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

Francesco Palumbo

Università degli Studi di Napoli Federico II

Outline

Outline

Framework

General notion

- 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)

Framework

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

Prototype Identification

Formally

Given

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

The prototipicality index is a function ϕ combing this two measures:

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

A prototype for a group C_k is the point that maximizes $T(\mathbf{x}, C_k)$

Prototypes for classification and clustering

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.
  - $\hookrightarrow$ internal resemblance and external dissimilarity match the classification and clustering objectives (homogeneity and separability).
- Δ and P are often the average (Lesot and Kruse, 2006)
  - $\hookrightarrow$ 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

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

Outline

Prototypes Identification

Our proposal

We propose a two-step procedure

1. **Maximize the external dissimilarity**
   - ↪ find the archetypes to exploit their properties (pureness, separabilty, 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 (Cutler and Breiman, 1994)

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)

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)

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 $\mathbf{a}_j$, $j = 1, \dots, m$, are those points such that:

$$\boxed{\mathbf{x}_i' = \boldsymbol{\gamma}_i' \mathbf{A}} \quad (1)$$

with $\gamma_{ij} \geq 0 \quad \forall i, j; \qquad \boldsymbol{\gamma}_i' \mathbf{1} = 1 \quad \forall i,$

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

$$\boxed{\mathbf{a}_j' = \boldsymbol{\beta}_j' \mathbf{X}} \quad (2)$$

with $\beta_{ji} \geq 0 \quad \forall j, i; \qquad \boldsymbol{\beta}_j' \mathbf{1} = 1 \quad \forall j,$

- where
  - $\mathbf{X}$ is the observed data matrix having generic row $\mathbf{x}_i$;
  - coefficients β_{ji} are the elements of the $\boldsymbol{\beta}_j'$ vectors, i,e, the weights of the n units;
  - $\mathbf{A}$ is the archetype matrix with $\mathbf{a}_j'$ its j-th row;
  - $\boldsymbol{\gamma}_i'$ is the vector of the convex combination coefficients of the m archetypes for the i-th data point, with generic elements γ_{ij}, $j = 1, \dots, m$,

Archetypal Analysis (Cutler and Breiman, 1994)

Formally:

- the archetypes $\mathbf{a}_j$, $j = 1, \ldots, m$, are those points such that:

$$\boxed{\mathbf{x}_i' = \boldsymbol{\gamma}_i' \mathbf{A}} \tag{1}$$

with $\gamma_{ij} \geq 0 \quad \forall i, j; \qquad \boldsymbol{\gamma}_i' \mathbf{1} = 1 \quad \forall i,$

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

$$\boxed{\mathbf{a}_j' = \boldsymbol{\beta}_j' \mathbf{X}} \tag{2}$$

with $\beta_{ji} \geq 0 \quad \forall j, i; \qquad \boldsymbol{\beta}_j' \mathbf{1} = 1 \quad \forall j,$

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

Archetypal Analysis (Cutler and Breiman, 1994)

Given m, the m archetypes $\mathbf{A}(m) = \mathbf{a}_1, \ldots \mathbf{a}_m$ are:

$$\mathbf{A}(m) = \arg \min_{\mathbf{\Gamma}(m), \mathbf{B}(m)} \|\mathbf{X} - \mathbf{\Gamma}(m)\mathbf{B}'(m)\mathbf{X}\|_F \qquad (3)$$

s,t, all the conditions on coefficients β_{ji} and γ_{ij}, and

$$\mathbf{\Gamma}(m) = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)', \qquad \mathbf{\Gamma}(m) \in \Re^{n \times m},$$
$$\mathbf{B}(m) = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m), \qquad \mathbf{B}(m) \in \Re^{n \times m},$$

where $\|\mathbf{Y}\|_F = \sqrt{Tr\left(\mathbf{Y}\mathbf{Y}'\right)}$ is the Frobenius norm for a generic matrix $\mathbf{Y}$,

Archetypal Analysis (Cutler and Breiman, 1994)

Given m, the m archetypes $\mathbf{A}(m) = \mathbf{a}_1, \ldots \mathbf{a}_m$ are:

$$\mathbf{A}(m) = \arg \min_{\boldsymbol{\Gamma}(m), \mathbf{B}(m)} \left\| \mathbf{X} - \boldsymbol{\Gamma}(m) \mathbf{B}'(m) \mathbf{X} \right\|_F \tag{3}$$

s,t, all the conditions on coefficients β_{ji} and γ_{ij}, and

$$\begin{aligned} \boldsymbol{\Gamma}(m) &= (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)', & \boldsymbol{\Gamma}(m) &\in \Re^{n \times m}, \\ \mathbf{B}(m) &= (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m), & \mathbf{B}(m) &\in \Re^{n \times m}, \end{aligned}$$

where $\left\| \mathbf{Y} \right\|_F = \sqrt{Tr\left(\mathbf{Y}\mathbf{Y}'\right)}$ is the Frobenius norm for a generic matrix $\mathbf{Y}$,

Example: Wheat varieties

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

Let us see AA at work

EXCEL

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 perfomance analysis (Heavlin, 2008)

Outline

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.

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.

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.

[1] D'Esposito M.R., Palumbo F. and Ragozini G. (2012). Interval Archetypes: A New Tool for Interval Data Analysis. *Statistical Analysis and Data Mining*, 5(4):322–335.

Outline

The study of uncertainty

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 study of uncertainty

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 study of uncertainty

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

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

precision/accuracy: to have exact measures of precise objets;

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

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 objets;**

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

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

precision/accuracy: to have exact measures of precise objets;

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.

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 objets;

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

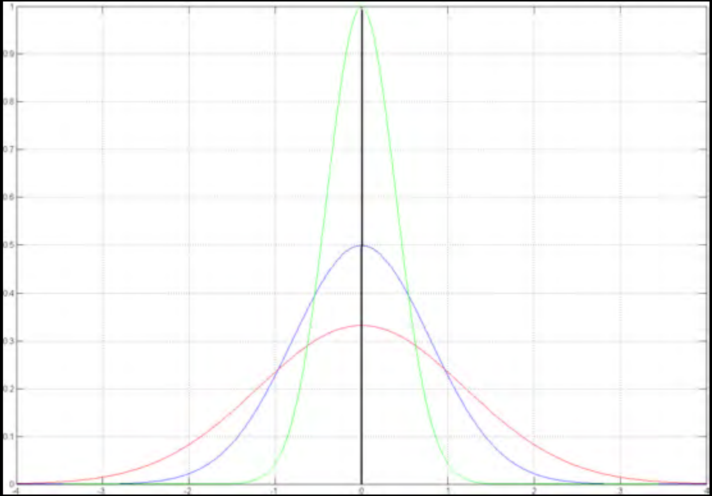
Coding of Imprecision and Vagueness

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.

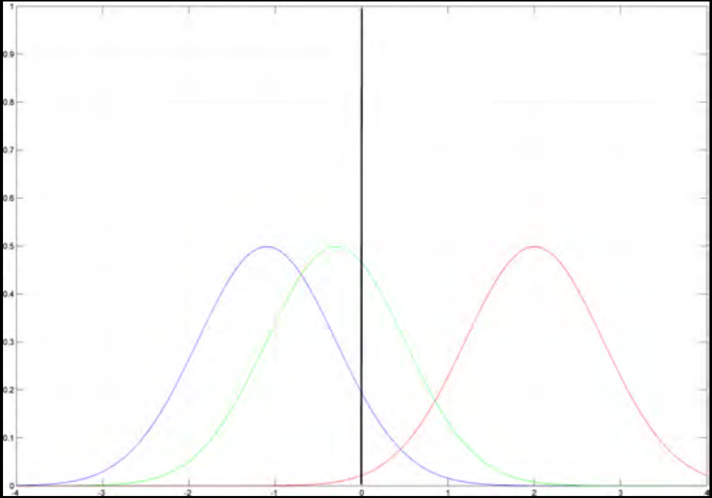
Same precision but different accuracy



Accuracy

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

Same accuracy but different precision



Outline

Interval Arithmetic

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*:

$$\begin{aligned} \mathbb{x} &= [\underline{x}, \overline{x}], \\ \mathbb{x} &= \{x \in \Re | \underline{x} \leq x \leq \overline{x}\}, \end{aligned}$$

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

The textbook of [Neumaier, 1990] is good reference to approach the IA.

Digression: Interval Analysis

- 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**

Interval Arithmetic: basic concepts

Given $\mathbb{x} = [\underline{x}, \overline{x}]$ and $\mathbb{y} = [\underline{y}, \overline{y}]$, the result of $\mathbb{x} \diamond \mathbb{y}$ is again an interval $\mathbb{z}$ with property:

$$\mathbb{x} \diamond \mathbb{y} = \mathbb{z} = \{z = x \diamond y | x \in \mathbb{x}, y \in \mathbb{y}\} \tag{4}$$

where $\diamond$ belongs to the set $\{+, -, \times, \div\}$,
Arithmetic operations on intervals are expressed in terms of ordinary arithmetics on their bounds as follows:

$$\begin{aligned}
\mathbb{x} + \mathbb{y} &= [\underline{x} + \underline{y}, \overline{x} + \overline{y}] \\
\mathbb{x} - \mathbb{y} &= [\underline{x} - \overline{y}, \overline{x} - \underline{y}] \\
\mathbb{x} \times \mathbb{y} &= [\min([\underline{x} \times \underline{y}, \underline{x} \times \overline{y}, \overline{x} \times \underline{y}, \overline{x} \times \overline{y}), \\
&\quad \max(\underline{x} \times \underline{y}, \underline{x} \times \overline{y}, \overline{x} \times \underline{y}, \overline{x} \times \overline{y})] \\
\mathbb{x} \div \mathbb{y} &= [\min([\underline{x} \div \underline{y}, \underline{x} \div \overline{y}, \overline{x} \div \underline{y}, \overline{x} \div \overline{y}), \\
&\quad \max(\underline{x} \div \underline{y}, \underline{x} \div \overline{y}, \overline{x} \div \underline{y}, \overline{x} \div \overline{y})]
\end{aligned}$$

with the extra condition $0 \notin \mathbb{y}$ for the division,

Midpoint/Range notation

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

$$\begin{aligned} \check{x} &= \frac{1}{2}(\underline{x} + \overline{x}), \\ \Delta x &= \frac{1}{2}(\overline{x} - \underline{x}), \end{aligned}$$

The interval $\mathbb{x}$ can also be written as:

$$\mathbb{x} = [\underline{x}, \overline{x}] = [\check{x} - \Delta x, \check{x} + \Delta x] = \{\check{x}, \Delta x\},$$

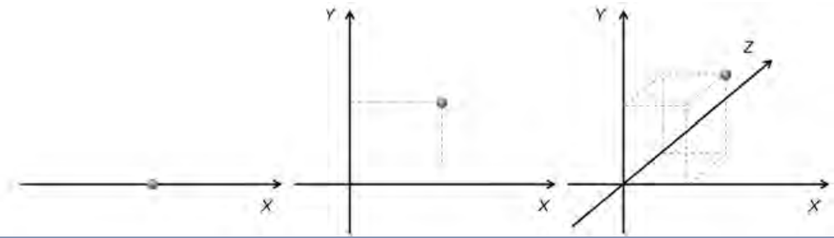
Dealing with intervals

- 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,

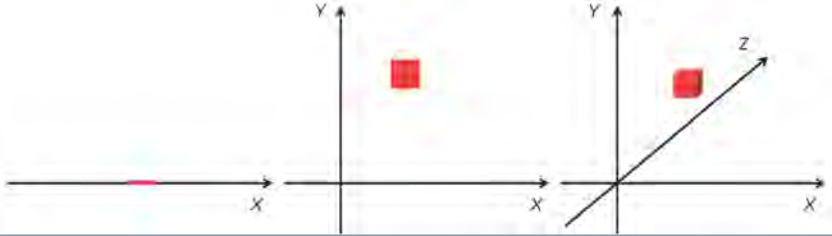
A point is a point in the $\Re^p$, $\forall p \geq 1$,



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 $\Re^p$, $\forall p \geq 1$, varies according to p: *line* for $p = 1$, *rectangle* for $p = 2$, *parallelotope* 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:

$$H(A,B) = \max\{\mid \overline{a} - \overline{b} \mid, \mid \underline{a} - \underline{b} \mid\} = \mid \check{a} - \check{b} \mid + \mid \Delta a - \Delta b \mid,$$

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 $\mathbb{R}$,

The Hausdorff Distance in $\mathbb{R}$

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

$$H(A,B) = \max\{\mid \overline{a} - \overline{b} \mid, \mid \underline{a} - \underline{b} \mid\} = \mid \check{a} - \check{b} \mid + \mid \Delta a - \Delta b \mid,$$

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 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:

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

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

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

i) $H(A, A) = 0 \Leftrightarrow A = A$, $\forall A$, 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:

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

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

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

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

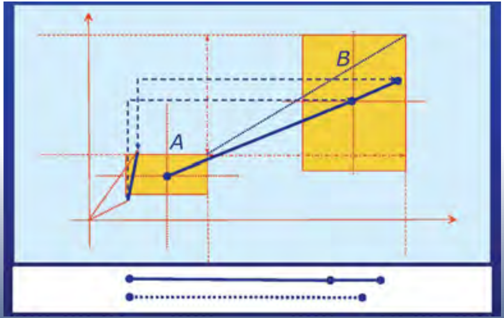
ii) $H(A,B) = H(B,A)$ (*Symmetry*)

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 $\mathbb{IR}^p$ introduced, for $\alpha = 2$ can also be expressed in terms of *midpoints* and *ranges*:

$$H(A,B) = \sqrt{\sum_{j=1}^{p} \left[(\check{a}_j - \check{b}_j)^2 + (\Delta a_j - \Delta b_j)^2 + 2 \mid \check{a}_j - \check{b}_j \mid \mid \Delta a_j - \Delta b_j \mid \right]},$$



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 $\check{A} = [\check{a}_j]$ $(j = 1, \ldots, p)$ 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:

$$H(A, B) = \sqrt{\sum_{j=1}^{p} \left(\check{a}_j - \check{b}_j\right)^2} + \mid r_A - r_B \mid \tag{5}$$

Where $\check{a}_j$ and $\check{b}_j$ indicate the generic centers coordinates of A and B, and r_A and r_B are the respective radii,

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],

The Hausdorff Distance for parallalotopes

Given the interval data matrix $\mathbb{X}$ and two parallelotopes $\mathbb{x}'_i$ and $\mathbb{x}'_{i'}$ in $\mathbb{IR}^p$, the Hausdorff distance between them is:

$$\begin{aligned} H(\mathbb{x}'_i, \mathbb{x}'_{i'}) &= \sum_{k=1}^{p} \max\left\{\left|\overline{x}_{ik} - \overline{x}_{i'k}\right|, \left|\underline{x}_{ik} - \underline{x}_{i'k}\right|\right\} \\ &= \sum_{k=1}^{p} \left(\left|\check{x}_{ik} - \check{x}_{i'k}\right| + \left|\Delta x_{ik} - \Delta x_{i'k}\right|\right), \end{aligned}$$

[Chavent, 2004]

Variability

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

$$d(\mathbb{x}_i, \mathbb{x}_{i'}) = \sqrt{\sum_{j=1}^{p} d^2(\mathbb{x}_{ij}, \mathbb{x}_{i'j})},$$

The centered interval-valued data matrix $\mathbb{Y} = \mathbb{X} - \mathbf{u}\bar{x}'$, 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 $\mathbb{Y}$ can be also written in midpoints and ranges notation: $\mathbb{Y} \equiv \{\check{\mathbf{Y}}, (\Delta\mathbf{Y})\}$,

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:

$$\mathbf{V}_{\mathbb{Y}} = \frac{1}{N}\left\{\check{\mathbf{Y}}'\check{\mathbf{Y}} + (\Delta\mathbf{Y})'(\Delta\mathbf{Y}) + \left[\mid \check{\mathbf{Y}}'(\Delta\mathbf{Y}) \mid + \mid (\Delta\mathbf{Y})'\check{\mathbf{Y}} \mid\right]\right\},$$

Outline

Archetypes for Interval Data

In analogy with the single value case, we define the *archetypes* $\mathbb{A}$ 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 γ_i' are imposed to be the same in the two spaces.
- Hence the γ_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].

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 $\mathbb{A}(m)$ can be determined by minimizing

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

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

- Thus, given m and the quantity:

$$\mathbb{RSS}(m) = \left\| d\left(\mathbb{X}, \tilde{\mathbb{X}}(m)\right) \right\|_F = \left\| \mathbb{X} - (\boldsymbol{\Gamma}(m)\mathbf{B}(m)\mathbb{X}) \right\|_F .$$

- the m archetypes solve the minimization problem:

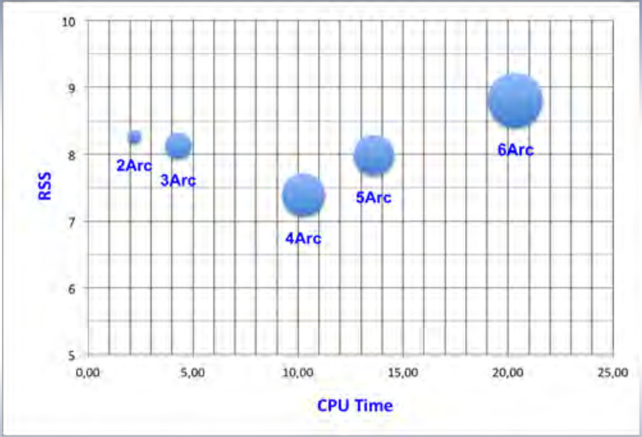
$$\min_{\boldsymbol{\Gamma}(m),\mathbf{B}(m)} \mathbb{RSS}(m) = \min_{\boldsymbol{\Gamma}(m),\mathbf{B}(m)} \left\| \mathbb{X} - (\boldsymbol{\Gamma}(m)\mathbf{B}(m)\mathbb{X}) \right\|_F$$

Example: Wheat varieties

Interval valued data

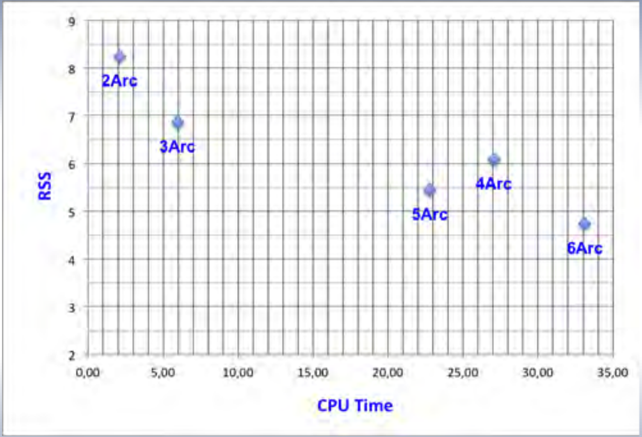
Wheat variety	Humidity	Weight	Protein	Ash	Glutine	iGlut	iYell
Anco Marzio	11.2 - 13.4	82 - 85	11.0 - 13.5	1.51 - 2.01	9.0 - 11.4	72 - 92	20.5 - 24.2
Ciccio	11.2 - 12.1	83 - 86	9.7 - 14.2	1.80 - 2.06	6.3 - 10.5	67 - 95	20.6 - 24.3
Claudio	11.1 - 12.8	81 - 85	10.8 - 14.2	1.82 - 2.16	8.6 - 12.8	62 - 91	21.5 - 24.1
Creso	11.0 - 13.4	81 - 85	10.8 - 14.0	1.72 - 2.00	8.9 - 13.2	51 - 87	18.5 - 24.2
Duilio	11.2 - 13.0	79 - 84	9.8 - 15.2	1.77 - 2.09	8.1 - 11.4	60 - 84	21.2 - 22.7
Iride	11.3 - 13.2	77 - 84	10.5 - 14.2	1.79 - 2.20	7.3 - 11.4	65 - 93	22.0 - 25.2
Levante	11.3 - 13.0	79 - 85	11.4 - 13.7	1.87 - 2.28	9.4 - 12.0	71 - 93	23.9 - 27.2
Orobel	10.9 - 12.7	79 - 84	11.7 - 13.7	2.05 - 2.20	9.2 - 12.3	42 - 89	20.7 - 26.7
Quadrato	11.5 - 11.7	81 - 83	11.5 - 12.5	1.80 - 2.12	9.5 - 10.8	90 - 96	23.8 - 25.3
Rusticano	11.9 - 13.6	81 - 85	11.6 - 14.5	1.86 - 2.07	10.0 - 11.9	62 - 96	21.5 - 24.0
San Carlo	11.7 - 14.0	77 - 84	11.9 - 14.4	1.77 - 2.20	8.4 - 13.0	62 - 95	22.8 - 25.8
Saragolla	11.2 - 12.8	78 - 82	11.1 - 15.6	2.08 - 2.21	7.5 - 12.4	75 - 98	21.7 - 26.6
Simeto	10.2 - 13.4	81 - 86	8.7 - 12.8	1.67 - 2.00	6.4 - 11.3	52 - 95	21.5 - 25.7
Svevo	11.1 - 12.7	77 - 84	11.7 - 15.4	1.83 - 2.17	8.3 - 13.6	53 - 90	24.1 - 28.0

How many archetypes?



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?



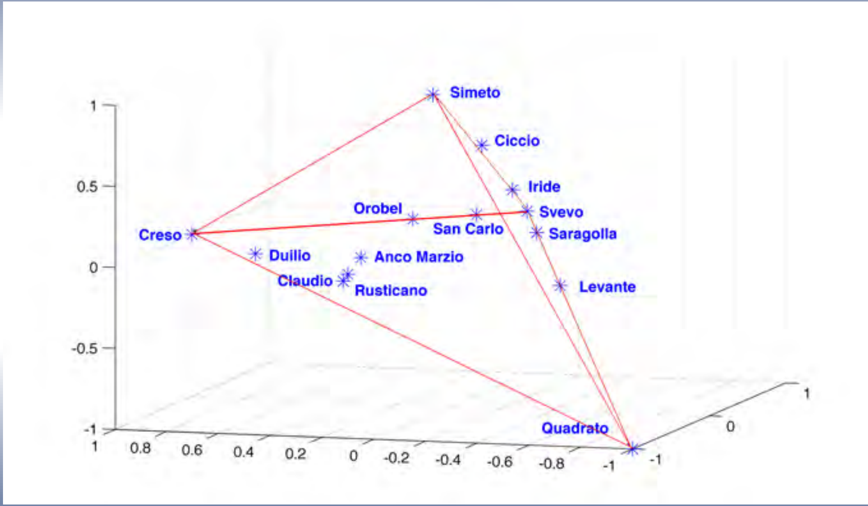
Data represent the best RSS value reached over 20 random start solutions and the corresponding CPU time (mill-seconds on this notebook), for any given m.

Weighting coefficients

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

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

Weighting coefficients



Clusters in the data

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

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

Clusters in the data

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

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

Clusters in the data

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

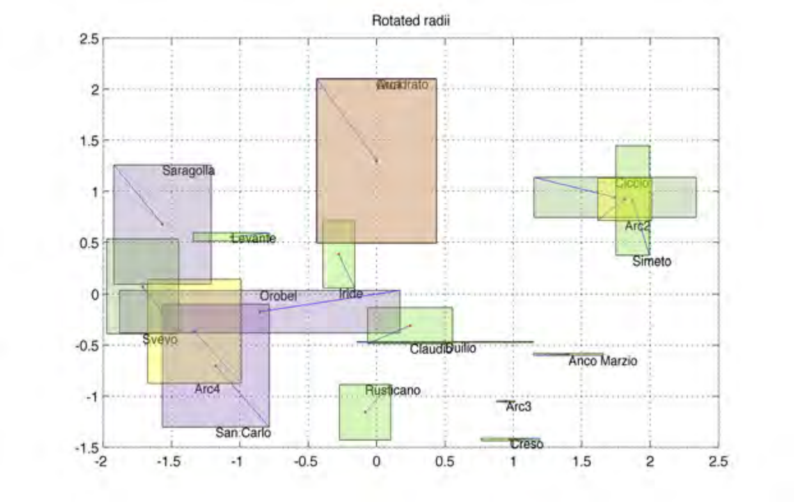
Wheat	$\gamma_1(4)$	$\gamma_2(4)$	$\gamma_3(4)$	$\gamma_4(4)$
Anco Marzio	0.000	0.528	0.213	0.259
Ciccio	0.000	0.000	0.828	0.172
Claudio	0.167	0.599	0.018	0.217
Creso	0.000	1.000	0.000	0.000
Duilio	0.062	0.841	0.000	0.097
Iride	0.670	0.000	0.272	0.058
Levante	0.687	0.000	0.000	0.313
Orobel	0.659	0.341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

Clusters in the data

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

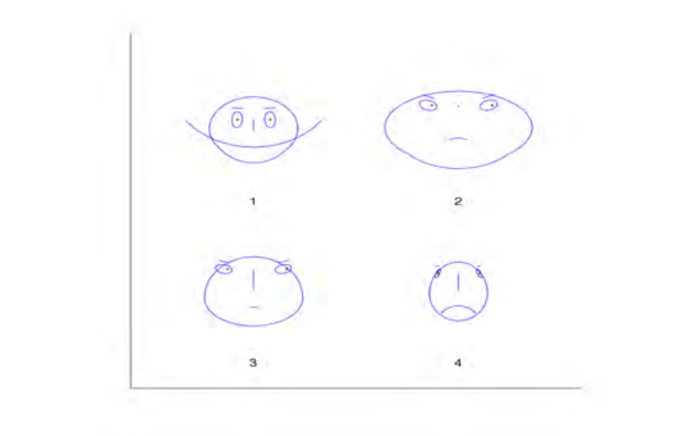
Wheat	$\gamma_1(4)$	$\gamma_2(4)$	$\gamma_3(4)$	$\gamma_4(4)$
Anco Marzio	0.000	0.528	0.213	0.259
Ciccio	0.000	0.000	0.828	0.172
Claudio	0.167	0.599	0.018	0.217
Creso	0.000	1.000	0.000	0.000
Duilio	0.062	0.841	0.000	0.097
Iride	0.670	0.000	0.272	0.058
Levante	0.687	0.000	0.000	0.313
Orobel	0.659	0.341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

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



Four archetypes (in yellow) are represented as supplementary rectangles.

Archetypes interpretation



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.

Outline

Archetypes and Prototypes

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 ($\mathbf{A}(K)$ or $\mathbb{A}(K)$);
- Represent data in the space spanned by the archetypes through the coefficient γ;
- Detect clusters and data structure through the γ coefficients;
- Find the prototypes $\mathcal{P}_h$ solving a minimization problem based on a appropriate distance function in this space;
- Revert to the original space ($\Re^p$ or $\mathbb{R}^p$) of the data to determine the prototypes $\mathbf{p}_k$.

The distance function

γ coefficient as compositional data

- Note that the γ 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:

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

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

$$\left[\sum_{j=1}^{p} \left(\log \frac{x_{ij}}{g(\mathbf{x}_i)} - \log \frac{x_{i'j}}{g(\mathbf{x}_{i'})} \right)^2 \right]^{\frac{1}{2}}$$

with $g(\mathbf{x}_\cdot) = \left(\prod_{j=1}^{p} x_{\cdot j} \right)^{\frac{1}{p}}$ [Aitchison et al., 2000].

Prototype identification in the γ space

Recall

Given a partition $(C_1, \ldots, C_K)$ of Ω in K clusters of sizes $n_1, \ldots, n_K$ and a dissimilarity measure $d(\cdot, \cdot)$, a prototype $\mathbf{p}_h$, $h = 1, \ldots, K$ minimizes the function $\sum_{i \in C_h} d(\mathbf{x}_i, \mathbf{p}_h)$

In our case

- $d(\cdot, \cdot)$ is the compositional distance in the space spanned by the archetypes $(\mathcal{S}^K)$
- the prototype in the γ 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 $g_{hj} = \left(\prod_{i \in C_k} \gamma_{ij} \right)^{\frac{1}{n_h}}$ [Martýn-Fernandez et al., 1998].

Prototype identification in the original space

Remember that:

$$\mathbf{x}'_i = \gamma'_i \mathbf{A}$$

$$\mathbb{x}'_i = \gamma'_i \mathbb{A}$$

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

$$\mathbf{p}'_h = \mathcal{P}'_h \mathbf{A} = \mathcal{P}'_h \mathbf{B}\mathbf{X}$$

$$\mathbb{p}'_h = \mathcal{P}'_h \mathbb{A} = \mathcal{P}'_h \mathbf{B}\mathbb{X}$$

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

Outline

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;

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;

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

Aitchison, J. (1982).
The statistical analysis of compositional.
Journal of the Royal Statistcal Society B, 44:139–177.

Aitchison, J., Barcelo-Vidal, C., Martyn-Fernandez, J. A., and Pawlowsky-Glahn, V. (2000).
Logratio analysis and compositional distance.
Mathematical Geology, 32(271-275).

Alefeld, G. and Herzerberger, J. (1983).
Introduction to Interval computation.
Academic Press, New York.

Braun, D., Mayberry, J., Powers, A., and Schlicker, S. (2003).
The geometry of the Hausdorff metric.
Available at:
`http://faculty.gvsu.edu/schlicks/Hausdorff_ paper.pdf`.

Chavent, M. (2004).
A hausdorff distance between hyper-rectangles for clustering interval data.
In Banks, D., House, L., McMorris, F. R., Arabie, P., and Gaul, W., editors, *Classification, Clustering, and Data Mining Applications*, volume 0 of *Studies in Classification, Data Analysis, and Knowledge Organization*, pages 333–339. Springer Berlin Heidelberg.
10.1007/978-3-642-17103-1_32.

Corsaro, S. and Marino, M. (2010).
Archetypal analysis of interval data.
Archetypal Analysis of Interval Data, 14(2):105–116.

de Carvalho, F. A., de Souza, R. M., Chavent, M., and Lechevallier, Y. (2006).
Adaptive hausdorff distances and dynamic clustering of symbolic interval data.
Pattern Recognition Letters, 27(3):167 – 179.

D'Esposito, M. R., Palumbo, F., and Ragozini, G. (2006).
Archetipal analysis for interval data in marketing research.
Statistica Applicata - Italian Journal of Applied Statistics Statistica Applicata - Italian Journal of Applied Statistics Statistica Applicata - Italian Jour. of Appl. Stat., 18(2):343–358.

D'Esposito, M. R., Palumbo, F., and Ragozini, G. (2011).
Interval archetypes: a new tool for interval data analysis.
Statistical Analysis and Data Mining.

Hickey, T., Ju, Q., and Van Emden, M. H. (2001).
Interval arithmetic: From principles to implementation.
Journal of the ACM, 48(5):1038–1068.

Kearfott, R. B. (1996).
Rigorous Global Search: Continuous Problems.
Kluver, Dordrecht.

Martýn-Fernandez, J. A., Barcelo-Vidal, C., and Pawlowsky-Glahn, V. (1998).
Measures of diference for compositional data and hierarchical clustering methods.
In *IAMG*, pages 526–531.

Moore, R. E. (1966).
Interval Analysis.
Prentice-Hall, Englewood Cliffs, NJ.

Neumaier, A. (1990).
Interval methods for systems of Equations.
Cambridge University Press, Cambridge.

Palumbo, F. and Lauro, C. N. (2003).
A PCA for interval valued data based on midpoints and radii.
In Yanai, H., Okada, A., Shigemasu, K., Kano, Y., and Meulman, J., editors, *New developments in Psychometrics*, Tokyo. Psychometric Society, Springer-Verlag.

Zadeh, L. (1965).
Fuzzy sets.
Information and Control, 8:338–353.