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

Paula Brito (*)
Fac. Economia / LIAAD - INESC TEC, Universidade do Porto, PORTUGAL
mpbrito@fep.up.pt

A. Pedro Duarte Silva
Fac. Economia e Gestão / CEGE, Universidade Católica Portuguesa, Porto, PORTUGAL,
psilva@porto.ucp.pt

José G. Dias
ISCTE - Instituto Universitário de Lisboa, BRU-IUL, Lisboa, PORTUGAL, jose.dias@iscte.pt

In this paper we propose a model-based approach to the clustering of interval data building on recently proposed parametric models for interval data. 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 parameterizations that allow representing 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 real data sets the pertinence of the methodology proposed is made clear. In particular, its flexibility to identify the most suitable configuration of restricted covariance structures (e.g., class invariant, block matrices…) is put in evidence. The method effectively selected heterocedastic models when they are the most suitable, even in situations with limited information. 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 analyzing interval data.

Key-Words: Finite mixture models, Interval-valued variable, Intrinsic variability, Symbolic data