusters and data structure through the $\gamma$ coefficients;
- Find the prototypes $P_h$ solving a minimization problem based on an appropriate distance function in this space;
- Revert to the original space ($\mathbb{R}^p$ or $\mathbb{R}^p$) of the data to determine the prototypes $p_k$. 
The distance function

\( \gamma \) coefficient as compositional data

- Note that the \( \gamma \) coefficient are compositional data by their definition.
- Compositional data [Aitchison, 1982] consist of vectors of positive numbers summing to a unit, or in general to some fixed constant for all vectors. These vectors span a simplex, defined as:

\[
S^p = \left\{ \mathbf{x} = [x_1, \ldots, x_p] \in \mathbb{R}^p \mid x_i > 0, i = 1, \ldots, p; \sum_i x_i = 1 \right\}
\]

- Given two compositions \( \mathbf{x}_i \in S^p \) and \( \mathbf{x}_{i'} \in S^p \) a proper distance function is defined as:

\[
\left[ \sum_{j=1}^{p} \left( \log \frac{x_{ij}}{g(x_i)} - \log \frac{x_{i'j}}{g(x_{i'})} \right)^2 \right]^{1/2}
\]

with \( g(x_i) = \left( \prod_{j=1}^{p} x_{ij} \right)^{1/p} \) [Aitchison et al., 2000].
Recall

Given a partition \((C_1, \ldots, C_K)\) of \(\Omega\) in \(K\) clusters of sizes \(n_1, \ldots, n_K\) and a dissimilarity measure \(d(\cdot, \cdot)\), a prototype \(p_h, h = 1, \ldots, K\) minimizes the function \(\sum_{i \in C_h} d(x_i, p_h)\).

In our case

- \(d(\cdot, \cdot)\) is the compositional distance in the space spanned by the archetypes \((S^K)\).
- The prototype in the \(\gamma\) space \(\mathcal{P}_h = (\mathcal{P}_{h1}, \ldots, \mathcal{P}_{hK})\) which solves the minimization is the **compositional geometric mean**:

\[
\mathcal{P}_h = \left( \frac{g_{h1}}{\sum_{j=1}^{K} g_{hj}}, \ldots, \frac{g_{hK}}{\sum_{j=1}^{K} g_{hj}} \right)
\]

where \(g_{hj} = \left( \prod_{i \in C_k} \gamma_{ij} \right)^{1/n_h} \) [Martýn-Fernandez et al., 1998].
Remember that:

\[ x'_i = \gamma'_i A \]
\[ x'_i = \gamma'_i A \]

The \( P_h \)'s are themselves prototypes in the space spanned by the archetypes, but are also the barycentric coordinate of the prototypes \( p_h \) with respect to the archetypes. The prototypes \( p_h \) in the original space (\( R^p \) or \( \mathbb{R}^p \)) will be then:

\[ p'_h = P'_h A = P'_h B X \]
\[ p'_h = P'_h A = P'_h B X \]

Note that the prototypes are function of the archetypes and are also a combination of the observed data.
Outline

1 Framework: Prototypes
   - Notion
   - Definition
   - Identification

2 Our proposal

3 A Two-Step Procedure

4 Methodological Advances on AA

5 The study of uncertainty

6 Interval Valued Variables and Statistics

7 Archetypal Analysis for Interval Data

8 Archetypes and Prototypes

9 Final remarks

10 Main references
Some Remarks:

1. Archetypes can identify well separated prototypes (they could be too extreme in some cases!);
2. Combining soft clustering procedure and compositional data analysis offers a good alternative to classical clustering procedures;
3. AA potentially works on any data types;
4. AA potentially works with any distance measure;
Some Remarks:

1. Archetypes can identify well separated prototypes (they could be too extreme in some cases!);
2. Combining soft clustering procedure and compositional data analysis offers a good alternative to classical clustering procedures;
3. AA potentially works on any data types;
4. AA potentially works with any distance measure;
Some Remarks:

1. Archetypes can identify well separated prototypes (they could be too extreme in some cases!);
2. Combining soft clustering procedure and compositional data analysis offers a good alternative to classical clustering procedures;
3. AA potentially works on any data types;
4. AA potentially works with any distance measure;
Final remarks

Some Remarks:

1. Archetypes can identify well separated prototypes (they could be too extreme in some cases!);
2. Combining soft clustering procedure and compositional data analysis offers a good alternative to classical clustering procedures;
3. AA potentially works on any data types;
4. AA potentially works with any distance measure;
Final remarks

Working group

M.R. D’Esposito  University of Salerno
M. Marino  University of Naples Federico II
F. Palumbo  University of Naples Federico II
G. C. Porzio  University of Cassino
G. Ragozini  University of Naples Federico II
Outline

1. Framework: Prototypes
   - Notion
   - Definition
   - Identification

2. Our proposal

3. A Two-Step Procedure

4. Methodological Advances on AA

5. The study of uncertainty

6. Interval Valued Variables and Statistics

7. Archetypal Analysis for Interval Data

8. Archetypes and Prototypes

9. Final remarks

10. Main references
The statistical analysis of compositional.

Logratio analysis and compositional distance.
Mathematical Geology, 32(271-275).

Introduction to Interval computation.

The geometry of the Hausdorff metric.
Available at:

A hausdorff distance between hyper-rectangles for clustering interval data.
10.1007/978-3-642-17103-1_32.

Archetypal analysis of interval data.

Adaptive hausdorff distances and dynamic clustering of symbolic interval data.
Archetypal analysis for interval data in marketing research.

Interval archetypes: a new tool for interval data analysis.
*Statistical Analysis and Data Mining*.

Interval arithmetic: From principles to implementation.

*Rigorous Global Search: Continuous Problems*.
Kluver, Dordrecht.

Measures of difference for compositional data and hierarchical clustering methods.
In *IAMG*, pages 526–531.

*Interval Analysis*.
Prentice-Hall, Englewood Cliffs, NJ.

*Interval methods for systems of Equations*.

A PCA for interval valued data based on midpoints and radii.