VarSelLCM: an R/C++ package for variable selection in model-based clustering of mixed-data with missing values

Matthieu Marbac[a] and **Mohammed Sedki**[b]

[a]CREST - Ensai
Campus de Ker-Lann, 35172 Bruz, France
matthieu.marbac-lourdelle@gmail.com

[b]University of Paris-Sud and UMR Inserm-1181
Paris, France
mohammed.sedki@u-psud.fr

Keywords : Missing values, Mixed data, Model-based clustering, Variable selection.

Summary: VarSelLCM permits a full model selection (detection of the relevant features for clustering and selection of the number of clusters) in model-based clustering, according to classical information criteria. Data to analyzed can be composed of continuous, integer and/or categorical features. Moreover, missing values are managed, without any pre-processing, by the model used to cluster with the assumption that values are missing completely at random. Thus, VarSelLCM also permits data imputation by using mixture models. A Shiny application is implemented to easily interpret the clustering results. A vignette is available online at http://varsellcm.r-forge.r-project.org/ and the package is freely available to download from the CRAN repository at https://CRAN.R-project.org/package=VarSelLCM/.

Feature selection in clustering facilitates the result interpretation and permits to reduce the variance of the estimators, by assuming that only a subset of the variables explains the unobserved partition. In a non model-based framework, Witten and Tibshirani (2010) propose a feature selection for K-means and hierarchical clustering by using Lasso-type penalty. This method is implemented in the R package *sparcl* (Witten and Tibshirani, 2013) for analyzing continuous data. In a model-based framework, the approaches mainly focus on a single type of variables and only few of them can deal with a large number of features, for computational reasons. We present here the main model-based approaches for model selection, but more references are presented in the review of Fop and Murphy (2017b).
Raftery and Dean (2006) consider the variable selection as a problem of model selection which can be done according to the BIC (Schwarz, 1978). To address the variable selection challenge, three types of variables are defined: the relevant variables (which explains the partition), the redundant variables (which are independent to the partition conditionally on the relevant variables) and the irrelevant variables (which are independent to the partition). This method is implemented in the R package *clustvarsel* (Scrucca and Raftery, 2014) which models the data distribution by a Gaussian mixture. Model selection consists in a maximization of the BIC achieved either via a stepwise greedy search or a headlong algorithm. In order to deal with large numbers of variables, Marbac and Sedki (2017) use a more constrained model which permits to perform model selection before parameter estimation. This model selection is done with respect to the clustering purpose, by an extension of the ICL criterion (Biernacki et al., 2000) named MICL. This method allows large number of features but, for computational reasons, quite large samples (*e.g.,* this approach has been used to cluster $1,318$ individuals described by

160, 470 SNPs).
Dean and Raftery (2010) then Fop et al. (2017) propose Bayesian approaches for feature selection in categorical data clustering. Their method is implemented in the R package *LCAvarsel* (Fop and Murphy, 2017a). Alternatively, Marbac et al. (2018) extend their MICL-based approach to cluster categorical data with large number of features. They also use a modified version of the EM algorithm (Green, 1990) to simultaneously perform model selection with BIC and maximum likelihood inference. These two propositions are complementary, because the second one is designed for large numbers of observations. Finally, their work is extended to clustering of mixed-data (data with continuous, integer, and/or categorical features) where missing values are allowed.
In this talk, we present the R package *VarSelLCM* implementing the two approaches of feature selection proposed by Marbac et al. (2018). To the best of our knowledge, *VarSelLCM* is the only R package which performs a full model selection (selection of the relevant features and the number of clusters) in clustering of mixed-data having possibly missing values.

References

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Dean, N. and Raftery, A. E. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, 62(1):11.

Fop, M. and Murphy, T. B. (2017a). *LCAvarsel: Variable selection for latent class analysis.*

Fop, M. and Murphy, T. B. (2017b). *Variable Selection Methods for Model-based Clustering.*

Fop, M., Smart, K., and Murphy, T. B. (2017). Variable selection for latent class analysis with application to low back pain diagnosis. *Annals of Applied Statistics.*, 11:2085–2115.

Green, P. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):443–452.

Marbac, M., Patin, E., and Sedki, M. (2018). Variable selection for mixed data clustering: Application in human population genomics. *Journal of Classification*, to appear.

Marbac, M. and Sedki, M. (2017). Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, 27(4):1049–1063.

Raftery, A. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Scrucca, L. and Raftery, A. (2014). clustvarsel: A Package Implementing Variable Selection for Model-based Clustering in R. *(submitted to) Journal of Statistical Software.*

Witten, D. and Tibshirani, R. (2010). A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association*, 105(490):713–726.

Witten, D. and Tibshirani, R. (2013). *sparcl: Perform sparse hierarchical clustering and sparse k-means clustering.* R package version 1.0.3.