Advances in Data Analysis and Classification (2019) 13:1–5
https://doi.org/10.1007/s11634-019-00355-w

EDITORIAL

Special issue on "Advances on model-based clustering and classification"

Preface by the Guest Editors

Sylvia Frühwirth-Schnatter[1] · Salvatore Ingrassia[2] · Agustín Mayo-Iscar[3]

Published online: 1 March 2019

This Special Issue of ADAC is devoted to recent developments in Model-Based Clustering and Classification which is an increasingly active area in both theoretical and applied research. This area has attracted the interest of a growing number of researchers due to the high potential of such approaches in applications, and new different topics have been investigated by several authors. This special issue contains 13 papers, that have been accepted for publication after a blinded peer-reviewed process, dealing with quite different topics like mixture models for both continuous and discrete data, co-clustering, robust approaches to clustering data, clustering time series, agreement measure for class partitions, and various applications in the framework of model-based clustering. Below, we provide a short overview of the papers published in this special issue.

The first paper, entitled "Unifying Data Units and Models in (Co-)Clustering" by Christophe Biernacki and Alexandre Lourme, draws attention on links between data and statistical models. The authors point out that the interpretation of ("classical") models is usually unit dependent and that models should thus be revisited as a couple (units, models). This can provide an opportunity for cheap, wide and meaningful enlarging of "classical" model families. The paper focuses especially on clustering and co-clustering (a simultaneous clustering of rows and columns) in the case of

✉ Salvatore Ingrassia
s.ingrassia@unict.it

Sylvia Frühwirth-Schnatter
sylvia.fruehwirth-schnatter@wu.ac.at

Agustín Mayo-Iscar
agustin@med.uva.es

1 Department of Finance, Accounting, and Statistics, Vienna University of Business and Economics, Vienna, Austria

2 Department of Economics and Business, University of Catania, Catania, Italy

3 IMUVA & Department of Statistics and Operational Research, University of Valladolid, Valladolid, Spain

possibly mixed data, but the approach could be extended to other models. Questions about the definition of new units are also addressed, in particular the possibility for the user to propose so-called "meaningful" units and also the possibility for the statistician to propose so-called "technical" units.

Subsequently, the paper entitled "From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering" by Sylvia Frühwirth-Schnatter and Gertraud Malsiner-Walli approaches sparse finite mixtures, introduced quite recently for Gaussian mixtures within the framework of Bayesian model-based clustering as a bridge between standard finite mixture and Dirichlet process mixture models. The concept of sparse finite mixtures is extended here to cluster a broad range of non-Gaussian data, in particular discrete data and continuous multivariate data arising from non-Gaussian clusters. Moreover, properties of sparse finite mixtures are investigated and compared to Dirichlet process mixtures with respect to their ability to identify the number of clusters in applications based on Poisson mixtures, mixtures of generalized linear models, and latent class models.

The paper entitled "Clustering Via Finite Nonparametric ICA Mixture Models" by Xiaotian Zhu and David R. Hunter proposes a novel extension of nonparametric multivariate finite mixture models by dropping the standard conditional independence assumption and incorporating the independent component analysis (ICA) structure instead. This innovation extends nonparametric mixture model estimation methods to situations in which conditional independence, a necessary assumption for the unique identifiability of the parameters in such models, is clearly violated. The authors formulate an objective function in terms of penalized smoothed Kullback–Leibler distance and introduce the nonlinear smoothed majorization-minimization independent component analysis (NSMM-ICA) algorithm for optimizing this function and estimating the model parameters. This algorithm is useful for fully unsupervised clustering problems in a multivariate setting and for image processing and has been implemented in the R package icamix.

In statistical analysis, particularly in econometrics, finite mixtures of regression models based on the normality assumption are popular for analyzing censored data. In the paper entitled "Finite Mixture of Regression Models for Censored Data Based on Scale Mixtures of Normal Distributions" by Camila Borelli Zeller, Celso Rômulo Barbosa Cabral, Víctor Hugo Lachos and Luis Benites an extension of this model is proposed by considering scale mixtures of normal distributions. This approach allows to model data with great flexibility, accommodating multimodality and heavy tails at the same time. The main virtue of considering finite mixture of regression models for censored data under the class of scale mixtures of normal distributions is that this model class has a convenient hierarchical representation that allows easy implementation of statistical inference. To perform maximum likelihood (ML) inference for the parameters of the proposed model, a simple EM-type algorithm is implemented and examined for several simulation studies. The method is applied to data on working hours for married women.

Methods which deal with clustering in matrices of data are often based on mathematical techniques such as distance-based algorithms or matrix decomposition and eigenvalues. However, in general, it is not possible to perform statistical inference or to select the appropriateness of a model via information criteria with these techniques

because there is no underlying probability model. The paper entitled "Finite mixture biclustering of discrete type multivariate data" by Daniel Fernández, Richard Arnold, Shirley Pledger, Ivy Liu, Roy Costilla reviews some recent model-based methodologies for matrices of binary, count, and ordinal data, which are modelled under a unified statistical framework using finite mixtures to group the rows and/or columns. The model parameters depend on a linear predictor of parameters and covariates through link functions. This likelihood-based one-mode and two-mode fuzzy clustering provides maximum likelihood estimation of parameters and the options of using likelihood information criteria for model comparison. Additionally, a Bayesian approach is implemented which estimates the parameters and the number of clusters simultaneously from their joint posterior distribution. Visualization tools focusing on ordinal data show the fuzziness of the clustering structures and extend various standard plots used in multivariate analysis.

In a quite theoretical framework, the paper entitled "Finite mixtures, projection pursuit and tensor rank: a triangulation" by Nicola Loperfido is placed in the framework of dimension reduction according to the projection pursuit approach. The paper theoretically motivates skewness-based projection pursuit for mixtures of multivariate distributions and relates it to the linear discriminant function when more than two skewed components are present. The approach is algebraic in nature and deals with the symmetric tensor rank of the third multivariate cumulant. The paper also derives closed-form expressions for the symmetric tensor rank of the third cumulants of several multivariate mixture models, including mixtures of skew-normal distributions and mixtures of two symmetric components with proportional covariance matrices.

The paper entitled "Clustering Space–Time Series: FSTAR as a flexible STAR Approach" by Edoardo Otranto and Massimo Mucciardi investigates the problem of modelling the dynamics of variables recorded at several locations at the same time. The STAR approach is widely used to represent such data and its advantage compared to space–time VAR structures is parsimony with respect to the number of parameter, as a single coefficient is introduced for each time and spatial lag. As this hypothesis can be very restrictive, the paper adds a certain degree of flexibility to the STAR model, by providing the possibility for coefficients to vary in groups of locations. This new class of models is compared to the classical STAR and the space–time VAR by simulation studies and an application to birth rate data from districts in the province of Caserta in Italy.

The paper entitled "Robust clustering for functional data based on trimming and constraints" by Diego Rivera-García, Luis Angel García-Escudero, Agustín Mayo-Iscar and Joaquín Ortega relies on robust approaches in clustering functional data, in particular when the sample of curves to be clustered contains contaminating data. Here an approach that relies on an approximation to the "density function" for functional data is proposed. The robustness follows from the joint application of data-driven trimming, for reducing the effect of contaminated observations, and constraints on the variances, for avoiding spurious clusters in the solution. A feasible algorithm to perform clustering and outlier detection simultaneously by maximizing a trimmed "pseudo" likelihood is then presented; simulated and real data example are presented to illustrate the performance of the proposed methodology.

Francesca Torti, Domenico Perrotta, Marco Riani and Andrea Cerioli investigate robust proposals for linear regression clustering in their contribution "Assessing trimming methodologies for clustering linear regression data". Robustness of the analyzed proposals is based on the joint application of trimming and constraints. An issue regarding the application of these robust methodologies is related to the choice of input parameters. In this paper, the authors analyze the role of these input parameters and propose a new approach for automatically estimating the trimming level. Additionally, they developed a tool for generating artificial data from mixtures of linear models that can control both the level of overlapping between components and the degree of overlapping with outliers. The authors provide implementations of all the proposals suggested in the paper in the FSDA toolbox of MATLAB.

The paper entitled "Variable selection in model-based clustering and discriminant analysis with a regularization approach" by Gilles Celeux, Cathy Maugis-Rabusseau and Mohammed Sedki, considers the problem of variable selection in the contexts of supervised and unsupervised model-based approaches to grouping data. Variable selection is an important issue in many situations. In this paper, the authors try to obtain a classification of the explanatory variables in relation to their relevance for identifying groups and their correlation with the set of relevant variables. Approaches for obtaining such a classification based on stepwise procedures are available, but tend to fail when applied to datasets with a high number of variables. The proposal presented here uses a lasso regularization strategy for avoiding a stepwise search procedure. It shows a high level of performance in the simulation studies. This methodology is available from the authors in the R package SelvarMix.

The paper entitled "Random effects clustering in multilevel modeling: choosing a proper partition" by Claudio Conversano, Massimo Cannas, Francesco Mola and Emiliano Sironi presents a novel criterion for estimating a latent partition of the observed group using samples of random effects from a hierarchical model. The criterion is based on a loss function combining the Gini income inequality ratio and the predictability index of Goodman and Kruskal in order to achieve maximum heterogeneity of random effects across groups and maximum homogeneity of predicted probabilities inside estimated clusters. The index is compared to alternative approaches in a simulation study and is applied in a case study concerning the role of hospital level variables in deciding for a cesarean section.

A quite different topic is presented in the paper entitled "ARI: a soft agreement measure for class partitions incorporating assignment probabilities" by Abby Flynt, Nema Dean and Rebecca Nugent. Here an extension of the ARI called soft adjusted Rand index (sARI) is proposed, having a similar intuition and interpretation. This new index also incorporates information from one or two soft partitions, coming from techniques like model-based clustering that include information about the certainty of allocation of objects through class membership probabilities. This index can be used in conjunction with the ARI, comparing the similarities of hard-to-soft or soft-to-soft partitions to the similarities of the mapped hard partitions. Applications show that the sARI more accurately reflects the cluster boundary overlap commonly seen in real data.

Finally, the paper "Studying crime trends in the USA over the years 2000–2012" by Volodymyr Melynkov and Xuwen Zhu is devoted to modelling matrix-valued data

in a flexible way, given by a mixture of skew-transformations of normal distributions. Methodology for addressing matrix type data is relatively recent, and is extended in this paper to handle skewed data. With this aim, the authors use the Manly transformation, providing a very competitive alternative when compared to other proposals commonly applied in other context for managing skewness, as skew-normal and skew-t distributions. It is easy to understand, as it is applied in a univariate way, and at the same time powerful for getting the desired flexibility. Moreover, the authors provide parsimonious parameterizations for reducing the high number of parameters included in the model. Their approach is motivated through the analysis of crime data in the United States.