Identifying Special Structures in Interval-Data via Model-Based Clustering

Paula Brito[1,4], A. Pedro Duarte Silva[2], and José G. Dias[3]
[1]FEP & LIAAD INESC TEC, Universidade do Porto, Porto, PORTUGAL
[2]FEG & CEGE, Universidade Católica Portuguesa at Porto, Porto, PORTUGAL
[3]ISCTE - Instituto Universiário de Lisboa, BRU-IUL, Lisboa, PORTUGAL
[4]Corresponding author: Paula Brito, e-mail: mpbrito@fep.up.pt

Abstract

In this paper we present a model-based approach to the clustering of interval data building on recently proposed parametric models. These methods consider configurations for the variance-covariance matrix that take the nature of the interval data directly into account. The proposed framework relies on parametrizations considering the inherent variability of the relevant data units and the relation that may exist between this variability and the corresponding value levels. Using both synthetic and real data sets the pertinence of the proposed methodology is shown, as the method effectively selects heterocedastic models with restricted covariance structures when they are the most suitable, even in situations with limited information. Moreover, considering special configurations of the variance-covariance matrix, adapted to nature of interval data, proves to be the adequate approach. The presented study also makes clear the need to consider both the information about position (conveyed by the MidPoints) and intrinsic variability (conveyed by the Log-Ranges) when analysing interval data.

Keywords: cluster, finite mixture models, interval-valued variable, intrinsic variability, symbolic data

1. Introduction

Symbolic Data Analysis (see, e.g., Bock and Diday (2000), Diday and Noirhomme (2008), Noirhomme and Brito (2011)) is a research area focusing on the analysis of complex data with intrinsic variability, which needs to be explicitly taken into account. Such data occur, in particular, when huge data bases are aggregated on the basis of some descriptors defining groups of interest - the new statistical units. This also occurs when the entities to be analysed are not single elements, but classes or concepts, for instance, not a particular flight, but the airport traffic, not the particular flower I am picking, but the flower species. The alternative of representing variable values by central measures like averages, medians or modes often leads to an unacceptable loss of information. New variable types were introduced that allow for the representation of the inherent variability of the data.

As in the classical case, symbolic variables may be either numerical or categorical. A numerical variable is single-valued if it takes one single value of an underlying domain for each entity (this is the classical case). 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 an empirical distribution over a set of subintervals is given, the variable is called a histogram-valued variable. As in the classical context, data are presented in a data-array, now called a “symbolic data table”, each row of which corresponds to a group, or concept, i.e., the entity of interest, and each column corresponds to a “symbolic variable”.

In Brito and Duarte Silva (2012) parametric models for interval data are proposed, which rely on Multivariate Normal or Skew-Normal distributions for the MidPoints and Log-Ranges of the interval-valued variables. A key feature of that proposal is the parametrization of the variance-covariance matrix that takes into account the particular nature of interval data.

In this paper we propose a model-based approach for clustering interval data, using the Gaussian models proposed in Brito and Duarte Silva (2012) in the model-based clustering context. The performance of the proposed methodology is illustrated with synthetic and real data sets.

In recent years finite mixture models have been extensively used as a clustering technique. It is assumed that there is discrete population heterogeneity with K subpopulations or clusters that can be un-mixed. Since each cluster or component is characterized by a specific density function, this approach has been called model-based clustering.

Clustering of interval data has been addressed by several authors, using various methodologies. Methods based on dissimilarities, generally adaptations of k-means, have been developed, as well as their fuzzy extensions. Other proposals use Kohonen maps, Poisson processes, or monothetic approaches. For a survey, see Noirhomme and Brito (2011). However, none of these has taken a model based approach.

The remaining of the paper is organized as follows. In Section 2 interval-valued variables are formally introduced and different representation of interval-data are considered. Parametric modelling of interval data, which will be used in the sequel, is recalled. Section 3 describes the proposed methodology for clustering interval data; Section 4 illustrates the procedure using a synthetic data set. Section 5 reports the application of the method to real datasets with different characteristics. The article ends by highlighting the main conclusions, advantages of this model-based clustering method, and desirable further extensions.

2. Interval data

Interval data occur in various contexts. When describing ranges of variable values, as for daily stock prices, we obtain interval data; in the aggregation of huge data bases into groups, real values describing the individual observations (the *microdata*) lead to intervals describing the groups formed. Let $S = \{s_1, \dots, s_n\}$, be the set of n entities under analysis. Formally, an interval-valued variable is defined by an application $Y : S \to T$ such that $s_i \to Y(s_i) = [l_i, u_i]$ where T is the set of 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-dimensional vector of intervals, $I_i = (I_{i1}, \dots, I_{ip}), i = 1, \dots, n$, with $I_{ij} = [l_{ij}, u_{ij}], j = 1, \dots, p$. The value of an interval-valued variable Y_j for each $s_i \in S$ is naturally defined by the lower and upper bounds l_{ij} and u_{ij} of $I_{ij} = Y_j(s_i)$. For modelling purposes an alternative parameterization consisting in representing $Y_j(s_i)$ by the MidPoint $c_{ij} = \dfrac{l_{ij} + u_{ij}}{2}$ and Range $r_{ij} = u_{ij} - l_{ij}$ of I_{ij} may also be useful.

In Brito and Duarte Silva (2012), a parametric modelling for interval data is proposed, considering multivariate Normal or Skew-Normal distributions for the MidPoints and Log-Ranges of the interval-valued variables. The Gaussian model consists in assuming a multivariate Normal distribution for MidPoints C and the logs of the Ranges R, $R^* = ln(R), (C, R^*) \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_C^t, \boldsymbol{\mu}_{R^*}^t\right]^t$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{CC} & \boldsymbol{\Sigma}_{CR^*} \\ \boldsymbol{\Sigma}_{R^*C} & \boldsymbol{\Sigma}_{R^*R^*} \end{pmatrix}$ where $\boldsymbol{\mu}_C$ and $\boldsymbol{\mu}_{R^*}$ are p-dimensional column vectors of the mean values of, respectively, the MidPoints and Log-Ranges, and $\boldsymbol{\Sigma}_{CC}, \boldsymbol{\Sigma}_{CR^*}, \boldsymbol{\Sigma}_{R^*C}$ and $\boldsymbol{\Sigma}_{R^*R^*}$ are $p \times p$ matrices with their variances and covariances.

One advantage of the Gaussian model is that it allows for the application of classical inference methods; however it is important to keep in mind that the MidPoint c_{ij} and the Range r_{ij} of the value of an interval-valued variable $I_{ij} = Y_j(s_i)$ are two quantities related to one same variable, and must therefore be considered together. Therefore, the global covariance matrix should take into account the link that may exist between MidPoints and Ranges of the same or different variables. In this paper, we shall consider the following parameterizations:

1. Non-restricted case: allowing for non-zero correlations among all MidPoints and Log-Ranges;

2. Interval-valued variables Y_j are independent, but for each variable, the MidPoint may be correlated with its Log-Range: $\boldsymbol{\Sigma}_{CC}, \boldsymbol{\Sigma}_{CR^*} = \boldsymbol{\Sigma}_{R^*C}, \boldsymbol{\Sigma}_{R^*R^*}$ all diagonal;

3. MidPoints (Log-Ranges) of different variables may be correlated, but no correlation between MidPoints and Log-Ranges is allowed: $\boldsymbol{\Sigma}_{CR^*} = \boldsymbol{\Sigma}_{R^*C} = \mathbf{0}$;

4. All MidPoints and Log-Ranges are uncorrelated, both among themselves and between each other: $\boldsymbol{\Sigma}$ diagonal.

From the Gausian assumption it obviously follows that imposing non-correlations with Log-Ranges is equivalent to imposing non-correlations with Ranges. It should be remarked that in Cases 2, 3 and 4, $\boldsymbol{\Sigma}$

can be written as a diagonal by blocks matrix, after a possible rearrangement of rows and columns. This is particularly important for maximum likelihood estimation.

3. Model-based clustering of interval data

The finite mixture model with K components for $\mathbf{x}_i = (x_{i1},...,x_{i,2p})$ is defined by $f(\mathbf{x}_i;\boldsymbol{\varphi}) = \sum_{k=1}^{K} \tau_k f_k(\mathbf{x}_i;\boldsymbol{\theta}_k)$, where component proportions τ_k are positive and sum to one; and $\boldsymbol{\theta}_k$ denotes parameters of the conditional distribution of cluster k. Model parameters are $\boldsymbol{\varphi} = (\boldsymbol{\tau},\boldsymbol{\theta})$, with $\boldsymbol{\tau} = (\tau_1,...,\tau_{K-1})$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_K)$. The number of free parameters in vectors $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ are $d_{\boldsymbol{\tau}} = K-1$ and $d_{\boldsymbol{\theta}}$, respectively. The number of free parameters is $d_{\boldsymbol{\varphi}} = d_{\boldsymbol{\tau}} + d_{\boldsymbol{\theta}}$.

For continuous metric data, finite mixtures of Gaussian distributions have been extensively applied (McLachlan and Peel (2000)). For this specification, the conditional distribution is given by $\mathcal{N}(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix, respectively. For instance, heteroscedastic Case 1 contains $d_{\boldsymbol{\varphi}} = Kp(2p+3)+K-1$ free parameters.

Maximum likelihood (ML) parameter estimation involves the maximization of the log-likelihood function: $\ell(\boldsymbol{\varphi};\mathbf{x}) = \sum_{i=1}^{n} \ln f(\mathbf{x}_i;\boldsymbol{\varphi})$, a problem that can be tackled by the Expectation-Maximization (EM) algorithm (Dempster *et al* (1977)). E-step computes the joint conditional distribution of the missing data given observed data and provisional estimates of model parameters. In the M-step, standard complete data ML methods are used to update the unknown parameters using an expanded data matrix with the estimated densities of the missing data (posterior cluster probabilities) as weights. In our implementation, to try avoiding local optima, each search of the EM algorithm is replicated many times from different starting points.

An important modelling issue is the selection of the number of components (K). We use the Bayesian Information Criterion (BIC) (Schwarz (1978)) given by BIC$= -2\ell(\hat{\boldsymbol{\varphi}};\mathbf{x}) + d_{\boldsymbol{\varphi}} ln(n)$, where $d_{\boldsymbol{\varphi}}$ is the number of free parameters in the model. We notice that BIC is a consistent criterion, whereas the AIC is a biased estimate of the true number of components (see Hurvich and Tsai (1989), Dias (2006)).

4. A synthetic data set example

To illustrate model-based clustering of interval-value data, we use a synthetic data set. As its structure is known, we can further understand the performance of the proposed method. The data set contains 1000 observations (n) and two variables (p). We assume three components (K) with sizes $\boldsymbol{\tau} = (0.5, 0.3, 0.2)$. This synthetic data set is simulated under Case 2, i.e., Midpoints and Log-Ranges are associated only within each variable. The setting assumes homoscedasticy, i.e., clusters share the same covariance matrix. Thus, conditional on the cluster (k), the simulated values are given by $\mathbf{X}_i|k \sim \mathcal{N}(\boldsymbol{\mu}_k,\boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -5 \\ -0.5 \\ -5 \\ -0.5 \end{bmatrix}, \ \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ -0.5 \\ 0 \\ -0.5 \end{bmatrix}, \ \boldsymbol{\mu}_3 = \begin{bmatrix} 5 \\ -0.5 \\ 5 \\ -0.5 \end{bmatrix}, \ and\ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.6 & - & - \\ 0.6 & 1 & - & - \\ - & - & 1 & 0.6 \\ - & - & 0.6 & 1 \end{bmatrix}.$$

Figure 1 depicts the first 50 observations of this data set and the centroids of the three components.

As competing models to the data generating process described above, we allow different homoscedastic configurations (Cases 1, 3, and 4) under the same number of components (K). BIC identifies the correct configuration, i.e., Case 2 with minimum value of BIC.

Model estimates show that the component sizes are well retrieved, i.e., proportions estimates ($\hat{\tau}_k$) are 0.535, 0.289, and 0.176. Mean vector and covariance matrix estimates are:

$$\hat{\boldsymbol{\mu}}_1 = \begin{bmatrix} -4.94 \\ -0.43 \\ -4.94 \\ -0.52 \end{bmatrix}, \ \hat{\boldsymbol{\mu}}_2 = \begin{bmatrix} -0.01 \\ -0.55 \\ 0.03 \\ -0.52 \end{bmatrix}, \ \hat{\boldsymbol{\mu}}_3 = \begin{bmatrix} 4.95 \\ -0.53 \\ 4.97 \\ -0.52 \end{bmatrix}, \ \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.97 & 0.60 & - & - \\ 0.60 & 0.96 & - & - \\ - & - & 1.00 & 0.62 \\ - & - & 0.62 & 1.03 \end{bmatrix}.$$

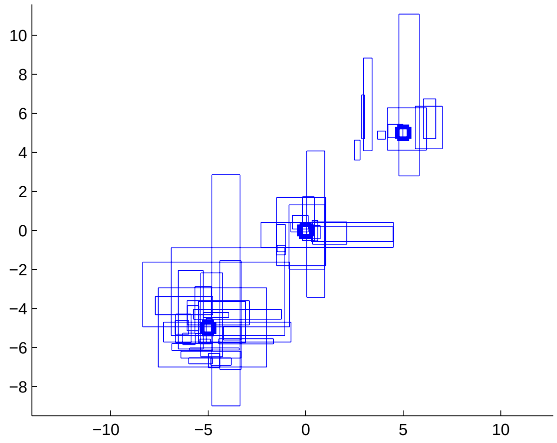

Figure 1: First 50 observations and three-cluster means (Synthetic data set)

Overall, this model-based clustering method retrieves close approximations of the true values.

5. Applications

5.1 Income-debt data

We applied our method to survey data included as sample in the SPSS package (named “customer.dbase”). Among the large set of variables available, we selected income and debt variables: Household Income (HI), Debt to Income Ratio (X 100) (DIR), Credit Card Debt (CCD) and Other Debts (OD); DIR is expressed as a percentage, while HI, CCD and OD are measured in thousand US dollars. The 5000 individual observations have been aggregated on the basis of Gender (F, M), Age Category (18-24, 25-34, 35-49, 50-64, more than 65 years old), Level of education (did not complete high school, high-school degree, some college, college degree, post-undergraduate degree), and Job Category (managerial and professional, sales and office, service, agricultural and natural resources, precision production, craft, repair, operation, fabrication, general labour), leading to 297 groups described by the intervals of observed values on the four considered variables. Table 1 shows the observed data for some of the groups.

Table 1: Income and Debt interval data

Group	HI	DIR	CCD	OD
Male, 8-24 High school degree, Service	$[15,61]$	$[0.1,23.4]$	$[0.0,6.57]$	$[0.02,7.71]$
Male, 35-49, College degree, Sales and Office	$[19,190]$	$[1.4,20.4]$	$[0.04,16.6]$	$[0.12,15.39]$
Female, 25-34, Some college Managerial and Professional	$[17,100]$	$[0.8,31.7]$	$[0.05,6.57]$	$[0.09,7.65]$

The lowest BIC value is observed for the solution in nine components, with a heterocedastic setup and Case 2, i.e., independent interval-valued variables. The analysis of component mean-values shows that in order to distinguish clusters, both Midpoints and Log-Ranges need to be considered. Furthermore, the estimated variance-covariance matrices are clearly different across groups, with noteworthy though different correlations between MidPoints and Log-Ranges of the same interval-valued variables. All these correlations are positive, showing that the higher the Midpoint of an interval-valued variable the higher the corresponding intrinsic variability.

5.2 Labour force survey data

This application concerns the Portuguese Labour Force Survey, from the 1^{st} semester of 2008. The quite large original micro data set comprehends a total of 42226 records. Only people who were unemployed at the time of the survey (had no job and were looking for one) were considered, i.e., 1540 cases, and we focused on the two following variables: activity time, in years (AT) and unemployment time, in months (UT). These micro data were gathered on the basis of Gender (M, F), Region (North, Centre, Lisbon and Tagus Valley (TV), South), Age-Group (15-24, 25-44, 45-64, over 65) and Education (Basic or less, Secondary, Higher), leading to 58 sociological groups, the statistical units of interest to be analysed.

The lowest BIC value is reached for the solution in five components, with a heterocedastic setup and Case 2, i.e., independent interval-valued variables.

In this application, where the number of observations is relatively low, a restricted though heterocedastic, model has been identified as best fit. This clearly illustrates the point that the method chose the best parameters for clustering, preferring a heterocedastic (and therefore heavier in the number of parameters) model to a "lighter" homocedastic one, but picking up a restricted configuration for the variance-covariance matrix, where interval-valued variables are assumed independent. Choosing Case 2 as opposed to Case 3, also means that correlation between the two parts of the interval-variables is considered more important than correlation between different variables. As in the previous application, components can only be separated by considering simultaneously Midpoints and Log-Ranges.

5.3 USA meteorological data

This dataset records temperatures and pluviosity measured in 282 meteorological stations in the USA. We consider the temperature ranges in January and July, and the annual pluviosity range measured in each station. All these values are based on 30 years averages (1970-2000). Data were retrieved from the USA National Environmental Satellite, Data and Information Service, at http://www1.ncdc.noaa.gov/pub/data/ccd-data (files nrmmin.txt, nrmmax.txt, nrmpct.txt); Temperatures are represented in the Fahrnheit scale, Pluviosity is measured in Inches.

The lowest BIC value is observed for the unrestricted (Case 1) heterocedastic solution with six natural clusters: Alaska, Pacific Coast, Arid Inland West, Northeast and Midwest, Southeast, Pacific Islands and Puerto Rrico. Clusters are differentiated not only by the MidPoint but also by the Log-Range variables, putting in evidence the importance of taking intrinsic variability of data into account. Moreover, clusters display highly different variances. In particular, the Alaska cluster presents very high variance for the January MidPoint variable, while Arid Inland West and Pacific Coast clusters present high variances for the July MidPoint variable; moreover, the Alaska cluster has a high variance for the Log-Range of the pluviosity and the Pacific Coast cluster for the Log-Range of the temperature in July. This stark difference illustrates well the need of a heterocedastic setup for these data.

For comparison purposes, Ward hierarchical clustering, using the Euclidean distances on standardized data, and the SCLUST algorithm from the *SODAS* package (see Diday and Noirhomme (2008)), which is based on the k-means methodology, using the Hausdorff distance to compare interval observations, have also been applied to this data set. The partitions obtained using the Ward method do not separate cold Alaska from warm Pacific-Islands and Puerto-Rico, while in the SCLUST partition the Alaska and the arid regions are well identified, but the remaining clusters are difficult to interpret, with the Pacific Islands and the Pacific Coast scattered by several groups. This may be explained by the fact that these methods somehow impose a similar covariance structure for all clusters, while some natural clusters, the Alaska one being the most obvious example, require larger variances than others.

6. Conclusion

In this paper we present a model-based approach to the clustering of interval data building on models developed in Brito and Duarte Silva (2012). The proposed framework relies on parametrizations that take into account the inherent variability of the relevant data units and the relation that may exist between this variability and the corresponding value levels. Using both synthetic data and empirical data sets the pertinence of the methodology proposed is demonstrated. In particular, the experiments show the flexibility of the model in identifing heterocedastic models; moreover, considering special configurations of the variance-covariance matrix, adapted to specific nature of interval data, proved to be the adequate approach. The presented study also made clear the need to consider both the information about position (conveyed by the MidPoints) and intrinsic variability (conveyed by the Log-Ranges) when analysing interval data.

Further research can compare this proposed framework with a multilevel setting as MidPoints and Log-Ranges are clustered within variables. Other research lines comprise the use of alternative models for interval data and the extension to other kinds of symbolic data.

Acknowledgements This work is partly funded by the ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within projects “FCOMP - 01-0124-FEDER-022701” and PEst-OE/EGE/UI0731/2011”.
One author would like to thank the Fundação para a Ciência e a Tecnologia (Portugal) for its financial support (Grant PTDC/EGE-GES/103223/2008).

References

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

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

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statistical Society Series B-Methodological*, Vol. 39, Issue 1: 1-38.

Dias, J.G. (2006). Latent class analysis and model selection, in ’*From Data and Information Analysis to Knowledge Engineering*’, M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, W. Gaul (eds.), Springer-Verlag, Berlin: 95-102.

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

Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika*, Vol. 76, Issue 2: 25-43.

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

Noirhomme-Fraiture, M. and Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis, *Statistical Analysis and Data Mining*, Vol. 4, Issue 2: 157-170.

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