Advances in Data Analysis and Classification (2020) 14:231–234
https://doi.org/10.1007/s11634-020-00407-6

EDITORIAL

Special issue on "Innovations on model based clustering and classification"

Christophe Biernacki[1] · Luis Angel García-Escudero[2] · Salvatore Ingrassia[3]

Published online: 11 June 2020

This is the fourth Special Issue of ADAC dedicated to recent developments in Model-Based Clustering and Classification, an area which provides increasingly active research in both theoretical and applied domains. The Call for Papers for this special issue resulted in 46 manuscript submissions, among which 10 have been accepted for publication after a blinded peer-reviewing process.

This Special Issue contains papers dealing with quite different topics. The first three papers focus on the area of mixtures of regressions from different perspectives. The three following papers deal with mixture models for robust model-based clustering, data modeling with multiple partial right censoring points, and modeling under measurement inconsistency, respectively. The next two papers belong to the area of mixtures of skewed distributions. Two final papers concern particle Monte Carlo methods and stochastic block models. Below, we provide a short overview on the papers published in this special issue.

The paper "Seemingly unrelated clusterwise linear regression" by *Giuliano Galimberti* and *Gabriele Soffritti* presents a flexible class of finite mixtures of multivariate Gaussian linear regression models in which different vectors of predictors can be used for the dependent variables. The model is fit to data according to the maximum likelihood approach and some parsimonious model are introduced too. One of the main contributions of the paper consists in providing conditions for model identifiability and in the analysis of the consistency of the maximum likelihood estimator under the proposed class of models. In this framework, theoretical results and simulation studies of the behavior of the ML estimator under different scenarios are provided, considering both different sample sizes and overlap levels among regression models. The model is also illustrated by an application to real data.

The next paper by *Sijia Xiang* and *Weixin Yao* entitled "Semiparametric mixtures of regressions with single-index for Model Based Clustering" proposes two finite

✉ Christophe Biernacki
christophe.biernacki@inria.fr

1 Lille, France

2 Valladolid, Spain

3 Catania, Italy

semiparametric mixture-of-regression models that can incorporate high-dimensional predictors into the nonparametric components. Modeling approaches for the high-dimensional predictors case is an area of increasing relevance in statistics. The two proposals include various previously proposed semiparametric/nonparametric mixture regression models as particular cases. A computational procedure combining backfitting and a modified EM algorithm is introduced. Identifiability results for the two new models are established and their asymptotic properties are investigated, demonstrating that optimal convergence rates can be achieved for both parametric and nonparametric components.

In the area of mixtures of regressions, another problem is proposed in the paper "Gaussian parsimonious clustering models with covariates and a noise component" by *Keefe Murphy* and *Thomas Brendan Murphy*. The authors focus here on Mixture of Experts (MoE) models in which the parameters of the mixture are modeled as functions of fixed, potentially mixed-type covariates. The paper presents two main contributions. First, a unifying framework combining Gaussian MoE with the flexibility of the covariance constraints in the Gaussian parsimonious model family is presented; the resulting approach is here called MoEClust and is also associated with the R package MoeClust. The second main contribution belongs to the area of outlier detection and concerns the addition of a noise component for capturing outlying observations. The clustering performance of the proposed model is exemplified through well-know real datasets.

An important area in model-based clustering concerns robust approaches. In this framework, *Andrea Cappozzo*, *Francesca Greselin* and *Thomas Brendan Murphy* present a paper entitled "A robust approach to model-based classification based on trimming and constraints". Their proposal introduces robust estimation of a Gaussian mixture model with parsimonious structure, to account for both attribute and label noise. Techniques originating from the domain of robust statistics are exploited here through a semi-supervised approach to obtain a model-based classifier in which parameters are robustly estimated and outlying observations identified. Computational issues are investigated in depth. Finally, the effectiveness of this methodology is illustrated by means of a large numerical study based on both simulated data and an application to real data concerning midinfrared spectroscopy of Irish honey.

The fifth paper entitled "Mixture modeling of data with multiple partial right-censoring levels" and authored by *Semhar Michael, Tatjana Miljkovic* and *Volodymyr Melnykov* looks at finite mixture models for data with multiple partial right censoring points. In this approach it is assumed that a certain portion of the observations are censored while others are not. The approach is particularly interesting for insurance loss data with multiple partial censoring levels that often appear due to the heterogeneous nature of insurance claims. The assessment of the variability of the parameter estimates is analysed and details are provided on how to calculate two common risk measures in actuarial applications, such as the Value-at-Risk (VaR) and Tail-Value-at-Risk (TVaR).

It is well known that variations due to differences in data collection (different ways of recording/representing data or different measurement scales) can seriously affect the performance of typical model-based clustering methods. The paper entitled "Gaussian mixture modeling and model-based clustering under measurement inconsistency" by *Shuchismita Sarkar*, *Volodymyr Melnykov* and *Rong Zheng* addresses this important problem. A generalization of the Gaussian mixture models that face with this

measurement inconsistency is proposed. A k-means-like algorithm is introduced for clustering massive datasets. Various simulation studies and an application to a real data example are provided to illustrate the interest of the proposed methodology.

The paper entitled "Mixtures of Skewed Matrix Variate Bilinear Factor Analyzers" is written by *Michael P.B. Gallaugher* and *Paul D. McNicholas* and addresses an area of recent and increasing research in mixture modeling, namely model-based clustering of matrix variate or three-way data such as multivariate longitudinal data or images, particularly in the related high-dimensional data situation. However, the methods available in literature assume the normality of the matrix variate, which is not adequate if clusters present skewness or excess kurtosis. In this paper, mixtures of bilinear factor analyzers models using skewed matrix variate distributions are proposed. In this framework, four mixture models are presented, based on matrix variate skew-t, generalized hyperbolic, variance gamma and normal inverse Gaussian distributions, respectively. Related estimation procedures are given and numerical experiments illustrate the practical interest of the proposal.

A quite different topic is considered in the paper "Data projections by skewness maximization under scale mixtures of skew-normal vectors" by *Jorge M. Arevalillo* and *Hilario Navarro*. Here the focus is on multivariate scale mixtures of skew-normal distributions, which are flexible models that account for asymmetry departures from normality. The paper is placed in the framework of projection pursuit and provides theoretical foundations motivating the skewness based projection pursuit problem for vectors that follow a multivariate scale mixture of skew-normal distribution. The methodological results are finally illustrated by means of some numerical studies based on both a large simulation study and two applications to real datasets concerning genomic data and biomedical data.

Another topic in this Special Issue concerns the integrative cluster analysis of multiple datasets using particle Monte Carlo methods. In the paper entitled "particleMDI-Particle Monte Carlo methods for the cluster analysis of multiple datasets with applications to cancer subtype identification", the authors *Nathan Cunningham, Jim E. Griffin* and *David L. Wild* present a nonparametric Bayesian approach for cluster analysis where observational data units come from multiple sources. A particle Gibbs sampler approach for inference is presented in which cluster assignments are updated by using a conditional particle filter within a Gibbs sampler. The proposed approach is applied to real biological data from the Cancer Genome Atlas. The procedure can be easily generalized to other types of data.

The tenth paper entitled "A stochastic block model for interaction lengths" is written by *Riccardo Rastelli* and *Michael Fop* and introduces a simple, elegant and original stochastic blockmodel (SBM) for continuous-time interaction lengths, differently from some previous approaches that consider discrete-time for the same dynamic network context. The aim is to detect clusters of units (nodes) sharing similar interaction lengths. The underlying main idea is that both interaction times and non-interaction times characterize the groups of the SBM. The model has the advantage that there is no need for time discretization and also that the truncated (non-) interaction lengths are taken into account as censored data. The time of the interaction (and non-interaction) is modeled via an exponential distribution. A classical variational algorithm determines the unit groups and the model parameters, whereas model selection is performed

through the Integrated Classification Likelihood (ICL) criterion. Experiments with synthetic data assess the clustering performance for the nodes of the network. An interesting real dataset, consisting of interactions between students, is finally analyzed to illustrate the practical interest of the approach.

The Editors gratefully acknowledge the assistance of the following experts and colleagues in the process of reviewing the manuscripts that were submitted for this special issue:

Christophe Ambroise (France), Celso Romulo Barbosa Cabral (Brazil), Tatiana Benaglia (Brazil), Etienne Birmelé (France), Laurent Bordes (France), Charles Bouveyron (France), Gilles Celeux (France), Alain Célisse (France), William Chad Young (USA), Pietro Coretto (Italy), Marco Corneli (France), Sophie Dabo (France), Alessio Farcomeni (Italy), Michael Fop (Italy), Giuliano Galimberti (Italy), Claire Gormley (Ireland), Francesca Greselin (Italy), Bettina Grun (Austria), Karel Hron (Czech Republic), Julien Jacques (France), Pierre Latouche (France), Seokho Lee (Republic of Korea), Tsung-I Lin (Taiwan), Nicola Loperfido (Italy), Matthieu Marbac-Lourdelle (France), Catherine Matias (France), Paul McNicholas (Canada), Volodymyr Melnykov (USA), Cristian Preda (France), Antonio Punzo (Italy), Marco Riani (Italy), Gunter Ritter (Germany), Magdalena Strauss (UK), Hemant Tyagi (France), Vincent Vandewalle (France), Cinzia Viroli (Italy), Weixin Yao (USA), Riquan Zhang (China).