Short Course: Introduction to Symbolic Data Analysis

by

Professor Paula Brito
Faculdade de Economia & LIAAD – INESC TEC; Univ. Porto, Portugal
mpbrito@fep.up.pt; www.fep.up.pt/docentes/mpbrito

University of New South Wales, 12—14 March, 2018

Symbolic Data Analysis (SDA), introduced by Edwin Diday in the late eighties of the last century, 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 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 unit 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 recent years, the term “Big Data” emerged, referring to data sets so large and complex that they become difficult to process with traditional data analysis applications and in a reasonable amount of time. SDA, offering the possibility of aggregating data at the user's chosen degree of granularity while keeping the information on the intrinsic variability, and then analyse the resulting (symbolic) data arrays, may play an important role in this context.

In this short course we shall introduce and motivate the field of Symbolic Data Analysis. We present the new variable types, 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 the real line is recorded, focusing on the parametric modelling proposed in (Brito and Duarte Silva (2012)). Methods developed to analyse symbolic data are then presented and discussed.

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. This methodology is particularly interesting for Economics and Management studies, Marketing, Social Sciences, Geography, Official Data statistics, as well as for Biology or Geology Data Analysis.

It is assumed that the participants master classical Statistics and Multivariate Data Analysis.

Paula Brito is Associate Professor at the Faculty of Economics of the University of Porto, and member of the Artificial Intelligence and Decision Support Research Group (LIAAD) of INESC TEC, Portugal. She holds a doctorate degree in Applied Mathematics from the University Paris Dauphine. Her current research focuses on the analysis of multidimensional complex data, known as symbolic data, for which she develops statistical approaches and multivariate analysis methodologies. In this context, she has been involved in two European research projects. Paula Brito was president of the International Association for Statistical Computing (IASC) in 2013-2015. She has been invited speaker at several international conferences, is regularly member of international program committees, and has been chair of COMPSTAT 2008. Web-page: www.fep.up.pt/docentes/mpbrito.

Course Timetable

Day 1 (12th March)	
Red Centre Room 3085	
10:30am—12:00pm	(1) Introduction to Symbolic Data Analysis. (1.5h)
12:00pm—1:30pm	Lunch break
1:30pm—3:00pm	(2) Parametric Modelling of Interval Data – Part I (1.5h)

Day 2 (13th March)	
Red Centre Room 3085	
10:30am—12:00pm	(3) Parametric Modelling of Interval Data – Part II (1.5h)
12:00pm—1:30pm	Lunch break
1:30pm—3:00pm	(4) Regression (1.5h)

Day 3 (14th March)	
Red Centre Committee Room	
10:30am—12:00pm	(5) Principal Component Analysis (1h)
12:00pm—1:30pm	Lunch break
1:30pm—3:00pm	(6) Classification (1.5h)

Note: All times are Sydney-time. Start times are chosen to be as family-friendly as possible for those in different time zones.

Remote Attendance:

This short course will be broadcast via zoom for those not able to attend in person. To access this visit https://unsw.zoom.us/j/238146698 and follow the prompts. Its recommended that you install the zoom software and check your connection at least 10 minutes before the meeting starts.

Note: Please mute your microphone if not speaking in order to reduce feedback and noise in the system for all viewers.

Main References

Books

- Bock, H.-H. and Diday, E. (2000). Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data. Springer, Heidelberg.
- Billard, L., Diday, E. (2007). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley.
- Diday, E. and Noirhomme-Fraiture, M. (2008). Symbolic Data Analysis and the SODAS Software. Wiley.

Papers

- Brito, P. (2014): "Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics". WIREs Data Mining and Knowledge Discovery, Volume 4, Issue 4, July/August 2014, 281– 295. DOI: 10.1002/widm.1133
- Brito,P.,DuarteSilva,A.P.(2012):"Modelling Interval Data with Normal and Skew-Normal Distributions". Journal of Applied Statistics, Volume 39, Issue 1, 3-20.
- Noirhomme-Fraiture,M.,Brito,P.(2011):"Far Beyond the Classical Data Models: Symbolic Data Analysis. " Statistical Analysis and Data Mining Volume 4, Issue 2, 157-170.
- Brito, P. (2007): "Modelling and Analysing Interval Data". In: "Advances in Data Analysis", Decker, R., Lenz, H.-J. (Eds.), Series "Studies in Classification, Data Analysis and Knowledge Organization", Springer, Berlin, Heidelberg, New-York, 197-208.
- Brito, P. (2007): "On the Analysis of Symbolic Data". In: "Selected Contributions in Classification and Data Analysis", Brito, P., Bertrand, P., Cucumel, G., De Carvalho, F. (Eds.), Series "Studies in Classification, Data Analysis and Knowledge Organization", Springer,Heidelberg, 13-22.
- Duarte Silva, A. P. , Brito, P. (2006). "Linear Discriminant Analysis for Interval Data". Computational Statistics, 21, 2, 289-308.
- 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), pp. 470–487.