Master Internship in Data Science & Artificial Intelligence

# Large Scale Model-Based Clustering of Distributed Data

**Context of the internship:** This intership will be performed in the framework of the ANR project SMILES-**S**tatistical **M**odeling and **I**nference for unsupervised **L**earning at larg**E**-**S**cale, which is a collaborative fundamental research project funded by ANR in the framework of the plan of the French state towards Artificial Intelligence. SMILES gathers scientists from four research organisms, UMR CNRS LMNO, UMR CNRS LMRS, UMR CNRS LIS, and INRIA Modal. This internship will be held in LMNO - the Lab of Mathematics Nicolas Oresme in Caen, related to the questions currently studied in a PhD thesis (the one of Thien Nhat Pham), so primarily as a support to the PhD studies, but we encourage and will be open to any interesting propositions within the topics of SMILES.

**Subject description:** The internship will consist of research & development activities expected to be mainly guided by the following directions. The objective if to develop *proven large-scale model-based clustering via distributed Latent Data Models (LDM)*. The use of distributed processing is a natural way to proceed in the analysis of a big volume of data. This raises the key question of how to distribute data while controlling the quality of estimators. In this research direction, we investigate ensemble learning methods and collaborative mixtures for large-scale model-based clustering. We approximate the overall density of the heterogeneous data by mixtures thanks to their universal approximation property.

The problem of data clustering becomes the one of estimating the mixture model parameters on each distributed site and then aggregate the resulting local estimators, to provide an overall proven aggregated estimator. Recent results in supervised learning promote using ensemble learning (bootstrap) to the scaled analysis of massive data (Kleiner et al., 2014), which will be investigated in an unsupervised context. The robust aggregation of local estimates is another important issue and can be handled by optimizing similarity measures like the Kullback-Leibler divergence, or, through a hierarchical mixtures aggregation, for instance through a mixture of experts modeling (Jacobs et al., 1991). The practical question of distributed computing can be performed within standard frameworks such as MapReduce or Hadoop, etc.

Scalable mixture models will therefore be the guideline of this research. This will involve estimating mixtures for various types of data (continuous data to start with), from millions of individuals distributed (by nature or by (re)sampling) and thousands of possibly correlated variables by aggregating parsimonious local estimators constructed on parallel computing resources, and estimating the variances of the resulting estimators. We seek guarantees in model estimation and selection. First, the novelty is to aggregate these criteria $\text{Crit}(\mathscr{D}_1,\widehat{M}_1),\ldots,\text{Crit}(\mathscr{D}_B,\widehat{M}_B)$ associated to models $(M_1,\ldots,M_B)$ (e.g. represented by $\theta$'s for parametric mixtures) which will be built respectively from small sub-samples $(\mathscr{D}_1,\ldots,\mathscr{D}_B)$ on parallel computations, to have pseudo-criteria of large samples. Then, selection criteria like BIC (Schwarz, 1978) should be transposed to such a large-scale distributed framework for model selection.

**Required profile:** The successful candidate should have *i)* very good skills to communicate in English, *ii)* pursuing a master 2 or an engineering degree in an area related to mathematical/statistical sciences or computer science with *iii)* strong skills in statistical inference, Machine learning and proven `Matlab`, `Python` or `R` programming. Experience using Github, Slack or other collaborative tools, and big-data technologies (e.g Spark) are a strong plus.

**Dates:** Application deadline: 31 march 2021. Starting date expected by april 2021, for 4 to 6 months. Please note that as we will not be able to respond to all the applications, if you don't hear from us within a week after your application, that means we didn't find a strong match with your profile.

**Internship gratification:** $\sim 550$ € net per month.

**How to apply:** Please send, in **A SINGLE .pdf FILE**, your CV, transcripts and a recommendation letter(s) or reference name(s), to Thien Pham and Faïcel Chamroukhi on nhat-thien.pham@unicaen.fr, chamroukhi@unicaen.fr.

## References


Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.