# Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach

Christine Keribin*[1]

[1]Laboratoire de Mathématiques d'Orsay (LMO) – Centre National de la Recherche Scientifique : UMR8628 – Bâtiment 425 Faculté des Sciences d'Orsay Université Paris-Sud F-91405 Orsay Cedex, France

## Résumé

*Clustering conceptually reveals all its interest when the dataset size considerably increases since there is the opportunity to discover tiny but possibly worth clusters which were out of reach with more modest sample sizes. However, clustering is practically faced to computer limits with such high data volume, since possibly requiring extremely high memory and computation resources. In addition, the classical subsampling strategy, often adopted to overcome these limitations, is expected to heavily failed for discovering clusters in the highly imbalanced cluster case. Our proposal first consists in drastically compressing the data volume by just preserving its bin-marginal values, thus discarding the bin-cross ones. Despite this extreme information loss, we then prove identifiability property for the diagonal mixture model and also introduce a specific EM-like algorithm associated to a composite likelihood approach. This latter is extremely more frugal than a regular but unfeasible EM algorithm expected to be used on our bin-marginal data, while preserving all consistency properties. Finally, numerical experiments highlight that this proposed method outperforms subsampling both in controlled simulations and in various real applications where imbalanced clusters may typically appear.* Collaboration with Filippo Antonazzo and Christophe Biernacki.

*Intervenant