Symbolic Data Analysis: Past, Present and Future

Paula Brito

Fac. Economia, Universidade do Porto & LIAAD-INESC TEC, Portugal

Advances in Data Science for Big and Complex Data
Université Paris Dauphine 10-11 January 2019

Outline

Outline

Symbolic Data Analysis (Diday (1988))

Objectives:

- To consider data that contain information which cannot be represented within the classical data models
- Produce results directly interpretable in terms of the input descriptive variables

"Model" for data representation:
allow taking into account intrinsic **variability**,
representing with a same language, e.g., the elements of a set and clusters of this set.

First "models":

- logical approach
- representation of a given set by intent, i.e., by its properties

Symbolic Data Analysis

Symbolic object:

Description expressed by means of a conjunction of events in terms of the values taken by the variables.

Description of car **model** "AlfaRomeo":

$$s_{\text{Alfa}} = [\mathit{Price} \in [27806, 33596]] \wedge [\mathit{Engine\ Capacity} \in [1000, 3000]] \wedge$$
$$\wedge [\mathit{Colour} \sim \{\mathit{Red}(30\%), \mathit{Black}(70\%)\}]$$

Symbolic Data Analysis

Focus on the duality between

- intent - the description
- extent - the set of "individuals" verifying this description

A *concept* is defined as a pair (extent, intent).

For appropriately defined generalization operators, and mappings f to get the intent of a given set and g to get the extent of a given description, the set of concepts forms a Galois lattice: **concept lattice**.

Using those operators, conceptual clustering methods have been defined.

- Brito, P., Polaillon, G. (2005). Structuring Probabilistic Data by Galois Lattices
- Brito, P. (1995). Symbolic Objects: Order Structure and Pyramidal Clustering.
- Brito, P. and Diday, E. (1990). Pyramidal Representation of Symbolic Objects.

Symbolic Data Analysis

- Need of a standardised data representation
- Distance-based methodologies
- Statistically-rooted models

$\rightarrow$ Tabular data representation, where n units in rows take "values" for p variables in columns:

Model	Price	Eng. Capacity	Colour
Alfa Romeo	$[27806, 33596]$	$[1000, 3000]$	{ Red (30%), Black (70%)}

Symbolic Data-Analysis $\longrightarrow$ Analysis of Symbolic Data

The Data

Symbolic Data Analysis:
to take into account **variability** inherent to the data

Variability occurs when we have

- Descriptors on flights, but: analyse the airport - not each individual flight
- Descriptors on prescriptions, but: analyse patients, or doctors - not the individual prescriptions
- Official statistics - Descriptors on citizens, but: analyse the cities, the regions - not the individual citizens

$\Longrightarrow$ (symbolic) variable values are

sets, intervals

distributions on an underlying set of sub-intervals or categories

Micro-data $\longrightarrow$ **Macro-data**

The Data

- In most common applications, symbolic data arises from the aggregation of micro data
- Often reported as such: temperature min-max intervals , financial assets daily min-max or open-close values
- They also occur directly, in descriptions of concepts: diseases, biological species (plants, etc.), technical specifications,...
- Quantile lists: infant growth, plant measures, etc.

The Data

Flight activity in airports, over a given period

Airport	Passengers	Companies	Delay (minutes)
A	$[150, 200]$	$\{1, 2\}$	$\{[0, 10[\,, 0.25; [10, 30[\,, 0.65; [30, 60]\,, 0.10\}$
B	$[180, 300]$	$\{1, 2, 3\}$	$\{[0, 10[\,, 0.45; [10, 30[\,, 0.30; [30, 60]\,, 0.25\}$
C	$[200, 400]$	$\{1, 3\}$	$\{[0, 10[\,, 0.75; [10, 30[\,, 0.20; [30, 60]\,, 0.05\}$

Airport	Shares of arriving flights
A	{British (0.25), Lufthansa (0.40), Air France (0.35)}
B	{British (0.10), Lufthansa (0.15), Air France (0.60), Iberia (0.15)}
C	{Lufthansa (0.30), Air France (0.50), Iberia (0.20)}

Symbolic Variable Types

- Numerical (Quantitative) variables
  - Numerical single-valued variables
  - Numerical multi-valued variables
  - **Interval variables**
  - **Distributional variables: Histograms, Quantile lists**
- Categorical (Qualitative) variables:
  - Categorical single-valued variables
  - Categorical multi-valued variables
  - Distributional variables: Categorical modal - Compositions

- Brito, P. (2002). Hierarchical and Pyramidal Clustering for Symbolic Data. Journal of the Japanese Society of Computational Statistics, 15 (2), 231-244.
- Noirhomme-Fraiture, M., Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. Statistical Analysis and Data Mining, 4 (2), 157-170.

Symbolic Variable Types

$S = \{s_1, ..., s_n\}$: the set of n units to be analyzed.
Let $Y_1, \ldots, Y_p$ be the variables, O_j the underlying domain of Y_j
B_j the observation space of $Y_j, j = 1, \ldots, p$

$$Y_j : S \longrightarrow B_j$$

- Y_j classical (numerical or categorical) single-valued variable: $B_j \equiv O_j$
- Y_j numerical or categorical multi-valued variable: $B_j = P(O_j)$
- Y_j interval variable: B_j set of closed and bounded intervals of O_j
- Y_j categorical modal or histogram variable: B_j set of distributions on O_j

Outline

Interval-Valued Variables

$S = \{s_1, ..., s_n\}$: the set of n units to be analyzed

Interval-valued variable :

$Y_j : S \to B : \ Y_j(s_i) = [l_{ij}, u_{ij}], l_{ij} \leq u_{ij}$

B : the set of closed and bounded intervals of an underlying set $O \subseteq \boldsymbol{R}$

For modelling purposes $\rightarrow$ often preferable to represent $Y_j(s_i)$ by

- the midpoint $c_{ij} = \dfrac{l_{ij} + u_{ij}}{2}$
- the range $r_{ij} = u_{ij} - l_{ij}$

Interval Data

	Y_1	...	Y_j	...	Y_p
s_1	$[l_{11}, u_{11}]$	...	$[l_{1j}, u_{1j}]$	...	$[l_{1p}, u_{1p}]$
...	...		...		...
s_i	$[l_{i1}, u_{i1}]$	...	$[l_{ij}, u_{ij}]$	...	$[l_{ip}, u_{ip}]$
...	...		...		...
s_n	$[l_{n1}, u_{n1}]$	...	$[l_{nj}, u_{nj}]$	...	$[l_{np}, u_{np}]$

Interval Data

Analysis of Interval Data: two main approaches, within a non-parametric framework:

- Assuming a distribution within each observed interval - usually the Uniform, but not necessarily (e.g. Triangular)
- Represent an interval by two real numbers, the lower and upper bounds or the midpoint and (half) range, and develop models and methods using these two values

These are usually exploratory approaches, relying on distance-based criteria:

- Clustering
- Linear regression
- Dimension reduction
- Time-series analysis

Parametric Models for Interval Data

To allow for parametric inference methodologies
$\longrightarrow$ probabilistic models for interval-valued variables.

- Le Rademacher, Billard (2011): Assume Uniformity
  - Mean (midpoints) modelled by a Gaussian, Variance modelled by an Exponential, independence assumed
  - (Mean, Variance) jointly Gaussian

- Lima Neto, Cordeiro, de Carvalho (2011): Interval-valued variable as a bivariate random vector with a joint probability density function belonging to the bivariate exponential family, e.g. bivariate Gaussian, bivariate Gamma

- Brito & Duarte Silva (2008, 2012): Assume that the joint distribution of the midpoints C and the logs of the ranges R is multivariate Gaussian or multivariate Skew-Normal

Parametric Models for Interval Data

Brito & Duarte Silva modelling:

$R^* = ln(R), (C, R^*) \sim N_{2p}(\mu, \Sigma)$
OR
$(C, R^*) \sim SN_{2p}(\xi, \Omega, \alpha) \longrightarrow (C, R^*) \sim SN_{2p}(\mu, \Sigma, \gamma)$

$$\mu = [\mu_C^t, \mu_{R^*}^t]^t \; ; \; \Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{pmatrix} \; ; \; \gamma = [\gamma_C^t, \gamma_{R^*}^t]^t$$

μ_C and μ_{R^*} - p-dimensional column vectors of the mean values

$\Sigma_{CC}, \Sigma_{CR^*}, \Sigma_{R^*C}$ and $\Sigma_{R^*R^*}$ - $p \times p$ matrices

γ_C and γ_{R^*} - p-dimensional column vectors of the skewness coefficients

Parametric Models for Interval Data

However, for interval data:

Midpoint c_{ij} and Range r_{ij} of the value of an interval-valued variable are two quantities related to one only variable
$\rightarrow$ should not be considered separately

So: parametrizations of the global covariance matrix $\rightarrow$
take into account the link that may exist between midpoints and log-ranges of the same or different variables

Parametric Models for Interval Data

Most general formulation - allow for non-zero correlations among all Midpoints and Log-Ranges; other cases of interest:

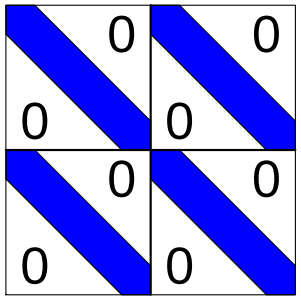
- The interval variables Y_j are non-correlated, but for each variable the Midpoint may be correlated with its Log-Range
- Midpoints (respectively, Log-Ranges) of different variables may be correlated, but no correlation between Midpoints and Log-Ranges is allowed
- All Midpoints and Log-Ranges are non-correlated

Config.	Characterization	Σ
1	Non-restricted	Non-restricted
2	Y_j's non correlated	$\Sigma_{CC}, \Sigma_{CR^*} = \Sigma_{R^*C}, \Sigma_{R^*R^*}$ all diagonal
3	C's non-correlated with R^*'s	$\Sigma_{CR^*} = \Sigma_{R^*C} = 0$
4	All C's and R^*'s are non-correlated	Σ diagonal

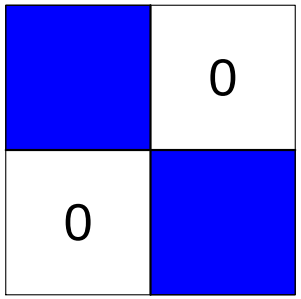
Parametric Models for Interval Data

$$\Sigma = \begin{bmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{bmatrix}$$

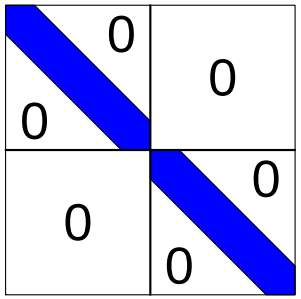
Configuration 1



Configuration 2



Configuration 3



Configuration 4

Brito, P., Duarte Silva, A.P. (2012).
Modelling Interval Data with Normal and Skew-Normal Distributions.
Journal of Applied Statistics, Volume 39, Issue 1, 3-20.

Multivariate Parametric Analysis of Interval Data

- ANOVA (likelihood-ratio approach)
- Discriminant analysis (linear and quadratic, location and general)
- Model-based clustering (finite-mixture modelling)
- Multivariate outlier detection (based on Mahalanobis distances)
- R Package MAINT.Data

Outline

Histogram-Valued Variables

Histogram-valued variable: $Y : S \to B$

B: the set of all possible partitions of any compact of $\boldsymbol{R}$ and all possible distributions over the (finite set of) corresponding sub-intervals.

$$Y(s_i) = (I_{i1}, p_{i1}; \ldots; I_{ik_i}, p_{iK_i})$$

$p_{i\ell}$: probability or frequency associated to $I_{i\ell} = [\underline{I}_{i\ell}, \overline{I}_{i\ell}[$
$p_{i1} + \ldots + p_{iK_i} = 1$

$Y(s_i)$ may be represented by the histogram:

$$H_{Y(s_i)} = ([\underline{I}_{i1}, \overline{I}_{i1}[, p_{i1}; \ldots; [\underline{I}_{iK_i}, \overline{I}_{iK_i}], p_{ijK_i})$$

Histogram Data

Example:

Airport	Delay (minutes)
C	$\{[0,10[\,,0.75;[10,30[\,,0.20;[30,60]\,,0.05\}$

In a multivariate framework:

	Y_1	...	Y_p
s_1	$\{[\underline{l}_{111},\overline{l}_{111}[,p_{111};\ldots;[\underline{l}_{11K_{11}},\overline{l}_{11K_{11}}],p_{11K_{11}}\}$	...	$\{[\underline{l}_{1p1},\overline{l}_{1p1}[,p_{1p1};\ldots;[\underline{l}_{1pK_{1p}},\overline{l}_{1pK_{1p}}],p_{1pK_{1p}}\}$
...	...		...
s_i	$\{[\underline{l}_{i11},\overline{l}_{i11}[,p_{i11};\ldots;[\underline{l}_{i1K_{i1}},\overline{l}_{i1K_{i1}}],p_{i1K_{i1}}\}$	...	$\{[\underline{l}_{ip1},\overline{l}_{ip1}[,p_{ip1};\ldots;[\underline{l}_{ipK_{ip}},\overline{l}_{ipK_{ip}}],p_{ipK_{ip}}\}$
...	...		...
s_n	$\{[\underline{l}_{n11},\overline{l}_{n11}[,p_{n11};\ldots;[\underline{l}_{n1K_{n1}},\overline{l}_{n1K_{n1}}],p_{n1K_{n1}}\}$	...	$\{[\underline{l}_{np1},\overline{l}_{np1}[,p_{np1};\ldots;[\underline{l}_{npK_{np}},\overline{l}_{npK_{np}}],p_{npK_{np}}\}$

Histogram-Valued Variables

- Assumption: within each sub-interval $[\underline{l}_{ij\ell}, \overline{l}_{i\ell}[$ the values of variable Y_j for observation s_i are uniformly distributed
- Proposed methods
  - based on sample moments derived from uniformity assumption
  - based on the representation of the histograms by the associated quantile functions, for which appropriate distances are considered
  - more recently: Dirichlet mixtures
- Interval-valued variables: particular case of histogram-valued variables: $Y_j(s_i) = [l_{ij}, u_{ij}] \rightarrow H_{Y_j(s_i)} = ([l_{ij}, u_{ij}], 1)$

Histogram-Valued Variables

$Y(s_i)$ can be represented by the inverse cumulative distribution function - quantile function

$$\Psi : [0,1] \longrightarrow \boldsymbol{R}$$

$$\Psi_i(t) = \begin{cases} \underline{I}_{i1} + \frac{t}{w_{i1}} r_{i1} \ \ if \ \ 0 \leq t < w_{i1} \\ \underline{I}_{i2} + \frac{t - w_{i1}}{w_{i2} - w_{i1}} r_{i2} \ \ if \ \ w_{i1} \leq t < w_{i2} \\ \vdots \\ \underline{I}_{ijK_i} + \frac{t - w_{iK_i - 1}}{1 - w_{iK_i - 1}} r_{iK_i} \ \ if \ \ w_{iK_i - 1} \leq t \leq 1 \end{cases}$$

where $\quad w_{ih} = \sum_{\ell=1}^{h} p_{i\ell}, h = 1, \ldots, K_i; r_{i\ell} = \overline{I}_{i\ell} - \underline{I}_{i\ell}$

for $\ell = \{1, \ldots, K_i\}$

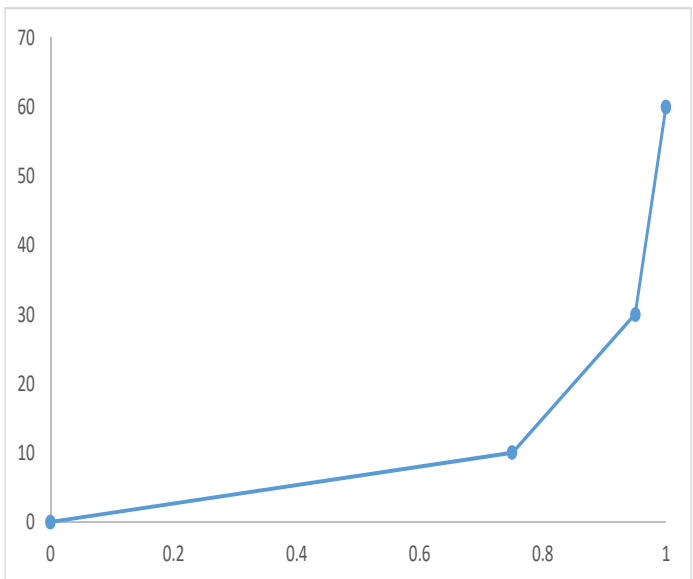
These are piecewise linear functions (from uniformity assumption).

Histogram-Valued Variables: Example

Airport	Delay (minutes)
C	$\{[0, 10[\,, 0.75; [10, 30[\,, 0.20; [30, 60]\,, 0.05\}$

The associated quantile function is

$$\Psi(t) = \begin{cases} \frac{t}{0.75} \times 10 = \frac{40t}{3}, & 0 \leq t < 0.75 \\ 10 + \frac{t-0.75}{0.20} \times 20 = & 0.75 \leq t < 0.95 \\ = 100t - 65 & \\ 30 + \frac{t-0.95}{0.05} \times 30 = & 0.95 \leq t \leq 1 \\ = 600t - 540 & \end{cases}$$



Histogram-Valued Variables: Distance Measures

Many measures proposed in the literature
(see e.g. Bock and Diday (2000), Gibbs (2002))

Distances based on the quantile functions:

- **Wasserstein distance**:
  $D_W(\Psi_{Y(i)}, \Psi_{Y(i')}) = \int_0^1 \left|\Psi_{Y(i)}(t) - \Psi_{Y(i')}(t)\right| dt$
- **Mallows distance**:
  $D_M(\Psi_{Y(i)}, \Psi_{Y(i')}) = \sqrt{\int_0^1 (\Psi_{Y(i)}(t) - \Psi_{Y(i')}(t))^2 dt}$

Histogram-Valued Variables: Distance Measures

Mallows distance:

Under the uniformity hypothesis,
and considering a fixed weight decomposition
(same weights, different intervals),
we have (Irpino and Verde (2006)):

$$D_M^2(\Psi_{Y(i)}, \Psi_{Y(i')}) =$$
$$= \sum_{\ell=1}^{K} p_\ell \left[(c_{Y(i)\ell} - c_{Y(i')\ell})^2 + \frac{1}{3}(r_{Y(i)\ell} - r_{Y(i')\ell})^2 \right]$$

For interval data:

$$D_M^2(\Psi_{Y(i)}, \Psi_{Y(i')}) = \left[(c_{Y(i)} - c_{Y(i')})^2 + \tfrac{1}{3}(r_{Y(i)} - r_{Y(i')})^2 \right]$$

Verde, Irpino (2008) : Mahalanobis-Wasserstein distance

Analysis of Histogram-Valued Data

Based on the Quantile Function representation, and using properties of the Mallows distance, multivariate analysis methods have been developed:

- Clustering - Dynamic clustering, Hierarchical, Divisive, Self-Organizing Maps
- Linear regression
- Discriminant analysis

Histogram-valued variables

What if uniformity cannot be assumed ?

Quantile representation of distributional data

Ichino (2008); Ichino, Brito (GfKI 2010)

- **Objective**: Obtain a common representation model for different variable types
- Allowing to apply multivariate methods to the full (originally) mixed data array
- Discrete approach : For each observation $Y_j(s_i)$ - use the m-quantiles of the underlying distribution of the observed data values $(min; Q_1; \ldots; Q_{m-1}; max)$ (Ichino, 2008)
- When quartiles are chosen $(m = 4)$: representation for each variable is defined by the 5-uple $(min; Q_1; Q_2; Q_3; max)$

Histogram-Valued Variables: Distance Measures

Mallows distance:
Discrete version :

For two empirical distributions f, g , with n quantiles, we obtain

$$d_M(f,g) = \left(\frac{1}{n} \sum_{i=1}^{n} \left| F_{(i)}^{-1} - G_{(i)}^{-1} \right|^2 \right)^{1/2}$$ (Levina & Bickel, 2001)

Outline

Challenges to be addressed: data aggregation

Aggregation of microdata:

- Most symbolic data result from aggregation of microdata
- Several aggregation options - intervals, histograms,...
- In each case, relying on assumptions
- Results depend on representation model and underlying assumptions

Challenges to be addressed: data aggregation

Aggregation by intervals is not robust:
Outlier effect in intervals bounds and range

Robust aggregation

e.g. using Inter-Quartile or $[Q_{0.1}, Q_{0.9}]$ intervals

Distributions allow keeping more information
But : Require more data at the micro level

Histogram variables : how many observations should be required ?
How many classes (bins) should be considered ? Which quantiles ?

Need for a theoretical background of complex aggregation
(Batagelj (SDA 2018))
$\rightarrow$ "which complex data types are compatible with merging of disjoint sets of units" ?

Challenges to be addressed

- Interval data: Distribution to consider within observed intervals
  - Usual assumption: within each observed interval, the underlying random variable follows an Uniform distribution
  - Alternative: a symmetrical distribution depending solely on two points (min, max), (midpoint, range), e.g., Symmetrical Triangular distribution
  - General Triangular Distribution: need of a third point - Mode
  - Gaussian
  - More general choices: Beta...
- Polygonal variables: Silva, Souza & Cysneiros (2019)

Recent developments

- Interval time series modelling and forecasting
- Clustering of interval time series
- Non-linear modelling - via data transformations using kernels
- Spatial analysis of georeferenced symbolic data
- Links with Compositional Data Analysis (CoDA)
  - For categorical modal variables
  - Analysing histogram data as compositional data ?
  - Order on categories ("parts")
  - Multivariate compositions

Applications of SDA

Classically ...

- Official statistics
- Botany and Zoology: data have intrinsic variability

Emerging fields of application :

- Large surveys, e.g., at European level - analysed by region
- Text analysis - distributions over topics
- Econometric studies
- Finance modelling
- Internet traffic
- Demography
- Meta analysis

Applications of SDA

Complex data structures

- Social network analysis
  - Variability on the links' weights
  - Super nodes
  - Attributed networks
  - Multilayer networks
- Data streams
- Sensor data $\rightarrow$ "Smart statistics"

Future Perspectives

- Analysis based on marginal empirical distributions - the empirical distribution for each variable is considered separately

  $\longrightarrow$ need to go one step further, and consider the joint observed distributions in the data representation and analysis

- Data aggregation by histograms :
  Results depend heavily on the chosen partition or weights decomposition

  $\longrightarrow$ From histograms to densities - density-valued variables
  See: Functional Data Analysis - e.g. works of P. Delicado in "analysis for data which are density functions"

Future Perspectives

- Spatial-temporal modelling and analysis
- Explore links with Multilevel statistical analysis
- Big Data analytics and deep learning approaches are still to be developed for symbolic data
- Latent Dirichlet allocation, to be combined with the analysis of generated distributions

Outline

Concluding Remarks

- Symbolic Data Analysis: a framework where the variability may effectively be considered in the data representation and analysis
- Relevant for the analysis of large data sets: from micro-data to macro-data
- Interval, distribution-valued,...new (?) variable types to take variability - and structure ? - into account
- New problems / challenges:
  intervals, distributions,... are not real numbers

Concluding Remarks

Symbolic Data Analysis:

- A fast developing area
- Theoretical and methodological levels
- Growing number of fields of application
- An approach to cope with Big Data

A most relevant area of Data Science