Clustering and variable selection for high-dimensional complex data

Michael Fop

Model-based clustering is a widely used approach for clustering multivariate data which has been successfully applied in a multitude of fields. The major advantage of this approach is the formulation of the clustering task as a modeling problem, which provides great flexibility and allows to use tools for statistical inference. High-dimensional data poses many challenges to model-based clustering, and the development of variable selection techniques has received a lot of attention in recent years. The main focus is the detection of those variables containing the most useful clustering information, with the purpose of reducing model complexity, enhance classification performance, and improve interpretability.

Despite the research effort, the current available variable selection methods are still prohibitively computationally expensive for high-dimensional data. Variable screening approaches can help to overcome the computational issues, but there is a lack of screening procedures for the effective clustering of high-dimensional and potentially ultrahigh-dimensional data. Moreover, little research has been devoted to the development of variable selection techniques for clustering non-standard data types and data presenting complex dependence structures.

The main goal of this project is to develop a suite of variable selection approaches for model-based clustering of complex and high-dimensional data. The research will focus on developing variable selection methods for clustering mixed-type and time-dependent data, and on developing variable screening techniques which enable the effective application of model-based clustering to datasets of large dimensions. Particular attention will be centered on modeling the complicated dependence structure usually present between variables in high-dimensional settings, as well as on the design of efficient algorithms which can reduce the computational burden of the selection procedure.

References

Fop, M. and Murphy, T. B. (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, 12:18–65.

Marbac, M., Sedki, M., and Patin, T. (2019). Variable selection for mixed data clustering: Application in human population genomics. *Journal of Classification*, pages 1–19.

Pan, R., Wang, H., and Li, R. (2016). Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. *Journal of the American Statistical Association*, 111(513):169–179.