DSSV 2023 Data Science, Statistics & Visualisation
ECDA 2023 European Conference on Data Analysis

Robust co-clustering for data exploration and anomaly detection in the high-dimensional setting

E. Fibbi[1,2], D. Perrotta[2], F. Torti[2], S. Van Aelst[1], and T. Verdonck[1]

[1]KU Leuven, Department of Mathematics, Celestijnenlaan 200B, Leuven 3001, Belgium
[2]European Commission, Joint Research Centre, Via Enrico Fermi 2749, Ispra 21027, Italy

The exploration and analysis of high-dimensional (HD) data sets calls for well-thought techniques to extract the salient information from the data. Co-clustering, which consists in simultaneously partitioning the rows and columns of a data matrix, can be particularly useful in the HD setting [1]. Latent block models cast co-clustering in a probabilistic framework that extends finite mixture models to the two-way partitioning case.

In addition to being high-dimensional, real-world data sets often contain anomalies which could be of interest *per se* and may make the results provided by standard, non-robust procedures unreliable. Also the estimation of latent block models can be heavily affected by contaminated data. Therefore, we propose a method to compute robust estimates for latent block models. The proposed algorithm combines impartial trimming [2] with a block Classification Expectation-Maximisation (CEM) algorithm, which aims to maximise the complete-data likelihood of the model.

Experiments on both simulated and real data sets show that our method is able to resist high levels of contamination and can provide additional insight into the data by highlighting possible anomalies. Moreover, the analysis can benefit from the model-based approach by leveraging the model likelihood and parameters to obtain further guidance in exploratory analysis, as will be shown with real-world applications. Integration with other exploratory techniques will also be discussed.

[1] G. Govaert and M. Nadif, *Co-Clustering: models, algorithms and applications*. London, England: ISTE Ltd, 2014.
[2] J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán, "Trimmed k-means: an attempt to robustify quantizers," *The Annals of Statistics*, vol. 25, no. 2, pp. 553 – 576, 1997.