Master 2 internship
Development of a deep latent block model for co-clustering

Duration: 6 months from March 2024
Profile : Master 2 or equivalent in statistics, data science or artificial intelligence
Contacts : Charles Bouveyron (charles.bouveyron@inria.fr), Marco Corneli (marco.corneli@univ-cotedazur.fr), and Vincent Vandewalle (vincent.vandewalle@inria.fr), members of the Inria Project Team MAASAI
Location: Centre Inria d'Université Côte d'Azur, 2004, route des Lucioles BP 93 06902 Sophia Antipolis Cedex
Gratification: About 600 euros per month.

Subject

The proposed internship is in the context of **co-clustering** which consists in simultaneously clustering the rows and the columns of an array of data [1], this is particularly useful to summarize large datasets (see Figure 1). A popular probabilistic model co-clustering is the **latent block model** [3](LBM), it assumes that the clusters in each row and each column are drawn independently from two multinomial distributions and that given these clusters all the entries of the data array are independent, and that each entry follows a distribution only depending on its clusters in row and column. In the internship, we propose to develop an extension of the LBM in the case of binary data by assuming that each row and each column can be encoded by a **latent position** in an Euclidean space and that the parameter of the distribution of each entry only depends on these latent positions similarly to [5]. This model performs both co-clustering and visualization of the data through the latent positions as in [2]. For the parameters inference we will consider a variational approach as in [2] by making use of a neural network architecture for the approximate posterior distribution of the latent variables.

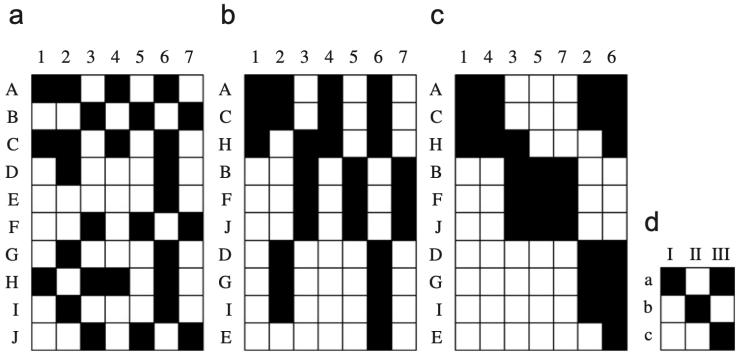

Figure 1: Binary data set (a), data reorganized by a partition on I (b), by partitions on I and J simultaneously (c) and summary binary data (d). Source [4]

Missions

The main mission of the internship is to write the mathematical model and its parameters inference, and perform its implementations on Python. Moreover, the accuracy of the proposed methodology will also be studied on real data sets.
A thesis subject may be proposed as a continuation of this internship.

References

[1] Christophe Biernacki, Julien Jacques, and Christine Keribin. A survey on model-based co-clustering: High dimension and estimation challenges. 2022.

[2] Rémi Boutin, Pierre Latouche, and Charles Bouveyron. The deep latent position topic model for clustering and representation of networks with textual edges, 2024.

[3] Vincent Brault and Mahendra Mariadassou. Co-clustering through latent bloc model: A review. *Journal de la Société Française de Statistique*, 156(3):120–139, 2015.

[4] Gérard Govaert and Mohamed Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, 52(6):3233–3245, 2008.

[5] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.