An Innovative Framework for Analysing Official Statistics: Symbolic Data Analysis

Paula Brito[1], A. Pedro Duarte Silva[2]

[1] Faculdade de Economia, Universidade do Porto & LIAAD, INESC TEC, Porto, Portugal

[2] Universidade Católica Portuguesa, Católica Porto Business School & CEGE, Porto, Portugal

Abstract

Symbolic Data Analysis provides a framework for the representation and analysis of aggregated data, allowing keeping information about their intrinsic variability. In this work, we consider data where individual units, resulting from the aggregation of large amounts of microdata, are described by distributions of numerical variables. Each distribution is represented by a central statistic, and the logarithm of inter-quantile ranges, for a chosen set of quantiles. Multivariate Normal distributions are assumed for the whole set of indicators, considering alternative structures of the variance-covariance matrix. This framework is applied to the analysis of Portuguese Household Budget Survey data, where microdata relating to individual households are aggregated into groups based on location and income, resulting in distributions of the numerical variables describing expenses on different items. Model-based Clustering then allows obtaining a typology of the formed sociological groups.

Keywords: aggregated data, household budget survey, histogram data, model-based clustering, symbolic data

1. Introduction

In classical Statistics and Multivariate Data Analysis data is typically represented in a data array where each row represents a statistical unit, for which one single value is recorded for each variable. This representation model is, however, too restricted when the data to be analysed comprises variability. That is the case when the entities under analysis are not single elements, but groups formed from the aggregation of the original statistical units where the observed variability within each group should be explicitly considered. To this aim, new variable types have been introduced, whose realizations are not single real values or categories, but sets, intervals, or distributions over a given domain. Symbolic Data Analysis (see e.g. Diday & Noirhomme-Fraiture (2008); Brito (2014)) provides a framework for the representation and analysis of such complex data, taking into account their inherent variability.

This framework is of particular relevance in the analysis of official statistics, where the interest often lies in statistical units at a higher aggregated level, rather than in single individuals, and confidentiality issues prevent the dissemination and analysis of the microdata. Therefore, data should be aggregated at an appropriate level of granularity, to preserve confidentiality and allow for analysis at the level of interest. Furthermore, if microdata is aggregated into the same groups, this approach allows for the combination of independent surveys, which would not be possible at the individual (microdata) level.

In this work we focus on the Portuguese Household Budget Survey. Microdata relating to individual households are aggregated into groups based on location and income. In the resulting symbolic data, units are then described by distributions of numerical attributes.

We assume parametric models for numerical distributional variables based on the representation of each distribution by a central statistic, and the logarithm transformation of inter-quantile ranges, for a chosen set of quantiles. Multivariate Normal distributions are assumed for the whole set of indicators, considering alternative structures of the variance-covariance matrix. This model then allows for Model-based Clustering of the defined groups, identifying sociological clusters. The identified structure is mostly connected to location rather than to income level.

The remainder of the paper is organized as follows. In Section 2 we detail the representation model for distributional data, we describe the aggregation of the Portuguese Household Survey, and present our parametric model for numerical distributional variables. Section 3 recalls Model-based Clustering, and discusses the clustering obtained on the Portuguese Household Survey Data. Section 4 concludes the paper, opening avenues for future research.

2. Modelling the Household Survey Aggregated Data

2.1 Representation of Distributional Data

Let Y_1, …, Y_p be the p distributional variables, defined on a set of units $S=\{s_1, \ldots, s_n\}$. We consider that for each unit, the descriptive variables are (in general) not constant, but present some variability, and we assume that a set of quantiles, a probability distribution, or a sample, from which quantiles may be derived, are given. We represent the "values" of a numerical distributional variable by an ordered vector of quantiles $(\psi_{0i}, \psi_{1i}, \ldots, \psi_{qi})$, where ψ_{0i} and ψ_{qi} are typically either the minimum and maximum, respectively, or small and large quantiles suitably chosen in order to disregard severe outliers.

The proposed model consists in representing $Y_j(s_i)$ by

- A central statistic C_{ij}, typically the median Med_{ij} or the Midpoint $M_{ij} = \frac{\psi_{0ij}+\psi_{qij}}{2}$;
- the $[\psi_0, \psi_1[$ range: $R_{1ij} = \psi_{1ij} - \psi_{0ij}$
- the $[\psi_1, \psi_2[$ range: $R_{2ij} = \psi_{2ij} - \psi_{1ij}$
- …
- the $[\psi_{q-1}, \psi_q[$ range: $R_{qij} = \psi_{qij} - \psi_{(q-1)ij}$

2.2 The Household Survey Data

This study concerns the Portuguese Household Budget Survey (HBS), analysing data from 2015, which is the most recent available. The considered original microdata consist of the expenses, for each household, on the following items: (i) Food products and non-alcoholic beverages; (ii) Clothing and footwear; (iii) Housing, water, electricity, gas, and other fuels; (iv) Home accessories, household equipment, and routine household maintenance; (v) Health; (vi) Transport; (vii) Communications; (viii) Leisure, recreation, and culture; (ix) Restaurants and hotels; (x) Miscellaneous goods and services.

For each item, we registered the proportion of the corresponding expense on the total household expenses. These data were then aggregated on the basis of

- Income class - considering 20 classes, based on equally-spaced quantiles
- Region - NUTS 2 (North, Centre, Lisbon Metropolitan Area, Alentejo, Algarve, Madeira, Azores)
- Type of area - Predominantly Rural (PRA), Medi-Urban (MUA), Predominantly Urban (PUA)

leading to 20 × 7 × 3 = 420 groups.

Each group is described by the distribution of each of the ten variables. Noting that for several variables there are many zeros at microdata level (sometimes extending beyond the 0.30 quantile), and that we observed upper outliers, we chose to represent each distribution by the corresponding Median, the Minimum, and the 0.40, 0.60, 0.80, and 0.99 quantiles. Table 1 shows the distribution of variable Food for a few groups.

Table 1: Partial view of distributional variable Food.

Group	Food
MU-North-IncQnt3	0.2369 ; {[0.00,0.22[,0.4; [0.22,0.24[,0.2; [0.24 , 0.28[,0.2; [0.28, 0.42],0.19}
MU-North-IncQnt4	0.2379 ; {[0.00,0.17[,0.4; [0.17,0.24[,0.2; [0.24 , 0.30[,0.2; [0.30, 0.62],0.19}
...	...
PUA-Madeira-IncQnt20	0.0980 ; {[0.04,0.09[,0.4; [0.09,0.10[,0.2; [0.10 , 0.13[,0.2; [0.13, 0.25],0.19}

2.3 Parametric Models for Distributional Data

The proposed model consists in assuming that the joint distribution of the central statistic C and the logarithms of the ranges $\mathrm{R}_\ell^* = \ln(\mathrm{R}_\ell), \ell = 1, \dots, \mathrm{q}$ is multivariate Normal:

$$\left(C, R_1^*, \ldots, R_q^*\right) \sim N_{q+1}(\mu, \Sigma)$$

where μ is the (q+1) dimensional mean vector and Σ is the (q+1)×(q+1) dimensional variance-covariance matrix.

In the most general formulation (configuration 1) we allow for non-zero correlations among all centres and log-ranges; for distributional variables there are however other cases of interest: the distributional-valued variables Y_j are non-correlated, but for each variable, the centre and all its log-ranges may be correlated among themselves (configuration 2); centres (respectively, log-ranges) of different variables may be correlated, but no correlation between centres and log-ranges is allowed (configuration 3); centres (respectively, each log-range) of different variables may be correlated, but no correlation between centres and log-ranges or between non-corresponding log-ranges is allowed (configuration 4); and, finally, all centres and log-ranges are non-correlated (configuration 5).

3. Analysis of the HBS Distributional Data

3.1 Model-based Clustering

To organize the sociological groups in a clustering structure, and then characterize the obtained typology, we resort to Model-based Clustering (Banfield & Raftery (1993); Fraley & Raftery (2002); McLachlan & Peel (2000)).

Model-based Clustering assumes the data arises from a distribution that is a mixture of several components. Each component is considered as a cluster, it is characterised by a conditional density/mass function, and has an associated probability or “weight”:

$$f(x_i, \varphi) = \sum_{h=1}^{k} \pi_h \, f_h(x_i, \Theta_h)$$

Here k is the number of components, π_h is the “weight” of component h, with all $\pi_h > 0$ and $\pi_1 + \ldots + \pi_k = 1$ *;* and f_h is the conditional distribution in component h, with parameters gathered in Θ_h. When the conditional probability is the multivariate Gaussian density, the probability model for clustering is a finite mixture of multivariate Normals, known as the Gaussian Mixture Model. In this case, Θ_h consists of the mean vectors and variance-covariance matrices of the descriptive variables – in our context, of all medians and log-ranges that describe the distributions of the original variables in each formed group.

To estimate the parameters in Θ_h as well as membership (posterior) probabilities of each unit, maximum likelihood estimation is employed, leading to the maximization of the log-likelihood function. This is usually done by the Expectation-Maximization (EM) algorithm (Dempster *et* *al.* (1977)). To avoid local optima each search of the EM algorithm is replicated from different starting points.

The selection of the model – in our case, the appropriate configuration of the variance-covariance matrices, and whether they are constant across components (homoscedastic model) or different in each component (heteroscedastic model) – and of the number of components (k), we employ the Bayesian Information Criterion (BIC). The BIC aims at identifying a solution which maximizes likelihood, but penalizes the number of parameters involved to avoid overfitting. For further details, see Brito *et al.* (2015).

3.2 Clustering the HBS data

The Model-based Clustering presented above was applied to the aggregated Portuguese household survey data described in Section 2.2. However, 183 of the original 420 groups lead to degenerate intervals in some variables and were discarded, resulting in a final data set with 287 groups and 10 distributional variables. The minimum value of the BIC was achieved by a heteroscedastic model with four components and diagonal variance-covariance matrices (configuration 5). We note that in this application, given the relatively large dimensionality of the unrestricted covariance matrices (a 50 by 50 matrix for each component) a parsimonious model was recommended, and the data suggested that taking into account different variances across components was more important than possible correlations between variable indicators. The groups are relatively well balanced across components, with 60 groups (20.9%) in component CP1, 67 (23.6%) in component CP2, 85 (29.2%) in component CP3, and 75 (26.3%) in CP4. The components differ by area type and region but not much by income classes. Tables 2 and 3 display the component compositions across area types and regions. Figure 1 shows a parallel coordinate plot of the means across components.

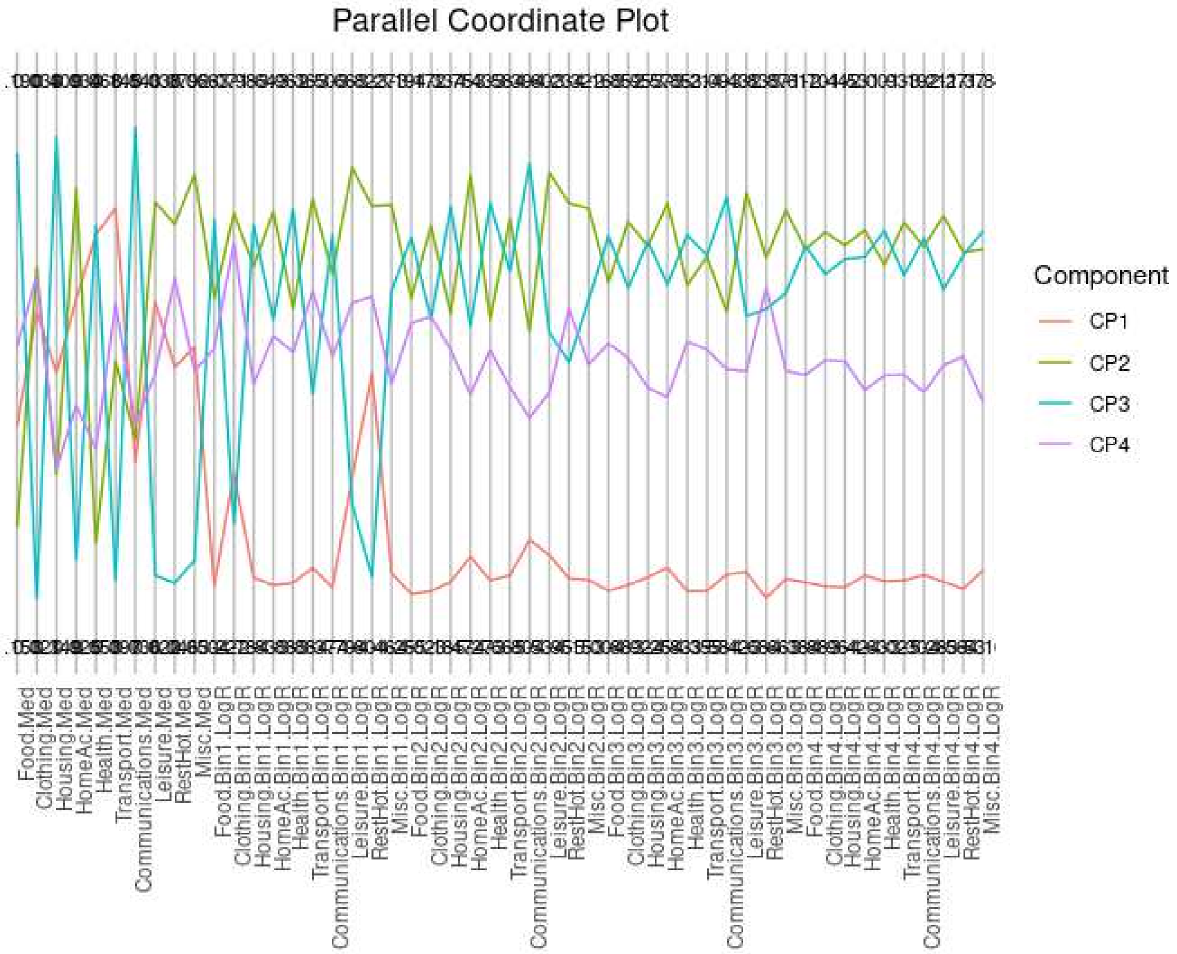
Table 2: Component composition across area types.

Component	Predominantly Urban	Medi-Urban	Predominantly Rural
CP1	0.4833	0.5167	0.0000
CP2	0.0896	0.0000	0.9104
CP3	0.1765	0.1882	0.6353
CP4	0.4667	0.5067	0.0267

Table 3: Component composition across regions.

Component	North	Centre	Lisbon MA	Alentejo	Algarve	Azores	Madeira
CP1	0.0833	0.0500	0.3833	0.0667	0.2000	0.0000	0.2167
CP2	0.2239	0.1791	0.1493	0.0746	0.1493	0.0746	0.1493
CP3	0.1647	0.2353	0.0941	0.2000	0.1059	0.0941	0.1059
CP4	0.2267	0.1333	0.0040	0.2133	0.2000	0.1333	0.0533

Figure 1: Parallel coordinate plot of the means across components.



We note that in addition to the region and area types, the components are mostly distinguished by the within group variable variability, and not so much by their medians. In particular:

- Component CP1 consists mostly of groups from urban areas, from Lisbon Metro Area, Algarve, and Madeira, and shows low overall variation, with high medians on transport expenses and negative skewness on leisure expenses.
- Component CP2 is dominated by groups from mainly rural areas; in this component, variables have an overall high variation, and relatively high median on home accessories, leisure, and restauration & hotels expenses.
- In component CP3 63% of the groups are from rural areas, mainly from the North, Centre, and Alentejo regions. In this component the variables display an overall high variation, with relatively high medians on food, housing and communications expenses.
- Finally, component CP4 is formed mostly by groups from urban areas, except the Lisbon Metro Area and Madeira; variables are characterized by an overall medium variation.

4. Conclusion and Perspectives

In this paper we have proposed a novel framework for the analysis of official aggregated data, relying on their empirical distribution, rather than on single central statistics. Appropriate parametric models are considered for the numerical distributional data, allowing for their multivariate analysis. Portuguese Household Survey data has been aggregated into sociological groups of interest and analysed under this framework. Model-based Clustering has provided a typology of those groups. Experimental results show the pertinence and usefulness of the proposed approach.

This framework is currently being extended, addressing robust estimation and (distributional) outlier detection, as well as other multivariate methodologies, such as MANOVA and Discriminant Analysis. An R Package is under development.

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within projects LA/P/0063/2020, DOI10.54499/LA/P/0063/2020 and UID/GES/00731/2019.

References

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

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

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

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

Fraley, C., & Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611-631.

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