Symbolic Data Analysis: Taking Variability in Data into Account

Paula Brito
Faculdade de Economia, Universidade do Porto & LIAAD - INESC TEC, Porto, Portugal

Symbolic Data, introduced by E. Diday in the late eighties of the last century, is concerned with 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; car models, rather than specific vehicles; species as a whole rather than individual specimen - 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 too important loss of information.

Symbolic Data Analysis (SDA) provides a framework allowing representing data with variability, using new variable types: the observed "values" for each case are not just single real values or categories, but finite sets of values, intervals or, more generally, distributions over a given domain. Methods for the (multivariate) analysis of such symbolic data have been developed, following different approaches and using distinct criteria, which allow taking the variability expressed in the data representation into account. SDA thereby offers the possibility of aggregating large datasets at the user's chosen degree of granularity while keeping the information on the intrinsic variability, and then analyse the resulting (symbolic) data arrays.

In this course we shall introduce and motivate the field of Symbolic Data Analysis, present into some detail the new variable types that have been introduced to represent variability, illustrating with some examples. We shall furthermore discuss some issues that arise when analysing data that does not follow the usual classical model, and present data representation models for some variable types. Then we recall some methods that have been developed to analyse symbolic data.

Parametric probabilistic models for interval-valued variables have been proposed and studied, based on the representation of each observed interval by its MidPoint and LogRange, for which Multivariate Normal and Skew-Normal distributions are assumed. The intrinsic nature of the interval-valued variables leads to different structures of the variance-covariance matrix, represented by different possible configurations. For all cases, maximum likelihood estimators of the corresponding parameters have been derived. This framework has been applied to different statistical multivariate methodologies, thereby allowing for inference approaches for symbolic data. More recently, this approach has been extended to distributional data, thereby keeping more information about the microdata distribution.

We then consider the case of aggregate numerical data described by empirical distributions, known as histogram data. Linear models for such distributional variables are proposed, which rely on the representation of histograms by the associated quantile functions. These then allow for linear regression as well as for linear discriminant analysis for histogram-valued data. An application of the proposed methodology will be presented.

The course is aimed at all potential data analysts who need or are interested in analyzing data with variability, e.g. data resulting from the aggregation of individual records into groups of interest, or data which represent abstract entities such as biological species or regions as a whole. It is assumed that the participants master classical Statistics and Multivariate Data Analysis.

References

- Duarte Silva, A.P., Brito, P., Filzmoser, P., Dias, J. (2021). MAINT.Data: Modelling and Analysing Interval Data in R. *R Journal*, 13(2). DOI: 10.32614/RJ-2021-074.
- Dias, S., Brito, P., Amaral, P. (2021). Discriminant Analysis of Distributional Data via Fractional Programming. *European Journal of Operational Research*, 294(1), 206-218. DOI: doi.org/10.1016/j.ejor.2021.01.025.
- Duarte Silva, A.P., Filzmoser, P., Brito, P. (2018). Outlier Detection in Interval Data. *Advances in Data Analysis and Classification*, 12(3), 785–822. DOI: /10.1007/s11634-017-0305-y.
- Dias, S., Brito, P. (2017). Off the beaten track: a new linear model for interval data. *European Journal of Operational Research*, 258(3), 1118–1130. DOI: 10.1016/j.ejor.2016.09.006
- Duarte Silva A.P., Brito, P. (2015). Discriminant Analysis of Interval Data: An Assessment of Parametric and Distance-Based Approaches. *Journal of Classification*, 32(3), 516-541. DOI:10.1007/s00357-015-9189-8.
- Dias, S., Brito, P. (2015). Linear Regression Model with Histogram-Valued Variables. *Statistical Analysis and Data Mining*, 8(2), 75-113. DOI: 10.1002/sam.11260.
- Brito, P. Duarte Silva A.P., Dias, J.G. (2015). Probabilistic Clustering of Interval Data. *Intelligent Data Analysis*, 19(2), 293-313. DOI: 10.3233/IDA-150718.
- 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. DOI: 10.1002/widm.1133.
- Brito, P., Duarte Silva, A.P. (2012): Modelling Interval Data with Normal and Skew- Normal Distributions. *Journal of Applied Statistics*, 39(1), 3-20. DOI: 10.1080/02664763.2011.575125.
- Billard, L., Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. *JASA*, 98 (462), 470-487. DOI: https://doi.org/10.1198/016214503000242.
- Diday, E. (1988), The symbolic approach in clustering and relating methods of data analysis: The basic choices. In: *Classification and Related Methods of Data Analysis*, Proc. First Conference of the International Federation of Classification Societies (IFCS-87), pp. 673-684. North Holland.

Books

- Brito, P., Dias, S. (Eds.) (2022). *Analysis of Distributional Data*. Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781315370545
- Diday, E., Guan, R., Saporta, G., & Wang, H. (Eds.). (2020). Advances in Data Science: Symbolic, Complex, and Network Data. John Wiley & Sons.
- Diday, E., Noirhomme-Fraiture, M. (Eds.) (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.
- Billard, L., Diday, E. (2007). Symbolic Data Analysis: *Conceptual Statistics and Data Mining*. Wiley.
- Bock, H.-H.; Diday, E. (Eds.) (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer.