# FROM COMPRESSIVE CLUSTERING TO COMPRESSIVE LEARNING

**Supervisor:** Rémi Gribonval (remi.gribonval@inria.fr)
**Lab:** PANAMA project-team, Inria Rennes / IRISA CNRS UMR 6074.
http:team.inria.fr/panama

## 1. Context

It is often useful to fit a probability model to a data collection, in order to concisely represent the data, to feed learning algorithms that work on densities, to extract features or, simply, to uncover underlying structures. A particularly popular probability model is the **Gaussian Mixture Model** (GMM). Among many other applications, GMM form a central tool to build time-frequency models of audio data that are used for audio source separation [2], and is traditionally fitted through the Expectation-Maximization (EM) algorithm [3]. However, when the collection is voluminous, memory and computation time can be prohibitive.

In a recent paper [1], we proposed a framework to fit a Gaussian Mixture Model (GMM) using only a low-dimensional **sketch** of the data which represents a limited number empirical moments of the underlying probability distribution. The framework is inspired by **compressive sensing** [4], an approach that leverages **sparsity** and **random projections** to characterize and reconstruct high-dimensional vectors through a limited set of random linear projections, under a sparsity assumption. Compressive sensing allows drastic dimension reduction with controlled loss of information, and is associated to reconstruction algorithms of bounded complexity.

By analogy with the algorithms developed for compressive sensing [5,6], we derived a reconstruction algorithm and experimentally showed that it is possible to precisely estimate the mixture parameters of a on GMM with isotropic Gaussians, provided that the sketch is large enough [1]. This resulted in a **compressive clustering** algorithm, in the sense that the centroids of the main clusters of the training collection can be identified from the sketch. Unlike the classical EM approach to GMM estimation, the proposed compressive clustering algorithm consumes an amount of memory independent of the volume of the training collection. It also provides a privacy-preserving data analysis tool, since the sketch doesnt disclose information about individual elements of the training collection.

## 2. Goals

The goal of this internship is to extend and validate the compressive clustering algorithm [1] to handle GMMs with non-isotropic Gaussians. After an implementation and testing stage where the main computational bottlenecks will be identified and addressed, if time allows, comparisons with EM approaches for audio source separation [2,3] will be performed.

## 3. References

**Main reference**
[1] A. Bourrier, R. Gribonval, and P. Perez, Compressive Gaussian Mixture Estimation, ICASSP, Vancouver, Canada, 2013, pp. 60246028.
**Additional references**
[2] A. Ozerov, E. Vincent, and F. Bimbot, A general flexible framework for the handling of prior information in audio source separation, IEEE Trans. Audio, Speech and Language Processing, vol. 20, no. 4, pp. 11181133, 2012.
[3] J. Thiemann and E. Vincent, A fast EM algorithm for Gaussian model-based source separation, 21st European Signal Processing Conference, 2013.
[4] S. Foucart and H. Rauhut, A Mathematical Introduction to Compressive Sensing. Springer, 2012, pp. 1587.
[5] S. Foucart, Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants, Approximation Theory XIII: San Antonio 2010, pp. 6577, 2012.
[6] T. Blumensath, Sampling and Reconstructing Signals From a Union of Linear Subspaces, IEEE Trans. Information Theory, vol. 57, no. 7, pp. 46604671, 2011.
**See also:**
https://team.inria.fr/panama/projects/please/
http://bass-db.gforge.inria.fr/fasst/