Advances in Data Analysis and Classification (2022) 16:487–490
https://doi.org/10.1007/s11634-022-00511-9

EDITORIAL

Editorial for ADAC issue 3 of volume 16 (2022)

Maurizio Vichi[1] · **Andrea Ceroli**[2] · **Hans A. Kestler**[3] · **Akinori Okada**[4] · **Claus Weihs**[5]

Published online: 16 August 2022

This issue 3 of volume 16 (2022) of the journal *Advances in Data Analysis and Classification (ADAC)* contains 10 articles that deal with association measures, finite mixture modeling, prediction of brand stories, maximal frequent rectangles, classification of bags, robust clustering, probabilistic model for random hypergraphs, outlier detection in big functional data, estimation of high-density regions, model-based recursive partitioning.

The first paper, written by *M. Rosário Oliveira Margarida Azeitona*, *António Pacheco, Rui Valadas*, is entitled "Association measures for interval variables". Authors propose a model linking the micro-data with the macro data of interval-valued symbolic variables. In this model the micro-data is defined as a random vector which is a function of the centers and ranges associated with the macro-data and random weights characterizing the structure of the micro-data given the associated macro-data. The model defines two scenarios where the various definitions of symbolic covariance matrices already proposed in the literature arise as particular cases. These scenarios correspond to two extreme situations: in the first one the weights are independent random variables and in the second one they are equal variables (almost surely). In both scenarios, the weights have zero mean and uncorrelated latent variables. The conditions on the random weights imply micro-data assumptions that may be too stringent, raising applicability concerns. Authors admit that more research is required in this area. These cases also highlight that, in the context of current definitions, a null symbolic covariance cannot be interpreted as absence of association. This confirms the need for further research on how to measure associations between interval-valued variables.

✉ Maurizio Vichi
maurizio.vichi@uniroma1.it

[1] Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Roma, Italy

[2] Department of Economics and Management, University of Parma, Parma, Italy

[3] Institute of Medical Systems Biology, Ulm University, Ulm, Germany

[4] Rikkyo University, Tokyo, Japan

[5] Faculty for Statistics, TU Dortmund University, Dortmund, Germany

The second article is written by *Francisco H. C. de Alencar, Christian E. Galarza, Larissa A. Matos, Victor H. Lachos* and entitled "Finite mixture modelling of censored and missing data using the multivariate skew-normal distribution". Authors suggest a novel approach to analyse multiply censored and missing data. They use of finite mixtures of multivariate skew-normal distributions generalizing several previously proposed solutions for censored data. A simple and efficient EM-type algorithm was developed, which has closed-form expressions at the E-step and relies to the mean vector and covariance matrix of the multivariate truncated skew-normal distribution. The proposed EM algorithm was implemented as part of the R package CensMFM and is available at the CRAN repository. The experimental results and the analysis of a real dataset provide support for the usefulness and effectiveness of our proposal. Authors underline that the proposed method can be extended to other types of mixture distributions, such as the multivariate scale mixtures of skew-normal distributions (or generalized hyperbolic mixtures). They also emphasize that the methodology can be easily applied to other areas in which the data being analysed have censored and/or missing observations, for instance, factor analysis models and linear mixed models.

In the next paper entitled "Prediction of brand stories spreading on social networks*", Thi Bich Ngoc Hoang*, and *Josiane Mothe* propose a method that helps business managers to understand and predict how popular a given brand story is on social networks. Specifically, they address the problem of predicting the diffusion of a given brand story on Twitter. The problem is casted into a binary classification and multi-class classification (predict the level of retweets). Authors define new features including: user-profile-based, temporal-based, and content-based features for a total of 32 variables used in the model. Considering two types of collections: consumer-generated stories and company's official stories, authors show that the model improves the F-measure by about 4% compared to the state of art for both types of prediction.

In the forth paper entitled "Mining Maximal Frequent Rectangles", *Irani Hazarika and Anjana Kakoti Mahanta,* extended the problem of maximal frequent interval mining to the problem of maximal frequent rectangle mining. First the notion of maximal frequent rectangles was proposed. Then, authors prove some properties related to rectangles as well as maximal frequent rectangles that have been used by the proposed algorithm that has been implemented and tested on a number of synthetic interval data sets of various sizes for different minimum support values. Also a discussion on the extension of maximal frequent rectangles into 3-dimensional space as maximal frequent cube has been given.

In the fifth paper, written by *Matteo Spallanzani*, *Gueorgui Mihaylov*, *Marco Prato* and *Roberto Fontana* on "A *fingerprint* of a heterogeneous data set", the authors propose a technique for the classification of bags of mixed type measurements. The methodology is motivated by a complex real-world industrial problem, such as the classification of industrial plants starting from the measurements collected from their production lines. In this setting, the fingerprint method is formally developed to compare the mixture components of a given test bag with the corresponding mixture components associated with the different classes, identifying the most similar generating distribution. The suggested technique is then compared to other classification algorithms on several synthetic data sets and on the original industrial example, showing remarkable improvements in performance.

The sixth article is written by *Wan-Lun Wang* and *Tsung-I Lin* on "Robust clustering via mixtures of t factor analyzers with incomplete data". In this work the authors present an extension of mixture models for factor analyzers that can accommodate missing information in the data. Being based on the t distribution, the proposed model can deal simultaneously with missing values and mild outliers. In the first part of the paper, the authors derive the score vector and Hessian matrix of the proposed model with incomplete data to approximate the information matrix. They also establish some asymptotic properties under regularity conditions. Then, three expectation-maximization-based algorithms are developed for maximum likelihood estimation of the model with possibly missing values at random. Practical issues related to the recovery of missing values and clustering of partially observed samples are also investigated. The usefulness of the methodology is finally exemplified through the analysis of simulated and real data sets.

In the next paper entitled "Model-based clustering for random hypergraphs", *Tin Lok James Ng* and *Thomas Brendan Murphy* present a novel probabilistic model for random hypergraphs that can represent unary, binary and higher order interactions among objects. Examples of the need to model such complex interactions occur in many modern applications, such as the analysis of co-authorship on academic papers and the study of co-appearance in movie scenes. The authors first show that the proposed model is an extension of the latent class analysis model that introduces two clustering structures for hyperedges and captures variation in the size of hyperedges. They also study the distribution of the size of a random hyperedge under the proposed framework. In the second part of the paper, an expectation maximization algorithm with minorization maximization steps is developed to perform parameter estimation. Model selection is performed using Bayesian Information Criterion. The model is then applied to simulated data and two real-world data sets on co-appearance relationships in movie scenes and news articles, where interesting results are obtained.

The eighth article by *Oluwasegun Taiwo Ojo , Antonio Fernández Anta, Rosa E. Lillo, and Carlo Sguera* titled "Detecting and classifying outliers in big functional data", extends an existing method of outlier detection in univariate functional data. The two proposed methods Semifast-MUOD and Fast-MUOD are capable of identifying and classifying magnitude, amplitude, and shape outliers without the need for visualisation. Both reduce the time complexity of O($n^2\cdot$ d) of the original MUOD method. Semifast-MUOD achieves this by taking a sub-sample of the functional data, while Fast-MUOD reduces the complexity to O($n\cdot d$) by using the point-wise median in the computation of the three types of indices. The authors show both on simulated and real data the performance benefits of the proposed methods. Further comparisons to existing univariate functional outlier detection tools show comparable or superior results in correctly identifying potential outliers of different types. Furthermore, the authors have implemented their methods in R and made their code available on github.

In the ninth paper by *Paula Saavedra-Nieves and Rosa María Crujeiras* with the title "Nonparametric estimation of directional highest density regions", the authors extend the definition of high-density regions (HDRs) to directional data and propose a plug-in estimator based on a new bootstrap bandwidth selector that is focused on HDRs reconstruction. They show the method's utility through an extensive simulation study and two studies on real-world data which also shows the method's applicability for circular (behavioural plasticity of sand hoppers) and spherical data (distribution of earthquakes). The code is readily available in their package HDiR.

Finally, in the tenth article titled "Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning", *Cynthia Huber, Norbert Benda, and Tim Friede* propose a new model-based recursive partitioning (MOB) method for subgroup identification in individual participant data meta-analysis. Their method metaMOB utilises the generalized mixed effects model tree algorithm. The goal is to identify treatment by subgroup interactions while accounting for between-trial heterogeneity in these meta-analysis settings. They implement two versions that differ in their assumptions regarding baseline effects. In one of the models, it is assumed that the between-trial heterogeneity in the baseline effect is fixed; in the other, it is assumed to be random. In their extensive simulation study, they show the utility of their method and argue for their stratified intercept approach (fixed baseline effect) and the inclusion of a large number of trials for the application of metaMOB.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.