SYMBOLIC DATA ANALYSIS

Resumo da lição a apresentar no âmbito da candidatura
ao título de agregado na área de conhecimento
em Matemática Aplicada da Universidade do Porto

Maria Paula de Pinho de Brito Duarte Silva
Março 2018

Contents

1 Introduction

Symbolic Data Analysis (SDA), introduced by Edwin Diday in the late eighties of the last century (Diday (1988)), is concerned with representing and analysing data presenting intrinsic variability, which is to be explicitly taken into account. In classical Statistics and Multivariate Data Analysis, the elements under analysis are generally individual entities for which a single value is recorded for each variable - e.g., individuals, described by their age, salary, education level, marital status, etc. But when the elements of interest are classes or groups of some kind - the citizens living in given towns; teams, consisting of individual players - then there is variability inherent to the data. To reduce this variability by taking central tendency measures - mean values, medians or modes - obviously leads to a significant loss of information. Symbolic Data Analysis (Bock and Diday (2000); Billard and Diday (2003, 2006); Diday and Noirhomme-Fraiture (2008); Brito (2014)) provides a framework allowing representing data with variability, using new variable types. Also, methods have been developed which suitably take data variability into account. Symbolic data may be represented using the usual matrix-form data arrays, where each entity is represented in a row and each column corresponds to a different variable - but now the elements of each cell are generally not single real values or categories, as in the classical case, but rather finite sets of values, intervals or, more generally, distributions.

In this lesson, we introduce and motivate the field of Symbolic Data Analysis, providing a historical perspective. We then detail the new variable types that have been introduced to represent variability, illustrating with some examples. We consider in particular the case of interval-valued data, i.e., where for each entity under analysis an interval of $\mathbb{R}$ is recorded, focusing on the parametric modelling for interval data proposed in Brito and Duarte Silva (2012). This modelling then allows for multivariate parametric analysis of multidimensional interval-valued data (Brito *et al* (2015); Duarte Silva and Brito (2015)). Next we consider the case of numerical data described by empirical distributions, known as histogram data. We introduce alternative representations of histogram observations, observing that interval-valued data constitutes a special case of those. Methods for the multivariate analysis of histogram-valued data are presented (Brito and Chavent (2012); Dias and Brito (2015, 2017)). We conclude by discussing open issues and research perspectives.

2 Symbolic Data

Since its introduction by Diday (1988), Symbolic Data Analysis has known a considerable development. It emerged from the need to consider data that contain information which cannot be represented within the classical data models, together with the objective of designing methods that produce results directly interpretable in terms of the input descriptive variables. The "model" for data representation should allow taking into account intrinsic variability - therefore allowing representing with a same language, e.g., the elements of a set and clusters of this set. The first "models" used a logical approach, rooted on the idea of representation of a given set by intent, i.e., by its properties. The need to represent elements defined by intent led to the introduction of *symbolic objects*. Generally speaking, a symbolic object is a description expressed by means of a conjunction of events in terms of the values taken by the variables, as in the following description of car model "AlfaRomeo" (the car model, not a single car !), in terms of Price, Engine Capacity and Colour:

$$\begin{aligned} s_{\text{Alfa}} = [Price \in [27806, 33596]] \wedge [Engine\ Capacity \in [1000, 3000]] \wedge \\ \wedge [Colour \sim \{Red(30\%), Black(70\%)\}] \end{aligned} \tag{1}$$

Symbolic objects differ therefore from the the classical *numerical objects* (represented and treated as vectors in $\mathbb{R}^p$) both at the description and syntactic levels. At the description level, the main differences come from the fact that variables are allowed to take multiple values for a given unit (thereby taking variability into account), and links between variable values may even be considered; at the syntactic level, symbolic objects are conceived to represent *knowledge* and not only (but also !) single observations. As a consequence, differences arise at the analysis level: the focus will now be on the duality between intent (the description) and extent (the set of individuals verifying this description) of a symbolic object, and generalization techniques are widely used (Brito and Diday (1990); Brito (1995)).

With the European projects on Symbolic Data Analysis - SODAS, "Symbolic Objects Data Analysis System", then followed by ASSO, "Analysis System of Symbolic Official data" - came the need of a standardised data representation, to allow for analysis by different methods; at the same time distance-based methodologies, closer to classical data analysis approaches, followed by statistically-rooted models were developed. This naturally led to the progressive departure from the logic-based representation of "symbolic objects", as in (1), in favour of a tabular data representation, more familiar to data analysts and statisticians, where n entities in rows take "values" for p variables in columns, as in Table 1.

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

Table 1: Tabular representation of symbolic data.

The formal definition of new *symbolic* variable types, and the study of their properties, followed. The community had moved from Symbolic Data-Analysis (a different approach to analyse data) to Symbolic-Data Analysis (the analysis of Symbolic Data). Official statistics appeared as a natural field of application for SDA methodologies, since studies in this area generally rely on aggregated data, also due to confidentiality issues which prevent the dissemination of individual data (microdata).

To describe groups of individuals or concepts, variables may now assume other forms of realizations, which allow taking into account the intrinsic variability. These new variable types have been called *symbolic variables*, and they may assume multiple, possibly weighted, values for each unit. As in the classical statistics framework, we are dealing here with random variables, which may be observed in a given population; the term "symbolic" is used to stress the fact that the values they take are of a different nature. In the next section, we define the (new) different variable types, providing illustrating examples in each case.

3 Types of Variables

To represent data variability, new variable types have been introduced in SDA, whose realizations are now not restricted to real values (in the numerical case) or individual categories (in the qualitative case). The different variable types, including the classical ones - which may be considered special cases of the symbolic types (see Brito (2014)) - are presented below.

As in classical Statistics, we distinguish numerical and categorical variables (Stevens (1946)). A numerical (or quantitative) variable is single valued (real or integer), as in the classical framework, if it takes one single value of an underlying domain for each unit. It is multi-valued if its values are finite subsets of the domain and it is an interval-valued variable if its values are intervals of $\mathbb{R}$. When a distribution over a set of sub-intervals is given, the variable is called a histogram-valued variable. A categorical (or qualitative) variable is single-valued (ordinal or

nominal), when it takes one category from a given finite category set $O = \{m_1, \ldots, m_k\}$ for each unit; multi-valued, if its values are finite non-empty subsets of O. A categorical modal variable Y with a finite domain $O = \{m_1, \ldots, m_k\}$ is a multi-valued variable where, for each unit, we are given a category set and, for each category m_ℓ, a frequency or probability which indicates how frequent or likely that category is for this unit Bock and Diday (2000).

Let $Y_1, \ldots, Y_p$ be the set of variables, O_j the underlying domain of Y_j and B_j the set where Y_j takes its value for each unit, for $j = 1, \ldots, p$. A description d is defined as a p-tuple $d = (d_1, \ldots, d_p)$ with $d_j \in B_j$, $j = 1, \ldots, p$.

Let $S = \{s_1, \ldots, s_n\}$ be the set of units under analysis (e.g. individuals or even classes of individuals) under analysis, then $Y_j(s_i) \in B_j$ for $j = 1, \ldots, p$; $i = 1, \ldots, n$. The data array to be analysed consists of n descriptions, one for each $s_i \in S : d_i = (Y_1(s_i), \ldots, Y_p(s_i))$, $i = 1, \ldots, n$.

3.1 Classical Variables

3.1.1 Quantitative Single-Valued Variables

Given the set of n units $S = \{s_1, \ldots, s_n\}$, a quantitative single-valued variable Y is defined by an application $Y : S \to O$ such that $s_i \longmapsto Y(s_i) = c \in O \subseteq I\!R$. This is the classical numerical case, and B is identical to the underlying set O, $B \equiv O$.

3.1.2 Categorical Single-Valued Variables

A categorical single-valued variable is a standard categorical variable. Given $S = \{s_1, \ldots, s_n\}$, and a finite set of categories, $O = \{m_1, \ldots, m_k\}$ a categorical single-valued variable is defined by an application $Y : S \to O$ such that $s_i \longmapsto Y(s_i) = m_\ell$ (i.e., in this case, again, $B \equiv O$). If the categories of O are naturally ordered, the variable is called ordinal, otherwise it is nominal. Such a categorical variable may be used to build new concepts or entities, by aggregating the cases sharing the same category.

3.2 New Variable Types

3.2.1 Quantitative Multi-Valued Variables

Given the set S, a quantitative multi-valued variable Y is defined by an application

$$\begin{aligned} Y : S &\longrightarrow B \\ s_i &\longmapsto Y(s_i) = \{c_{i1}, \ldots, c_{in_i}\} \end{aligned}$$

Here B is the power set of an underlying set $O \subseteq I\!R$ (excepting the empty set $\emptyset$). $Y(s_i)$ is now a finite non-empty set of real numbers.

3.2.2 Interval-Valued Variables

Given $S = \{s_1, \ldots, s_n\}$, an interval-valued variable is defined by an application

$$\begin{aligned} Y : S &\longrightarrow B \\ s_i &\longmapsto Y(s_i) = [l_i, u_i] \end{aligned}$$

B is the set of closed and bounded intervals of an underlying set $O \subseteq I\!R$. Let I be an $n \times p$ matrix representing the values of p interval-valued variables on S. Each $s_i \in S$ is represented by

a p-tuple of intervals, $I_i = (I_{i1}, \ldots, I_{ip})$, $i = 1, \ldots, n$, with $I_{ij} = [l_{ij}, u_{ij}]$, $j = 1, \ldots, p$ (see Table 2).

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

Table 2: Matrix I of interval data.

The value of an interval-valued variable Y_j for each $s_i \in S$ is usually defined by the lower and upper bounds l_{ij} and u_{ij} of $I_{ij} = Y_j(s_i)$. For modelling purposes, however, it may be useful to represent $Y_j(s_i)$ by the midpoint $c_{ij} = (l_{ij} + u_{ij})/2$ and range $r_{ij} = u_{ij} - l_{ij}$ of I_{ij}.

Example:

Consider a dataset containing information about arriving flights at some airports; Table 3 presents data for three airports. In airport A, for instance, the number of passengers in arriving flights ranged from 150 to 200, the number of codesharing companies involved in each flight was either 1 or 2. Here, the number of passengers is an interval-valued variable whereas the number of codesharing companies involved is a multi-valued quantitative variable. A similar description may be obtained for the remaining airports. It should be stressed that in this example the units under analysis are the airports, for each of which we have aggregated information, and not the individual flights.

Airport	Passengers	Companies
A	$[150, 200]$	$\{1, 2\}$
B	$[180, 300]$	$\{1, 2, 3\}$
C	$[200, 400]$	$\{1, 3\}$

Table 3: Data for airports (1).

$\bowtie$

3.2.3 Histogram-Valued Variables

When real-valued data are aggregated by means of intervals, the information on the distribution inside the intervals is not taken into account. One way to keep more detailed information is to define sub-intervals between the global lower (LB) and upper (UB) bounds and compute frequencies for these intervals. We obtain for each case a histogram[1] with k classes (and k frequencies) where k is the number of the considered sub-intervals. Naturally, to aggregate numerical microdata by means of a histogram implies that a reasonably large number of observations are available at the micro level.

Given $S = \{s_1, \ldots, s_n\}$, a histogram-valued variable is defined by an application

[1] We use here the term "histogram" in an informal way, to denote the empirical distribution over a set of sub-intervals, although this does not correspond to the statistical formal definition of a histogram.

$$Y : S \longrightarrow B$$
$$s_i \longmapsto Y(s_i) = \{\left[\underline{I}_{i1}, \overline{I}_{i1}\right[, p_{i1}; \left[\underline{I}_{i2}, \overline{I}_{i2}\right[, p_{i2}; \ldots; \left[\underline{I}_{ik_i}, \overline{I}_{ik_i}\right], p_{ik_i}\}$$

where $I_{i\ell} = \left[\underline{I}_{i\ell}, \overline{I}_{i\ell}\right[$ or $\left[\underline{I}_{i\ell}, \overline{I}_{i\ell}\right]$, $\ell = 1, \ldots, k_i$ are the sub-intervals considered for observation s_i, $\underline{I}_{i\ell} = \overline{I}_{i\ell-1}, \ell = 2, \ldots, k_i$ and $p_{i1} + \ldots + p_{ik_i} = 1, i = 1, \ldots, n$. B is now the set of all possible partitions of any compact of $\mathbb{R}$ and all possible distributions over the (finite set of) corresponding sub-intervals. It is assumed that for each unit s_i values are uniformly distributed within each sub-interval. For different observations, the number and length of sub-intervals considered may naturally be different.

When $k = 1$ a histogram reduces to an interval: interval-valued variables may therefore be considered special cases of histogram-valued variables.

Example:

Consider again the airports example, with a new variable which records the delay (in minutes) of each arriving flight. In this case, information is recorded for three time lengths (0 to 10 minutes, 10 to 30 minutes, 30 minutes to one hour), the corresponding variable is therefore a histogram-valued variable - see Table 4.

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	$[200, 400]$	$\{1, 3\}$	$\{[0, 10[, 0.75; [10, 30[, 0.20; [30, 60], 0.05\}$

Table 4: Data for airports (2).

$\bowtie$

3.2.4 Categorical Multi-Valued Variables

A categorical multi-valued variable is defined by an application

$$Y : S \to B$$

where B is the set of non-empty subsets of $O = \{m_1, \ldots, m_k\}$. The "values" of $Y(s_i)$ for $s_i \in S$ are now finite sets of categories.

3.2.5 Categorical Modal Variables

A categorical modal variable Y with a finite domain $O = \{m_1, \ldots, m_k\}$ is a multi-valued variable where, for each unit, we are given a category set and, for each category m_ℓ, a weight, frequency or probability p_ℓ which indicates how frequent or likely that category is for this unit. In this case, B is the set of distributions (probability, frequency, or other) over O, and its elements are denoted $\{m_1(p_1), \ldots, m_k(p_k)\}$.

Example:
Consider again the airports example, and the information on the companies' shares of arriving flights. We have then a categorical modal variable, as shown in Table 5.

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)}

Table 5: Data for airports (3).

⋈

Categorical modal variables are similar to histogram-valued variables for the quantitative case, in that their values are both characterized by classes or categories and weights. In SDA, "distributional data" refers to both types, as opposed to "set-valued" variables, when no distribution is given. Nevertheless, from a mathematical point of view, they are of different nature.

4 Analysis of Symbolic Data

To represent data taking into account variability intrinsic to each observation, new variable types have been defined, whose values assume new forms. As expected, definitions of basic statistical notions do not apply automatically, and well established properties are no longer straightforward. To apply statistical and multivariate data analysis techniques to symbolic data then requires proper consideration, and often the design of appropriate tools.

Consider the case of numerical variables, where the evaluation of dispersion is a central question and the consequences of different possible choices in the design of multivariate methods has to be addressed. Also, many multivariate methodologies are defined by linear combinations of the descriptive variables, and on the properties of dispersion measures under linear transformations. The question then arises of how should a linear combination of symbolic numerical variables be defined, and which properties remain valid.

Different approaches have been considered by various authors to address these and other issues and propose symbolic extensions of multivariate data analysis methods. Most existing methods for the analysis of such data still rely on non-parametric descriptive approaches.

Interval-valued data is the most investigated case for which more methods have been developed. Those methods follow two distinct approaches within a non-parametric framework. These consist either in (i) assuming a distribution, usually the Uniform, within each observed interval, derive sample moments from this assumption (see Bertrand and Goupil (2000), Billard and Diday (2003)), and design methods based on such moments (Billard and Diday (2000)) or (ii) represent an interval by two real numbers, the lower and upper bounds or the midpoint and (half) range, and propose methods using these two values. These are usually exploratory approaches, relying on distance-based criteria - see, e.g., De Carvalho *et al* (2006); Chavent *et al* (2006), Duarte Silva and Brito (2006), Neto and De Carvalho (2008), Neto and De Carvalho (2010).

The statistical analysis of distributional data, and in particular of histogram-valued data, has received more attention in the past few years. The approaches developed rely on assuming a Uniform distribution within each sub-interval of each observed histogram. The proposed methods are either based on sample moments derived from such assumption or on the representation of the histograms by the associated quantile functions, for which appropriate distances are considered,

which allow defining appropriate criteria - see Irpino and Verde (2006, 2008), Brito and Chavent (2012), Dias and Brito (2015).

However, for the specific case of interval-valued variables, probabilistic approaches have been proposed and investigated (see Brito and Duarte Silva (2012), Neto *et al* (2011), Le-Rademacher and Billard (2011)), opening new paths: statistical modelling of symbolic variables then allows for estimation and hypothesis testing.

In the next section we present the parametric model for interval-valued variables proposed by Brito and Duarte Silva (2012), and show how it opens the way to multivariate parametric analysis of interval data. In Section 6 we focus on the representation of histogram data by means of the associated quantile functions. Considering now interval-valued data in this framework, we obtain the very first results deduced for this type of variables as a special case. Non-parametric exploratory methods based on this representation are presented, which may be applied to data of either (or both) types.

5 Modelling and Analysing Interval Data

5.1 Parametric Models for Interval Data

Let $S = \{s_1, \ldots, s_n\}$ be the set of n units under analysis. Consider that each $s_i \in S$ is represented by a p-dimensional vector of intervals, $(I_{i1}, \ldots, I_{ip}), i = 1, \ldots, n$, with $I_{ij} = [l_{ij}, u_{ij}], j = 1, \ldots, p$, as in Table 2.

Brito and Duarte Silva (2012) proposed parametric models for interval data, relying on Multivariate Normal or Skew-Normal distributions for the MidPoints C, with $c_{ij} = \dfrac{l_{ij} + u_{ij}}{2}$, and Log-Ranges $R^* = \ln R$ with $r_{ij} = u_{ij} - l_{ij}$, of the interval-valued variables.

The Gaussian model consists in assuming a multivariate Normal distribution for the MidPoints C and the logs of the Ranges, R^*, with mean vector $\mu = [\mu_C^t \ \mu_{R^*}^t]^t$ and covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{pmatrix}$ where μ_C and μ_{R^*} are p-dimensional column vectors of the mean values of, respectively, the MidPoints and Log-Ranges, and $\Sigma_{CC}, \Sigma_{CR^*}, \Sigma_{R^*C}$ and $\Sigma_{R^*R^*}$ are $p \times p$ matrices with their variances and covariances.

This model has the advantage that it allows for a straightforward application of classical multivariate methods. It is important to keep in mind, however, that the MidPoint c_{ij} and the Log-Range $r^*_{ij} = \ln(r_{ij})$ of the value of an interval variable $I_{ij} = Y_j(s_i)$ are related to the same variable, and must therefore be considered together. The link that may exist between MidPoints and Log-Ranges of the same or different interval-valued variables should be taken into account by appropriate configurations of the global covariance matrix. Intermediate parameterizations between the non-restricted and the non-correlation setup considered for real-valued data are therefore relevant for the specific case of interval data.

The most general formulation allows for non-zero correlations among all MidPoints and Log-Ranges (configuration C1); in another setup, interval variables Y_j are independent, but for each variable, the MidPoint may be correlated with its Log-Range (configuration C2); a third situation allows for MidPoints (respectively, Log-Ranges) of different variables to be correlated, but no correlation between MidPoints and Log-Ranges is allowed (configuration C3); finally, all MidPoints and Ranges are uncorrelated, both among themselves and between each other (configuration C4). Table 6 summarizes the different considered configurations. We note that from the Normality assumption it follows that, in this particular framework, imposing non-correlations with Log-Ranges is equivalent to imposing non-correlations with Ranges.

Configuration	Characterization	Σ
C1	Unrestricted	Unrestricted
C2	Y_j's independent	$\Sigma_{CC}, \Sigma_{CR^*} = \Sigma_{R^*C}$, $\Sigma_{R^*R^*}$ all diagonal
C3	C's uncorrelated with R^*'s	$\Sigma_{CR^*} = \Sigma_{R^*C} = 0$
C4	All C's and R^*'s are uncorrelated	Σ diagonal

Table 6: Different cases for the variance-covariance matrix.

It should be remarked that for configurations C2, C3 and C4, Σ can be written as a block diagonal matrix, after a possible rearrangement of rows and columns.

The Gaussian model has many advantages, which explains its generalized use in multivariate data analysis; in particular, it allows for a direct modelling of the covariance structure between the variables. Nevertheless it does present some limitations, namely the fact that it imposes a symmetrical distribution on the MidPoints and a specific relation between mean, variance and skewness for the Ranges. A more general model that overcomes these limitations may be obtained by considering the family of Skew-Normal distributions (see, for instance, Azzalini (1985); Azzalini and Dalla Valle (1996)). The Skew-Normal generalizes the Gaussian distribution by introducing an additional shape parameter, while trying to preserve some of its mathematical properties.

The density of a p-dimensional Skew-Normal distribution is given by

$$f(y;\alpha,\xi,\Omega) = 2\phi_p(y-\xi;\Omega)\Phi(\alpha^t\omega^{-1}(y-\xi)), y \in I\!R^p \tag{2}$$

where ξ and α are p-dimensional location and shape parameter vectors respectively, Ω is a symmetric $p \times p$ positive-definite matrix, ω is a diagonal matrix formed by the square-roots of the diagonal elements of Ω, ϕ_p is the density of a $N_p(0,\Omega)$ and Φ is the distribution function of a standard Gaussian variable.

Notice that the Skew-Normal model encompasses mixed models with marginal Normal random variables, for which the corresponding shape parameter is null.

The mean, variance-covariance matrix and skewness coefficients of a p dimensional Skew-Normal distribution are given by (Azzalini (2005)):

$$\mu = E(Y) = \xi + \omega\mu_Z \tag{3}$$

$$\Sigma = Var(Y) = \Omega - \omega\mu_Z\mu_Z^t\omega \tag{4}$$

$$\gamma_{1,j} = \frac{E[(Y_j - E(Y_j))^3]}{Var(Y_j)^{3/2}} = \frac{4-\pi}{2}\frac{\mu_{Z;j}^3}{(1-\mu_{Z;j}^2)^{3/2}} \quad , \; j = 1,\ldots,p \tag{5}$$

where μ_Z is a vector of expected values for standard Skew-Normal variables, which are defined by

$$\mu_Z = \sqrt{\frac{2}{\Pi}}\delta \text{ with } \delta = \frac{\omega^{-1}\Omega\omega^{-1}\alpha}{\sqrt{(1+\alpha^t\omega^{-1}\Omega\omega^{-1}\alpha)}}.$$

As an alternative to the Gaussian model, it may be considered that (C, R^*) follow jointly a $2p$-multivariate Skew-Normal distribution, for which the different alternative configurations of

the Σ matrix may be assumed. Given (4), a null covariance $\Sigma(j,j')$ implies that $\Omega(j,j') = \Omega(j,j)^{\frac{1}{2}}\mu_{Z_j}\Omega(j',j')^{\frac{1}{2}}\mu_{Z_{j'}}$ or, equivalently $\Omega(j,j') = \frac{2}{\pi}\ \frac{1}{1+\alpha^t\omega^{-1}\Omega\omega^{-1}\alpha}\ \Omega_j^t\omega^{-1}\alpha\alpha^t\omega^{-1}\Omega_{j'}$ where Ω_j denotes the j^{th} column of Ω. This defines non-linear relations between the parameters in Ω and α.

5.1.1 Parameter Estimation

Gaussian Model

Let $Y_i = Y(s_i) = \left[{C_i}^t, {R_i^*}^t\right]^t$ be the $2p$ dimensional column vector comprising all the Mid-Points and Log-Ranges for unit s_i and $\bar{Y}$ be sample mean of the Y_i's. The maximum likelihood estimators of μ and Σ under the unrestricted configuration C1 are obviously the classical ones,

$$\hat{\mu} = \bar{Y} \tag{6}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})(Y_i-\bar{Y})^t := \frac{1}{n}E \tag{7}$$

For the restricted configurations, the maximum likelihood estimators of μ and Σ are obtained from the non-restricted estimators simply replacing by zeros the null parameters in the model for Σ (see Brito and Duarte Silva (2012)). For all configurations, the log-likelihood can be written as

$$\ln L(\mu,\Sigma) = -np\ \ln(2\pi) - \frac{n}{2}\ \ln\det\Sigma\ - \frac{1}{2}\ trE\Sigma^{-1} - \frac{n}{2}\left(\bar{Y}-\mu\right)^t\Sigma^{-1}\left(\bar{Y}-\mu\right) \tag{8}$$

Since Σ^{-1} is symmetric positive definite, the quadratic form term will be a minimum only when μ is equal to $\bar{Y}$, so that the maximum-likelihood estimate of the mean vector is always $\bar{Y}$, as usual. In the restricted configurations, Σ can be written as a block diagonal matrix, after a possible rearrangement of rows and columns. Then the maximum can be obtained by separately maximizing with respect to each block of Σ.

Skew-Normal Model

Azzalini and Capitanio (see, e.g., Azzalini and Capitanio (1999); Azzalini (2005)) have obtained the log-likelihood of a p-dimensional Skew-Normal distribution as

$$\ln L(\xi,\Omega,\alpha) = \ \text{constant}\ - \frac{1}{2}n\ln\det\Omega - \frac{n}{2}tr(\Omega^{-1}V) + \sum_i \zeta_0(\alpha^t\omega^{-1}(Y_i-\xi)) \tag{9}$$

where $V = n^{-1}\sum_i(Y_i-\xi)(Y_i-\xi)^t$ and $\zeta_0(x) = ln(2\Phi(x))$. The maximization of (9) is performed in two steps by defining a new parameter, $\eta = \alpha^t\omega^{-1}$, and separating the maximization on ξ and η from the maximization on Ω given ξ, which has the analytical solution $\Omega = V$.

The optimal likelihood solution for the Skew-Normal model with restricted configurations may not be obtained by simply replacing corresponding entries in the appropriate matrices, because of the non-linear relations between the parameters in Ω and α. For the Skew-Normal model with restricted configurations, we rely on Valle and Azzalini (2008) centred parametrization, which employs directly the parameters μ, Σ and γ_1 given by (3), (4) and (5). The log-likelihood is maximized with respect to μ, the free elements in Σ and γ. This may be done using a quasi-Newton numerical algorithm and the gradients derived by Valle and Azzalini (2008).

Given an interval-valued data set, the choice among the different models and covariance configurations may be based on usual information criteria, such as the Bayesian Information Criteria (BIC) (Schwarz (1978)), or on pairwise likelihood ratio tests.

5.2 Multivariate Parametric Analysis of Interval Data

The models presented above allow for multivariate parametric analysis of interval data, by suitably extending and adapting the corresponding models for classical real-valued data. Analysis of Variance (Brito and Duarte Silva (2012)), Discriminant Analysis (Duarte Silva and Brito (2015)) and Model Based Clustering (Brito *et al* (2015)) have been addressed under this framework. The R-package MAINT.DATA (Duarte Silva and Brito (2017)), available on CRAN, allows modelling interval data under the proposed framework, providing functions and methods for parameter estimation, outlier detection, (M)ANOVA, discriminant analysis and model-based clustering.

5.2.1 ANOVA and MANOVA

ANOVA and MANOVA may be performed following a likelihood ratio approach.

Since each interval-valued variable Y_j is modelled by a pair $< C_j, R_j^* >$, an analysis of variance of Y_j is accomplished by a two-dimensional MANOVA of (C_j, R_j^*).

Let us assume a one-way design, where the single factor has k levels, and let n_ℓ be the number of units in group ℓ. Let again $Y_{ij} = \left[C_{ij}, R_{ij}^*\right]^t$ be the 2-dimensional column vector with the MidPoint and Log-Range of variable Y_j for unit s_i, let $\bar{Y}_{\bullet j\ell}$ and $\mu_{\bullet j\ell}$ be sample and population means of the Y_{ij}'s in group ℓ, and $\bar{Y}_{\bullet j\bullet}$ the corresponding global sample mean. The null hypothesis in this case consists in stating that all $\mu_{\bullet j\ell}$ are equal across groups.

Consider first the Gaussian model. For all covariance configurations, the likelihood ratio statistic is given by $\lambda = \left(\dfrac{\det E_{j,\text{alt}}}{\det E_{j,\text{null}}}\right)^{\frac{n}{2}}$ where $E_{j,\text{null}}$ and $E_{j,\text{alt}}$ are 2×2 matrices corresponding to the null and alternative hypothesis respectively. In the unrestricted case C1, these matrices are the classical ones, for the restricted covariance configurations, $E_{j,\text{null}}$ and $E_{j,\text{alt}}$ are obtained from their classical versions by replacing the null entries corresponding to each configuration.

In all cases, under the null hypothesis, $2 \ln \lambda$ follows asymptotically a Chi-square distribution with $n - k$ degrees of freedom.

For the Skew-Normal model, we need to maximize the log-likelihood for the null (mean vectors equal across groups) and the alternative hypothesis. Since no closed form is known for maximum likelihood estimates in this case, maximization has to be performed by numerical methods. For the unrestricted configuration C1, all covariance parameters are free, whereas for the restricted configurations, the corresponding ones are fixed to zero.

A simultaneous analysis of all the Y's interval-valued variables may be accomplished by a $2p$ dimensional MANOVA, following the same procedure.

5.2.2 Discriminant Analysis

The main goal of discriminant analysis is to obtain classification rules capable of assigning units of unknown origin to one of several well defined given groups, based on a vector of relevant attributes. The classical decision theoretic approach to this problem assumes that the attribute vectors follow some known distribution and derives an optimal rule that minimizes either the misclassification probability or the expected value of the misclassification cost.

Assume a problem with k groups, $G_\ell, \ell = 1, \ldots, k$ and denote the attribute vectors by $\mathbf{x}$, the *a priori* group membership probabilities by π_ℓ and the within group probability or density function by $f_\ell(\mathbf{x})$.

Assuming equal misclassification cost across groups, it is well-known that the optimal rule assigns an unit to the group G_ℓ for which $\pi_\ell \times f_\ell(\mathbf{x})$ is maximal - see, e.g., McLachlan (1992). These rules are usually expressed in terms of unknown parameters, that in practice must be estimated from observations with known group membership.

When f_ℓ is a Gaussian density, and the covariance matrices are equal across groups, the approach described above results in a linear classification rule, whereas when covariance matrices differ from group to group, a quadratic classification rule is obtained.

Consider the Gaussian model for interval data. Then, for each covariance configuration, an estimate of the optimum classification rule can be obtained by direct generalisation of the classical linear (10) and quadratic (11) discriminant classification rules, given by

$$\Gamma = \arg\max_{\ell}(\hat{\mu_\ell}^t \hat{\Sigma}^{-1} Y - \frac{1}{2}\hat{\mu_\ell}^t \hat{\Sigma}^{-1}\hat{\mu_\ell} + \ln\ \hat{\pi_\ell}) \tag{10}$$

$$\Gamma = \arg\max_{\ell}(-\frac{1}{2} Y^t \hat{\Sigma_\ell}^{-1} Y + \hat{\mu_\ell}^t \hat{\Sigma_\ell}^{-1} Y + \ln\ \hat{\pi_\ell} - \frac{1}{2}(\ln\det\hat{\Sigma_\ell} + \hat{\mu_\ell}^t \hat{\Sigma_\ell}^{-1}\hat{\mu_\ell})) \tag{11}$$

where $\Gamma \in \{1, \ldots, k\}$ denotes the group assignments, ℓ is a group index, $\hat{\mu_\ell}, \hat{\Sigma}, \hat{\Sigma_\ell}$ and $\hat{\pi_\ell}$ are the maximum likelihood estimates of $\mu_\ell, \Sigma, \Sigma_\ell$ and π_ℓ for the corresponding covariance configurations.

For the Skew-Normal model different alternatives may be considered: the groups differ only in terms of the location parameter ξ; the groups differ in terms of both ξ and Ω; the groups differ in terms of ξ, Ω and the shape parameter α.

We consider a Location Model in which the groups differ only in terms of the location parameter ξ, and a General Model, where the groups differ in terms of all parameters. The corresponding classification rules are, respectively,

$$\Gamma = \arg\max_{\ell}(\hat{\xi_\ell}^t \hat{\Omega}^{-1} Y - \frac{1}{2}\hat{\xi_\ell}^t \hat{\Omega}^{-1}\hat{\xi_\ell} + \ln\ \hat{\pi_\ell} + \zeta_0(\hat{\alpha}^t \hat{\omega}^{-1}(Y - \hat{\xi_\ell}))) \tag{12}$$

and

$$\Gamma = \arg\max_{\ell}(-\frac{1}{2} Y^t \hat{\Omega_\ell}^{-1} Y + \hat{\xi_\ell}^t \hat{\Omega_\ell}^{-1} Y + \ln\ \hat{\pi_\ell} - \frac{1}{2}(\ln\ \det\hat{\Omega_\ell} + \hat{\xi_\ell}^t \hat{\Omega_\ell}^{-1}\hat{\xi_\ell}) + \zeta_0(\hat{\alpha_\ell}^t \hat{\omega_\ell}^{-1}(Y - \hat{\xi_\ell}))) \tag{13}$$

where $\xi_\ell, \Omega, \Omega_\ell, \alpha, \alpha_\ell$ and location, scale, association and shape parameters (see Azzalini and Capitanio (1999)), ω and ω_ℓ are the square-root of the diagonal elements of the matrices Ω and Ω_ℓ, and $\zeta_0(x) = \ln(2\Phi(x))$.

5.2.3 Model-Based Clustering

Model-based clustering considers the data as arising from a distribution that is a mixture of two or more components (Banfield and Raftery (1993); McLachlan and Peel (2000)). Each component is described by a density function and has an associated probability or "weight" in the mixture. Typically it is assumed that components are p-variate Normal distributions, thus, the probability model for clustering will often be a finite mixture of multivariate Normals. Each component in the mixture will then be called a cluster. The problem then consists in estimating

the model parameters for each component, as well as the membership probabilities of each unit. To this purpose, the Expectation-Maximization (EM) algorithm is commonly used. This is an iterative method to find maximum likelihood estimates, when the model depends on unobserved variables - in this case the component membership probabilities. The method alternates between an expectation (E) step, which finds the expectation of the log-likelihood at the current parameter estimates, and a maximization (M) step, which estimates parameters maximizing the expected log-likelihood found on the E step.

Model-based clustering of interval data may be addressed by considering the Gaussian model presented above. For that purpose, the EM algorithm has been adapted to the likelihood maximization in our models, for different covariance configurations.

In model-based clustering of interval data, $Y_i = [C_i, R_i^*]$ is defined as the $2p$ dimensional vector comprising all the MidPoints and Log-Ranges for s_i, and the "complete" data are considered to be $v_i = (y_i, z_i)$, where $z_i = (z_{i1}, \ldots, z_{ik})$ is assumed as the "missing" data, with $z_{i\ell} = 1$ if $s_i \in$ cluster ℓ and $z_{i\ell} = 0$ otherwise. In the unrestricted case the M-step formulas for $\hat{\Sigma}$, $\hat{\Sigma}_\ell$ are the classical ones; for the restricted configurations $\hat{\Sigma}$ and $\hat{\Sigma}_\ell, \ell = 1, \ldots, k$ are obtained maximizing the likelihood for each block separately (see Brito and Duarte Silva (2012)).

For the selection of the appropriate model and the number of components k, the Bayesian Information Criterion (BIC) (Schwarz (1978)) may be used.

6 Analysis of Histogram Data

Given $S = \{s_1, \ldots, s_n\}$, if Y is a histogram-valued variable then

$$Y(s_i) = \{[\underline{I}_{i1}, \overline{I}_{i1}[\,, p_{i1}; [\underline{I}_{i2}, \overline{I}_{i2}[\,, p_{i2}; \ldots; [\underline{I}_{ik_i}, \overline{I}_{ik_i}]\,, p_{ik_i}\}, \; i = 1, \ldots, n$$

where $I_{i\ell} = [\underline{I}_{i\ell}, \overline{I}_{i\ell}[$ or $[\underline{I}_{i\ell}, \overline{I}_{i\ell}]$, $\ell = 1, \ldots, k_i$ are the sub-intervals considered for observation s_i, $\underline{I}_{i\ell} = \overline{I}_{i\ell-1}, \ell = 2, \ldots, k_i$ and $p_{i1} + \ldots + p_{ik_i} = 1$.

The values of a histogram-valued variable may equivalently be represented by the empirical distribution function or by its inverse, the quantile function. This latter option is often used, given that all quantile functions are defined on the same domain $[0, 1]$, which is convenient for comparisons and multivariate analysis.

Assuming a Uniform distribution within each subinterval $I_{i\ell}$, the quantile function associated with a histogram-valued observation $Y(s_i)$ is a piecewise linear function given by :

$$\Psi(t) = F^{-1}(t) = \begin{cases} \underline{I}_{i1} + \frac{t}{w_{i1}} r_{i1} & \text{if} \quad 0 \leq t < w_{i1} \\ \underline{I}_{I_{i2}} + \frac{t - w_{i1}}{w_{i2} - w_{i1}} r_{i2} & \text{if} \quad w_{i1} \leq t < w_{i2} \\ \vdots & \\ \underline{I}_{ik_i} + \frac{t - w_{ik_i - 1}}{1 - w_{ik_i - 1}} r_{ik_i} & \text{if} \quad w_{in_i - 1} \leq t \leq 1 \end{cases} \tag{14}$$

where $w_{il} = \sum_{\ell=1}^{l} p_{i\ell}$ if $l = 1, \ldots, k_i$ and $r_{i\ell} = \overline{I}_{i\ell} - \underline{I}_{i\ell}$ with $\ell \in \{1, \ldots, k_i\}$; k_i is the number of sub-intervals in $Y(s_i)$.

Example:

Consider the data in Table 4 and the delay distribution for airport C : $\{[0, 10[\,, 0.75; [10, 30[\,, 0.20; [30, 60]\,, 0.05\}$. The associated quantile function is

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

⋈

As noted above, interval-valued variables may be considered as a particular case of histogram-valued ones, when only one interval is allowed for in each observation, with weight equal to one. If Y is an interval-valued variable, then $Y(s_i) = [l_i, u_i] = \{[l_i, u_i], 1\}$. If a Uniform distribution is assumed in each observed interval, the quantile function associated with $Y(s_i)$ is a linear function $\Psi(t) = F^{-1}(t) = l_i + t(u_i - l_i)$, $t \in [0, 1]$. Notice however, that for interval-valued variables other distributions may be considered within the observed intervals, e.g., the Triangular distribution - see Dias and Brito (2017); Cheira *et al* (2017); Malaquias (2017).

6.1 Descriptive Statistics for Histogram-Valued Variables

Assuming a Uniform distribution within each sub-interval of $Y(s_i)$, $i = 1, \ldots, n$, $I_{i\ell} = [l_{i\ell}, u_{i\ell}]$, $\ell = 1, \ldots k_i$, we may derive sample moments of a histogram-valued variable. Billard and Diday (2003) obtained the symbolic sample mean

$$\overline{Y} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{\ell=1}^{k_i} [(l_{i\ell} + u_{i\ell}) p_{i\ell}] = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{k_i} (c_{i\ell}\, p_{i\ell}) \tag{15}$$

and the symbolic sample variance

$$\begin{aligned} S_Y^2 =& \frac{1}{3n} \sum_{i=1}^{n} \sum_{\ell=1}^{k_i} \left[(l_{i\ell}^2 + l_{i\ell} u_{i\ell} + u_{i\ell}^2) p_{i\ell}\right] - \overline{Y}^2 = \\ =& \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{k_i} \frac{(u_{i\ell} - l_{i\ell})^2}{12} p_{i\ell} + \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{k_i} \left(\frac{l_{i\ell} + u_{i\ell}}{2} - \overline{Y}\right)^2 p_{i\ell} \end{aligned} \tag{16}$$

Billard and Diday (2003) also obtained a formula for the covariance between two histogram-valued variables from the empirical joint density function:

$$Cov_1(Y_j, Y_{j'}) = \frac{1}{4n} \sum_{i=1}^{n} \left[\sum_{\ell 1=1}^{k_{ij}} p_{ij\ell 1} (l_{ij\ell 1} + u_{ij\ell 1}) \sum_{\ell 2=1}^{k_{ij'}} p_{ij'\ell 2} (l_{ij'\ell 2} + u_{ij'\ell 2}) \right] - \overline{Y_j}\, \overline{Y_{j'}} \tag{17}$$

Later, Billard (2008) derived a different formula, considering a decomposition into Within observations Sum of Products (WithinSP) and Between observations Sum of Products (BetweenSP):

$$\begin{aligned} Cov_2(Y_j, Y_{j'}) =& \underbrace{\frac{1}{n} \sum_{i=1}^{n} \sum_{\ell 1=1}^{k_{ij}} \sum_{\ell 2=1}^{k_{ij'}} p_{ij\ell 1} p_{ij'\ell 2} \frac{(u_{ij\ell 1} - l_{ij\ell 1})(u_{ij'\ell 2} - l_{ij'\ell 2})}{12}}_{\text{WithinSP}} + \\ &+ \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left[\sum_{\ell 1=1}^{k_{ij}} p_{ij\ell 1} \Big(\frac{l_{ij\ell 1} + u_{ij\ell 1}}{2} - \overline{Y_j} \Big) \sum_{\ell 2=1}^{k_{ij'}} p_{ij'\ell 2} \Big(\frac{l_{ij'\ell 2} + u_{ij'\ell 2}}{2} - \overline{Y_{j'}} \Big) \right]}_{\text{BetweenSP}} \end{aligned} \tag{18}$$

For the particular case of interval-valued variables, and assuming a Uniform distribution within each observed interval, we obtain the symbolic sample mean and variance as previously derived by Bertrand and Goupil (2000):

$$\overline{Y} = \frac{1}{2n}\sum_{i=1}^{n}(l_i + u_i) = \frac{1}{n}\sum_{i=1}^{n} c_i \tag{19}$$

$$S_Y^2 = \frac{1}{3n}\sum_{i=1}^{n}(l_i^2 + l_i u_i + u_i^2) - \overline{Y}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(u_i - l_i)^2}{12} + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{l_i + u_i}{2} - \overline{Y}\right)^2 \tag{20}$$

which is the sum of the variances of the Uniform distributions assumed within each observed interval, plus the variance of the intervals' midpoints.

Billard and Diday (2003)) derived a formula for the covariance between two interval-valued variables from the empirical joint density function:

$$Cov_1(Y_j, Y_{j'}) = \frac{1}{4n}\sum_{i=1}^{n}(l_{ij} + u_{ij})(l_{ij'} + u_{ij'}) - \overline{Y_j Y_{j'}} \tag{21}$$

Billard (2008) then obtained

$$Cov_2(Y_j, Y_{j'}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{(u_{ij} - l_{ij})(u_{ij'} - l_{ij'})}{12}}_{\text{WithinSP}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{l_{ij} + u_{ij}}{2} - \overline{Y_j}\right)\left(\frac{l_{ij'} + u_{ij'}}{2} - \overline{Y_{j'}}\right)}_{\text{BetweenSP}} \tag{22}$$

6.2 Distance Measures for Histogram-Valued Variables

Many distance measures to compare distributions have been proposed in the literature - see e.g., Bock and Diday (2000); Gibbs and Su (2002). Here we focus on specific cases of interest : the Euclidean distance, and distances based on the quantile function representation.

Let, as above, $Y(s_i) = \{[\underline{I}_{i1}, \overline{I}_{i1}[, p_{i1}; [\underline{I}_{i2}, \overline{I}_{i2}[, p_{i2}; \ldots; [\underline{I}_{ik_i}, \overline{I}_{ik_i}], p_{ik_i}\}$ be the "observation" of histogram-valued variable Y at $s_i \in S$.

Assuming that both histograms are defined on a fixed partition (same subintervals, $I_1, \ldots, I_k$) the (squared) Euclidean distance between two histogram observations $Y(s_i), Y(s_{i'})$ compares the respective weights:

$$D_E^2(Y(s_i), Y(s_{i'})) = \sum_{\ell=1}^{k}(p_{i\ell} - p_{i'\ell})^2 \tag{23}$$

The Wasserstein and the Mallows distances compare the quantile functions associated with the histograms, the former using a L_1 absolute value approach and the latter a L_2 approach.

The Wasserstein distance is defined as

$$D_W(Y(s_i), Y(s_{i'})) = D_W(\Psi_i, \Psi_{i'}) = \int_0^1 |\Psi_i(t) - \Psi_{i'}(t)|\, dt \tag{24}$$

while the Mallows distance is given by

$$D_M(Y(s_i), Y(s_{i'})) = D_M(\Psi_i, \Psi_{i'}) = \sqrt{\int_0^1 (\Psi_i(t) - \Psi_{i'}(t))^2 dt} \tag{25}$$

Under the uniformity hypothesis, and considering a fixed weight decomposition (same weights, different intervals), we have (Irpino and Romano (2007)):

$$D_M^2(Y(s_i), Y(s_{i'})) = D_M^2(\Psi_i, \Psi_{i'}) = \sum_{\ell=1}^{k} p_\ell \left[(c_{i\ell} - c_{i'\ell})^2 + \frac{1}{3}(r_{i\ell} - r_{i'\ell})^2 \right] \tag{26}$$

For the particular case of interval-valued observations (only one (sub)-interval for each observation), we obtain

$$D_M^2(Y(s_i), Y(s_{i'})) = D_M^2(\Psi_i, \Psi_{i'}) = (c_i - c_{i'})^2 + \frac{1}{3}(r_i - r_{i'})^2 \tag{27}$$

i.e., the squared Mallows distance between two observed intervals is defined as the squared Euclidean distance between the intervals' midpoints plus one third of the squared Euclidean distance between the intervals' half-ranges. For real-valued data, the ranges are null, and the (squared) Mallows distance coincides with the (squared) Euclidean distance.

Using a metric-based approach, Irpino and Verde (2015) obtained basic statistics for histogram-valued variables, with the interval-valued ones as special case. The Fréchet Mean, or *barycenter*, is defined by

$$M = \arg\min_x \sum_{i=1}^{n} w_i d^2(s_i, x) \tag{28}$$

When based on the Euclidean distance, the mean distribution or barycenter of a family of distributions is the finite uniform mixture of the given distributions.

For the Mallows distance, the barycenter is obtained from the mean quantile function, the Mallows *barycentric histogram* is the solution of the minimization problem

$$\min_{\Psi_b(t)} \quad \sum_{i=1}^{n} D_M^2(\Psi_i(t), \Psi_b(t)) \tag{29}$$

and it is defined by the quantile function where the centers and half ranges of each subinterval ℓ are the classical mean of the centers and half ranges of all observations.

Given a partition of S in K groups $C_1, \ldots, C_K$, the Mallows distance fulfils the *Huygens theorem* decomposition in Between and Within dispersion, as relates to the barycenters (as defined in (28) and (29)) (Irpino and Verde (2006)):

$$\begin{aligned} &\sum_{i=1}^{n} D_M^2(\Psi_i(t), \overline{\Psi_S}(t)) = \\ &\sum_{h=1}^{k} n_h D_M^2(\overline{\Psi_S}(t), \overline{\Psi_{C_h}}(t)) + \sum_{h=1}^{k} \sum_{s_i \in C_h} D_M^2(\Psi_i(t), \overline{\Psi_{C_h}}(t)) \end{aligned} \tag{30}$$

where n_h is the number of units in group C_h, $\overline{\Psi_S}(t)$ is the quantile function of the barycentric histogram in S and $\overline{\Psi_{C_h}}(t)$ is the quantile function of the barycentric histogram in $C_h, h = 1, \ldots, K$.

Given its (good) properties, the Mallows distance is the basis of many data analysis approaches for histogram-valued data.

6.3 Multivariate Analysis of Histogram Data

Several methods for multivariate analysis of histogram-valued data have been developed which rely on the representation of the histogram observations by quantile functions. We present here approaches developed for linear regression and clustering.

6.3.1 Linear Regression of Histogram Data

Dias and Brito (2015) proposed the *Distribution and Symmetric Distribution* Linear Regression model, where the distributions taken by the histogram-valued variables are represented by their quantile functions.

The space of quantile functions is a semi-linear space: the sum of two quantile functions is still a quantile function, but the product of a quantile function by a scalar is a quantile function if and only if the scalar is non-negative (for negative scalars we obtain a decreasing function, which cannot be a quantile function). This implies that a regression model may not be defined, as in the classical case, by a linear combination with real coefficients.

The solution proposed for this problem relies in considering a model with two terms for each independent variable X_j: the quantile function that represents the observed distribution(histogram) the variable takes, $\Psi_{X_j(s_i)}(t)$ and the quantile function that represents the distribution (histogram) of the respective symmetric histogram-valued variable, $\Psi_{\widetilde{X}_j(s_i)}(t)$. The obtained quantile function $\Psi_{\widehat{Y}(s_i)}(t)$ is then given by:

$$\Psi_{\widehat{Y}(s_i)}(t) = \gamma + \alpha_1 \Psi_{X_1(s_i)}(t) + \beta_1 \Psi_{\widetilde{X}_1(s_i)}(t) + \ldots + \alpha_p \Psi_{X_p(s_i)}(t) + \beta_p \Psi_{\widetilde{X}_p(s_i)}(t) \tag{31}$$

with $\alpha_j, \beta_j \geq 0$, $j = 1, 2, \ldots, p$ and $\gamma \in I\!R$.

Although non-negativity restrictions on the parameters are imposed, this does not imply a direct linear relationship because the model includes both the quantile functions that represent the distributions taken by the histogram-valued variables and the quantile functions that represent the distributions taken by the respective symmetric histogram-valued variables. Determination of the model requires solving a quadratic optimization problem, subject to non-negativity constraints on the unknowns. The Mallows distance is used to quantify the error, i.e., the difference between the observed and the predicted quantile function of $Y(s_i)$. The parameters of the model are an optimal solution of the minimization problem:

$$\text{Minimize} \qquad SE = \sum_{i=1}^{n} D_M^2(\Psi_{Y(s_i)}, \Psi_{\widehat{Y}(s_i)}) \tag{32}$$

$$\text{with } \alpha_j, \beta_j \geq 0,\ j = 1, 2, \ldots, p \text{ and } \gamma \in I\!R.$$

The Karush Kuhn Tucker optimality conditions allow defining a measure to evaluate the quality of fit of the model, with values in the unit interval, similarly to the coefficient of determination in classical linear regression for real-valued data.

The model has also been studied and applied to real problems in the particular case of interval-valued variables (Dias and Brito (2017)), considering either Uniform of Symmetric Triangular distributions within the observed and predicted intervals.

Along similar lines, Irpino and Verde (2008) have developed a linear regression model for histogram-valued data which minimizes the Mallows distance between the observed quantile function of the dependent variable, and the one derived from the linear model. The proposed

method relies in the exploitation of the properties of a decomposition of the Mallows distance by Irpino and Romano (2007); this is used to measure the sum of squared errors and rewrite the model splitting the contribution of the predictors in a part depending on the averages of the distributions and another depending on the centred quantile distributions.

6.3.2 Clustering of Histogram Data

The Huygens decomposition for the Mallows distance referred above (see Section 6.2) allows for the extension to histogram-valued data of clustering methods which use criteria based on the within and between cluster dispersion measured by a quadratic distance.

Brito and Chavent (2012) proposed a method for divisive clustering of histogram and interval data. The method provides a hierarchy on a set of objects together with a conjunctive characterization of each formed cluster. Starting from the (full) set under analysis, the method proceeds by performing a bipartition of one cluster at each step. At step m a partition of S in m clusters is present, one of which will be further divided in two sub-clusters; the cluster to be divided and the splitting rule are chosen to obtain a partition in $m+1$ clusters minimizing intra-cluster dispersion

$$Q(m) = \sum_{h=1}^{k} \sum_{s_i, s_{i'} \in C_h^{(m)}} D^2(s_i, s_{i'}) \tag{33}$$

with

$$D^2(s_i, s_{i'}) = \sum_{j=1}^{p} d^2(Y_j(s_i), Y_j(s_{i'})) \tag{34}$$

Here d must be quadratic distance between distributions, allowing for the Huygens decomposition, as the Mallows distance (see (25)). The bipartition to be performed at each step is defined by one single variable, considering conditions of the type $R_{j\ell} := Y_j \leq \overline{I}_{j\ell}, \ell = 1, \ldots, K_j - 1, j = 1, \ldots, p$. Then, sub-cluster 1 will consist of those elements $s_i \in S$ who verify condition $R_{j\ell}:= Y_j(s_i) \leq \overline{I}_{j\ell}$ and sub-cluster 2 of those $s_{i'}$ who do not $:Y_j(s_{i'}) > \overline{I}_{j\ell}$. It is considered that $s_i \in S$ verifies the condition $Y_j(s_i) \leq \overline{I}_{j\ell}$ if and only if $\sum_{\alpha=1}^{\ell} p_{ij\alpha} \geq 0.5$. The sequence of such conditions constitutes a necessary and sufficient condition for cluster membership; therefore the obtained clustering is *monothetic*: each cluster is represented by a conjunction of properties in the descriptive variables.

Exploring the same decomposition, Irpino and Verde developed Ward hierarchical clustering (Irpino and Verde (2006)) and dynamical clustering (Verde and Irpino (2007); Irpino and Verde (2008)) approaches for histogram data, treating interval data as a special case.

7 Concluding Remarks and Perspectives

Symbolic Data Analysis provides a framework where the variability observed may effectively be considered in the data representation, and methods be developed that take that variability into account. This approach is of particular and growing interest in the analysis of huge sets of data, recorded in very large databases, when the units of interest are not the individual records (the microdata), but rather some second-level entities. The multivariate statistical analysis of symbolic data, however, raises new problems, as intervals and empirical distributions are not real numbers: classical concepts do not apply directly, and usual properties on which established methods rely cannot be taken for granted.

Different approaches have been pursued to appropriately take the variability inherent to the data into account in the modelling and analysis. Parametric approaches, allowing for inferential studies and hypotheses testing have only been proposed for interval-valued variables, and are based on the decomposition of the intervals in two quantities, usually midpoints and ranges, to which concepts designed for real-valued data may apply. The consideration of a distribution within observed intervals (other than the Uniform) raises new problems, for which only preliminary approaches have been developed. This leads to the analysis of histogram-valued data, so far only considered from a non-parametric point of view. Moreover, all work to this day has been done on the basis of marginal empirical distributions, i.e., the empirical distribution for each variable is considered separately. An important effort will be necessary to go one step further, and consider the joint observed distributions in the data representation and analysis. Finally, kernel density estimation may be applied to the empirical distributions, leading to density-valued variables. This will be another fascinating line of research in the future of Symbolic Data Analysis.

References

Azzalini, A. (1985). A class of distributions which includes the Normal ones. *Scandinavian Journal of Statistics*, 12, 171–178.

Azzalini, A. (2005). The Skew-Normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32, 159–188.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate Skew-Normal distribution. *J. R. Statist. Soc. B*, 61(3), 579–602.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate Skew-Normal distribution. *Biometrika*, 83(4), 715–726.

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803–821.

Bertrand, P. and Goupil, F.(2000). Descriptive statistics for symbolic data. In: *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, H.-H. Bock and E. Diday (Eds.), 106–124, Springer-Verlag, Berlin-Heidelberg.

Billard, L. (2008). Sample covariance functions for complex quantitative data. In: *Proc. 4th World Conference of the International Association of Statistical Computing*, 157–163, Yokohama, Japan.

Billard, L. and Diday, E. (2000). Regression analysis for interval-valued data. In: *Data Analysis, Classification, and Related Methods*, Proc. Seventh Conference of the International Federation of Classification Societies (IFCS00), 369-374, Springer.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association* 98(462), 470-487.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chichester.

Bock, H.-H. and Diday, E. (Editors) (2000). *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin-Heidelberg.

Brito P. (2014). Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, 4(4), 281–295.

Brito, P. (1995). Symbolic Objects: Order structure and pyramidal clustering. *Annals of Operations Research*, 55, 277-297.

Brito, P. and Chavent, M. (2012). Divisive monothetic clustering for interval and histogram-valued data. In: *Proc. ICPRAM 2012 - 1st International Conference on Pattern Recognition Applications and Methods*, Vilamoura, Portugal.

Brito, P. and Diday, E. (1990). Pyramidal representation of symbolic objects. In: *Knowledge, Data and Computer-Assisted Decisions*, M. Schader and W. Gaul (Eds.), NATO ASI Series, Springer Verlag, Berlin-Heidelberg-New York, 3-16.

Brito P. and Duarte Silva A.P. (2012). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, 39(1), 3–20.

Brito, P., Duarte Silva A.P. and Dias, J.G. (2015). Probabilistic clustering of interval data. *Intelligent Data Analysis*, 19(2), 293-313.

Chavent, M., Lechevallier, Y. and Verde, R. (2006). New clustering methods for interval data. *Computational Statistics*, 21(2), 211-229.

Cheira, P., Brito, P. and Duarte Silva, A.P. (2017). Factor Analysis of Interval Data. arXiv:1709.04851 [stat.ME]. Web address: `http://arxiv.org/abs/1709.04851`.

De Carvalho, F.A.T., Brito, P. and Bock,H.-H. (2006). Dynamic clustering for interval data based on L_2 distance. *Computational Statistics*, 21(2), 231-250.

Dias, S. and Brito, P. (2015). Linear Regression model with histogram-valued variables. *Statistical Analysis and Data Mining*, 8(2), 75-113.

Dias, S. and Brito, P. (2017). Off the beaten track: A new linear model for interval data. *European Journal of Operational Research*, 258(3), 1118-1130.

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis: the basic choices. In: *Classification and Related Methods of Data Analysis, Proc. of the Conference of the International Federation of Classification Societies IFCS'87*, H.-H. Bock (Ed.), 673-684, North Holland, Amsterdam.

Diday, E. and Noirhomme-Fraiture, M. (Editors) (2008). *Symbolic Data Analysis and the Sodas Software*, Wiley, Chichester.

Duarte Silva, A.P. and Brito, P. (2006). Linear discriminant analysis for interval data. *Computational Statistics*, 21(2), 289-308.

Duarte Silva, A.P. and Brito, P. (2017). MAINT.Data - Modelling and Analysing Interval Data. *R* Package, version 1.2.1., available on CRAM at `https://cran.r-project.org/web/packages/MAINT.Data/index.html`.

Duarte Silva A.P. and Brito P. (2015). Discriminant analysis of interval data: an assessment of parametric and distance-based approaches. *Journal of Classification*, 32(3), 516–541.

Gibbs, A. and Su, F. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70(3), 419–435.

Irpino, A. and Romano, E. (2007). Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. In: M. Noirhomme and G. Venturini (Eds.), *Revue des Nouvelles Technologies de l'Information*, RNTI-E-9, Cépadués-Éditions, 99–110.

Irpino, A. and Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: *Data Science and Classification, Proc. of the Conference of the International Federation of Classification Societies (IFCS)*, V. Batagelj, H.-H. Bock, A. Ferligoj and A. Žiberna (Eds.), 185–192, Springer-Verlag, Berlin-Heidelberg.

Irpino, A. and Verde, R. (2008). Dynamic clustering of interval data using a Wasserstein-based distance. *Pattern Recognition Letters*, 29(11), 1648–1658.

Irpino, A. and Verde, R. (2015). Basic statistics for distributional symbolic variables: a new metric-based approach. *Advances in Data Analysis and Classification*, 9(2), 143-175.

Le-Rademacher, J. and Billard, L. (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. *Journal of Statistical Planning and Inference*, 141, 1593–1602.

Malaquias, P. (2017). *Modelo de Regressão Linear para Variáveis Intervalares: Uma Extensão do Modelo ID*. Master Dissertation in Data Analytics, Faculdade de Economia, Universidade do Porto.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Neto, E.A.L. and De Carvalho, F.A.T. (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, 52(3), 1500-1515.

Neto, E.A.L. and De Carvalho, F.A.T. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, 54(2), 333-347.

Neto, E.A.L., Cordeiro, G.M. and De Carvalho, F.A.T. (2011). Bivariate symbolic regression models for interval-valued variables. *Journal of Statistical Computation and Simulation* 81(11), 1727–1744.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: Symbolic Data Analysis. *Statistical Analysis and Data Mining*, 4(2), 157–170.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680.

Valle, R.B. and Azzalini, A. (2008). The centred parametrization for the multivariate Skew-Normal distribution. *Journal of Multivariate Analysis*, 99(4), 1362–1382.

Verde, R. and Irpino, A. (2007). Dynamic clustering of histogram data: using the right metric. In: *Selected Contributions in Data Analysis and Classification*, P. Brito (Ed.), 123-134, Springer, Berlin, Heidelberg.