Robust k-means-based clustering for high-dimensional data

P. Filzmoser[a], Š. Brodinová[a,b], T. Ortner[a], C. Breitender[a], and M. Rohm[a]

[a] *TU Wien, Austria,* [b] *Solvistas GmbH, Austria*

We introduce a robust k-means-based clustering method for high-dimensional data where not only outliers but also a large number of noise variables are very likely to be present. Although Kondo et al. [2] already addressed such an application scenario, our approach goes even further. Firstly, the introduced method is designed to identify clusters, informative variables, and outliers simultaneously. Secondly, the proposed clustering technique additionally aims at optimizing required parameters, e.g. the number of clusters. This is a great advantage over most existing methods. Moreover, the robustness aspect is achieved through a robust initialization [3] and a proposed weighting function using the Local Outlier Factor [1]. The weighting function provides a valuable source of information about the outlyingness of each observation for a subsequent outlier detection. In order to reveal both clusters and informative variables properly, the approach uses a lasso-type penalty [4]. The method has thoroughly been tested on simulated as well as on real high-dimensional datasets. The conducted experiments demonstrated a great ability of the clustering method to identify clusters, outliers, and informative variables.

Keywords: k-means, Outliers, High-dimensional data.

References

[1] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander (2000). LOF: Identifying density-based local outliers. In: *ACM Sigmod Record*, 29:93–104.

[2] Y. Kondo, M. Salibian-Barrera, and R. Zamar (2016). RSKC: An R package for a robust and sparse k-means clustering algorithm. *Journal of Statistical Software*, 72:1–26.

[3] A. H. Mohammad, C. Vineet, S. Saeed, and J. Z. Mohammed (2009). Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recognition Letters*, **30**(11):994–1002.

[4] D. M. Witten, and R. Tibshirani (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, **105**(490):713–726.