Robust Sparse Models and Outlier Detection for Multivariate Distributional Data

Pedro Duarte Silva[1], Peter Filzmoser[2], and Paula Brito[3]

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

psilva@ucp.pt

[2] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria

peter.filzmoser@tuwien.ac.at

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

mpbrito@fep.up.pt

The classical data representation model, where for each statistical unit a single value is recorded for each variable, is too restrictive when the data to be analysed are not real numbers or single categories but comprise variability. In this work, we focus on numerical distributional data, i.e., data where units are described by histogram or interval-valued variables, representing the intrinsic variability of the corresponding observations. In our model, each distribution is represented by a location measure and interquantile ranges, for a given set of quantiles; typical cases consist in using the median, or else the midpoint, as central statistics, and quartiles, or other equally-spaced quantiles. The proposed model consists in assuming that the joint distribution of the central statistic and the logarithms of the ranges is Gaussian. Alternative sparse structures of the variance-covariance matrix are considered, which allow modelling the possible relations between the different indicators. A multivariate outlier detection method is then proposed that is based on a sparse robust estimator of the inverse of the variance-covariance matrix. The computations rely on an efficient adaptation of the graphical lasso algorithm. The proposed methodology is evaluated in a controlled simulation experiment, and illustrated with real distributional-valued data.