CLADAG2017

Book of Short Papers

Editors: Francesca Greselin,
Francesco Mola and Mariangela Zenga

via Sottoriva, 9 - 46100 Mantova (MN)
P. IVA 02346110204
E-book (PDF version) published in September 2017
ISBN 978-88-99459-71-0

This book is the collection of the Abstract / Short Papers submitted by the authors of the International Conference of The CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS), held in Milan (Italy), University of Milano-Bicocca, September 13-15, 2017.

Euro 9,00

Contributed sessions

Classification of Multiway and Functional Data

A generalized Mahalanobis distance for the classification of functional data

Andrea Ghiglietti, Francesca Ieva, Anna Maria Paganoni

Classification methods for multivariate functional data with applications to biomedical signals

Andrea Martino, Andrea Ghiglietti, Anna M. Paganoni

A new Biclustering method for functional data: theory and applications

Jacopo Di Iorio, Simone Vantini

A leap into functional Hilbert spaces with Harold Hotelling

Alessia Pini, Aymeric Stamm, *Simone Vantini*

Sampling Designs and Stochastic models

Statistical matching under informative probability sampling

Daniela Marella, Danny Pfeffermann

Goodness-of-fit test for discrete distributions under complex sampling design

Pier Luigi Conti

Structural learning for complex survey data

Daniela Marella*, Paola Vicard*

The size distribution of Italian firms: an empirical analysis

Anna Maria Fiori, Anna Motta

Robust statistical methods

New proposal for clustering based on trimming and restrictions

Luis Angel Garcìa Escudero, Francesca Greselin, *Agustin Mayo Iscar*

Wine authenticity assessed via trimming

Andrea Cappozzo, Francesca Greselin

Robust and sparse clustering for high-dimensional data

Sarka Brodinova, Peter Filzmoser, Thomas Ortner, Maia Zaharieva, Christian Breiteneder

M-quantile regression for multivariate longitudinal data

Marco Alfo', *Maria Francesca Marino,* Maria Giovanna Ranalli, Nicola Salvati, Nikos Tzavidis

New proposals in Clustering methods

Reduced K-means Principal Component Multinomial Regression for studying the relationships between spectrometry and soil texture

Pietro Amenta, *Antonio Lucadamo*, Antonio Pasquale Leone

Comparing clusterings by copula information based distance

Marta Nai Ruscone

Fuzzy methods for the analysis of psychometric data

Isabella Morlini

Inverse clustering: the paradigm, its meaning, and illustrative examples

Jan W. Owsinski, Jaroslaw Stanczak, Karol Opara, Slawomir Zadrozny

NEW PROPOSALS FOR CLUSTERING BASED ON TRIMMING AND RESTRICTIONS

Luis Angel García Escudero [1], Francesca Greselin[2] and Agustín Mayo Iscar[1]

[1] Departmento de Estadística e I.O., Universidad de Valladolid, (e-mail: lagarcia@eio.uva.es, agustin@med.uva.es)

[2] Department of Statistics and Quantitative Methods,University of Milano-Bicocca, (e-mail: francesca.greselin@unimib.it)

ABSTRACT: TCLUST is a model-based clustering methodology, which employs trimming and restrictions for getting robust estimators. It is available in the *tclust* package at the CRAN website and in the FSDA Matlab library. Extensions of TCLUST modelling include clustering around linear subspaces, factor analyzers approaches and fuzzy proposals. Further research has been focused in allowing more flexible models for the components, based on the skew normal distribution. An important issue that may appear within TCLUST is the dependence of the obtained solutions from the input parameters. Therefore, a variety of tools have been developed to assist to the users in choosing these parameters. Theoretical and robustness properties for the TCLUST estimators have been proven, and many empirical evidences show the efficacy of the proposed methodology, in a wide variety of situations.

KEYWORDS: model based clustering, robustness, trimming, constraints

1 TCLUST methodology

TCLUST methodology is included in robust model-based clustering. The robustness of this approach is based on the joint application of trimming and constraints. Here, we firstly introduce the basic release of the methodology, tailored for samples drawn from a finite mixture of a known number of normal populations. Our approach allows for adopting different weights in the clusters/mixtures components and different patterns for their scatters. It yields a restricted maximum likelihood estimator, corresponding to the classification likelihood of the mixture model, based on a $(1-\alpha)$ proportion of the observations. The selected observations are the ones with the highest contribution to the likelihood. The associated trimming refers to the non-selected observations, to eliminate the influence of the observations lying far from the bulk of the data, for all the components. The application of the scatter constraints, along the estimation procedure, aims at avoiding all singularities in the likelihood, and to reduce the occurrence of spurious local maximizers. The constraints included in the basic version of TCLUST are expressed in terms of the eigenvalue restrictions. They control the relative size of all eigenvalues belonging to the components scatter matrices, by using a constant c, which is a parameter related with the level of strength of the restrictions. Therefore, the input parameters included in this basic version of the model are the number of components k, the level of trimming α, corresponding to the percentage of eliminated observations, and the constant c, related with the level of the restrictions. It has been proven that the optimization problem solved by the TCLUST estimator is well-posed, therefore the estimator exists and it is consistent to the solution of the corresponding theoretical optimization. The seminal paper of TCLUST methodology is García Escudero *et al.* (2008), where the main statistical properties of the methodology have been presented. García Escudero *et al.* (2014) extended this methodology to mixture model estimation. Robustness properties of TCLUST methodology related to the influence function and the breakdown point of the estimator can be found in Ruwet *et al.* (2012) and Ruwet *et al.* (2013). The TCLUST algorithm is a classification EM algorithm, which has been adapted for finding the maximum in the constrained parameter space. To get it, at each M step an explicit function in $2kp+1$ values has been provided (p is the dimension of the multivariate observations). The derived algorithm is a substantial improvement over the initial release, based on Dijkstra algorithm. Details about it can be found at Fritz *et al.* (2012). TCLUST is available in the CRAN website (in *tclust* package), and in the FSDA library of Matlab. The packages has been presented in Fritz *et al.* (2012) and Riani *et al.* (2012).

2 TCLUST extensions

An important branch of TCLUST methodology has been developed for estimating clusters around linear subspaces. The joint application of trimming and constraints, specifically adapted to the new framework, allows to get robust estimators. The first available proposals (García-Escudero *et al.* , 2009, 2010, 2017) include clustering/mixture model approaches based on orthogonal/regression errors. As expected, this methodology works very well in presence of noisy observations, and with concentrated contamination as well. When the contamination is close to the model and located outside the support of the explanatory variables, the estimation can be modified in a substantial way by the ouliers, even if it is still protected against breaking down. To avoid this undesirable effect, initial proposals included a second trimming step, applied on the surviving observations after the usual trimming. The second trimming step has been added for eliminating the more distant observations, in terms of the explanatory variables. Indeed, they are classically considered as the most influential ones for linear estimation. García-Escudero *et al.* (2017) proposed the robustification of the mixtures of regression by employing the cluster-weighted model. With this choice, a modelization for the explanatory variables is provided, apart from the usual Gaussian assumption on the regression errors. Therefore the cluster-weighted model allows to implement simultaneously both versions of trimming. This approach avoids not only the harmful influence of observations located far from the linear model, it also eliminates the potentially more dangerous effects of observations having outlying values of the explanatory variables.

A second important extension of TCLUST is related to Flury's paper (Flury, 1984), where the author motivated the utility of common principal components by showing several applications in different areas, ranging from Biometry to Industry. It corresponds to a constrained estimation of the eigenvectors of the covariance matrices, to get the same set of eigenvectors for all the components. This model is included in the collection of 14 parsimonious mixture models proposed by Celeux & Govaert (1995). Browne & McNicholas (2014) improved the classical available algorithm from Flury (1984) for the common principal components model. Now, we have implemented a robust version, based on trimming and constraints, for the estimation of the set of parsimonious models. To get it, the algorithms were modified, in a very natural way, by incorporating also constraints for the eigenvectors to the classical constraints for the eigenvalues.

Successful proposals for robustifying mixture model estimation in high dimensional settings, still based on trimming and restrictions, have also been presented in the literature. García-Escudero *et al.* (2016b) and Rivera-García *et al.* (2017) introduced robust methodologies for estimating mixtures of factor analyzers and for clustering of functional data, respectively.

Further, fuzzy extensions of TCLUST methodology are available, whose robustness is based on the joint implementation of trimming and constraints. The seminal paper is Fritz *et al.* (2013) for clustering observations in a multivariate space. More recent proposals have been focused on clustering around linear subspaces (Dotto *et al.* , 2016) and mixtures of factor analyzers (García-Escudero *et al.* , 2016b). In this fuzzy context, additional input parameters should be incorporated in the modelization. They are related with the level of fuzziness and the proportion of observations with hard assignment the user wants to have in the obtained solution.

An important issue in all of the previously mentioned proposals is related with the delicate choice of the input parameters. As expected, they have to be provided by the user. García-Escudero *et al.* (2011) proposed exploratory tools for chosing the number of clusters and the level of trimming. To decrease the dependence from the chosen level of trimming, Dotto *et al.* (2017) proposed a reweighted approach. García Escudero *et al.* (2015) and Cerioli *et al.* (2017) analyzed the role of the restriction level c in the estimation. They provided feasible strategies for facing this problem.

A remarkable extension to the TCLUST methodology has been focussed on widening the choice and the features of the component models, to increase its adaptability and to improve goodness of fit. Skew-symmetric models are now available at the component level, in mixture model estimation. Most of the recent research in this setting has been focussed in the skew-t distribution (Lee & McLachlan (2014) provides an in-depth review). This choice is very popular because, additionally, it increases the resistance to the outliers for mixture estimation, by accommodating them with its heavy tails. By applying trimming, the properties of the skew-t become not so crucial when considering model resistance to outliers. In this way, new approaches based on the "basic" skew model, that is the skew normal, have been proposed by García-Escudero, Greselin, Mayo-Iscar (2016).

References

BROWNE, R.P., & MCNICHOLAS, P.D. 2014. Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification*, **8**, 217–226.

CELEUX, G., & GOVAERT, G. 1995. Gaussian parsimonious clustering models. *Pattern Recognition*, **28**, 781–793.

CERIOLI, A., GARCÍA-ESCUDERO, L.A., MAYO-ISCAR, A., & RIANI, M. 2017. Finding the Number of Normal Groups in Model-Based Clustering via Constrained Likelihoods. *Submitted.*

DOTTO, F., FARCOMENI, A., GARCÍA-ESCUDERO, L.A., & MAYO-ISCAR, A. 2016. Robust Fuzzy Clustering via Trimming and Constraints. *Soft Methods for Data Science, Advances in Intelligent Systems and Computing*, **456**, 197–204.

DOTTO, F., FARCOMENI, A., GARCÍA-ESCUDERO, L.A., & MAYO-ISCAR, A. 2017. A Reweighting Approach to Robust Clustering. *Statistics and Computing*, **To appear**.

FLURY, B. N. 1984. Common principal components in k groups. *Journal of the American Statistical Association*, **79**, 892–898.

FRITZ, H., GARCÍA ESCUDERO, L.A., & MAYO-ISCAR, A. 2012. Tclust: An r package for a trimming approach to cluster analysis. *Journal of Statistical Software*, **12**, 1–26.

FRITZ, H., GARCÍA ESCUDERO, L.A., & MAYO-ISCAR, A. 2013. Robust constrained fuzzy clustering. *Information Sciences*, **245**, 38–52.

GARCÍA ESCUDERO, L.A., GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2008. A General Trimming Approach to Robust Cluster Analysis. *Annals of Statistics*, **36**, 1324–1345.

GARCÍA-ESCUDERO, L.A., GORDALIZA, A, SAN MARTÍN, R., VAN-AELST, S., & ZAMAR, R. 2009. FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Journal of the Royal Statistical Society Ser. B*, **71**, 301–319.

GARCÍA-ESCUDERO, L.A., GORDALIZA, A, SAN MARTÍN, R., & MAYO-ISCAR, A. 2010. Robust Clusterwise linear regression through trimming. *Computational Statistics and Data Analysis*, **54**, 3057–3069.

GARCÍA-ESCUDERO, L.A., GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2011. Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, **21**, 585–599.

GARCÍA ESCUDERO, L.A., GORDALIZA, A., & MAYO-ISCAR, A. 2014. A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification*, **8**, 27–43.

GARCÍA ESCUDERO, L.A., GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2015. Avoiding spurious local maximizers in mixture modeling. *Statistics and Computing*, **25**, 619–633.

GARCÍA-ESCUDERO, L.A., GRESELIN, F., & MAYO-ISCAR, A. 2016a. Robust estimation of mixture models with skew components via trimming and constraints. *CFE-CMStatistics 2016 Book of Abstracts*.

GARCÍA-ESCUDERO, L.A., GORDALIZA, A., GRESELIN, F., INGRASSIA, S., & MAYO-ISCAR, A. 2016b. The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. *Computational Statistics and Data Analysis*, **99**, 131–147.

GARCÍA-ESCUDERO, L.A., GORDALIZA, A., GRESELIN, F., INGRASSIA, S., & MAYO-ISCAR, A. 2017. Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *Statistics and Computing*, **27**, 377–402.

LEE, S., & MCLACHLAN, G.J. 2014. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, **24**, 181–202.

Riani, M., Perrotta, D., & Torti, F. 2012. FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, **116**, 17–32.

Rivera-García, D., García-Escudero, L.A., Mayo-Iscar, A., & Ortega, J. 2017. Robust Clustering for Time Series Using Spectral Densities and Functional Data Analysis. *Proc. 14th International Work-Conference on Artificial Neural Networks (IWANN 2017, Cadiz, Spain, June 14-16, 2017). Advances in Computational Intelligence: Lecture Notes in Computer Science-Springer*, **2**, 142–153.

Ruwet, C., García Escudero, L.A., Gordaliza, A., & Mayo-Iscar, A. 2012. The influence function of the TCLUST robust clustering procedure. *Advances in Data Analysis and Classification*, **6**, 107–130.

Ruwet, C., García Escudero, L.A., Gordaliza, A., & Mayo-Iscar, A. 2013. On the breakdown behavior of the TCLUST clustering procedure. *TEST*, **6**, 466–487.