Master Internchip in Data Science & Artificial Intelligence

Latent Data Models for Model-Based Clustering and Co-Clustering of High-Dimensional Data

Statistical Modeling and Inference for unsupervised Learning at largE-Scale (SMILES) is a collaborative fundamental research project funded by ANR (2018-2022) in the framework of the plan of the French state towards Artificial Intelligence (AI). Large-scale data analysis is an inherently multidisciplinary area and is becoming of broader interest for today's society. SMILES aims at introducing an unsupervised statistical modeling framework and scaled inference algorithms for transforming large-scale data into knowledge. It considers the large-scale context as a whole, with its main issues related to inference from a big volume of data of very high dimension and underlying complex hidden structures. The key tenet of SMILES is to introduce large-scale latent data models (LDM) for unsupervised data classification and representation. The knowledge extraction will namely consist in automatically retrieving hidden structures, summarizing prototypes, groups, sparse representations. We consider different data settings, including functional data, multimodal bioacoustical data, and biological data. SMILES gathers experts in statistical modeling, inference, optimisation, sparse representation, information processing and machine learning. The consortium is composed of four research organisms: UMR CNRS LMNO, UMR CNRS LMRS, UMR CNRS LIS, INRIA Modal.

LDM for clustering and segmentation of functional data: The first part of the internship consists of participating to the implementation, in `R` or in `C`, of unsupervised learning algorithms for clustering and segmentation of high-dimensional data written in Matlab and Python (see `https://github.com/fchamroukhi` and `https://chamroukhi.users.lmno.cnrs.fr/software.php` for details), which include:

- Latent data models for time series segmentation (RHLP, PWR, HMMR in the univariate and the multivariate case)
- Latent data models for simultaneous functional data clustering and segmentation (MixRHLP, MixHMMR, PWRM)
- Non-normal and robust mixtures-of-experts (NNMoE) (SNMoE, TMoE, STNMoE)
- Unsupervised learning of regression mixture models with unknown number of components (RobustEM-PRM, SRM, bSRM)

The submission of `R`-packages and software papers to the the `R`-software journal will be particularly encouraged.

Large-scale LDM and inference for functional data: The second expected part of the internship is to further consider the problem of high-dimensional functional data clustering, where each individual is described by a set of functions. To address this issue, we propose latent functional block models for co-clustering high-dimensional time series. Most of these statistical analyses in model-based co-clustering are multivariate. However, in many application domains, the observations are issued from underlying continuous functions (e.g., curves) and therefore a standard classical multivariate co-cluster analysis may be not adapted. This context is quite new and is

considered in the project. We consider this problem of co-clustering of functional data here and propose to deal with by functional latent block model (FLBM) to simultaneously cluster a sample of multivariate functions into a finite set of blocks, each block being an association of cluster over individuals and a cluster over functional variables (see our recent work [1]). In our proposal for dealing with large-scale functional data, we associate both model-based co-clustering to the framework of FDA to model the density of the observed discretized function (x,y) by the FLBM $f(Y|X;\theta)=\sum_{(z,w)\in\mathscr{Z}\times\mathscr{W}}\mathbb{P}(Z=z,W=w)f(Y|X,Z=z,W=w;\theta)$. This is very new paradigm and we are one of the first contributors [1, 3]. Furthermore, we are interesting to models which are able to discover more complex structure, that co-clustering the data, that is, segmenting each homogeneous cluster governed by a dynamical hidden structure, into regimes [1]. A regression model with a hidden process (e.g [2]) may be used as a conditional block density $f(Y|X,Z=z,W=w;\theta)$. The obtained functional latent block model is estimated by a variational EM algorithm or an MCMC sampling via a stochastic EM extension.

Additional information:
Required profile: Successful candidates are master 2 students in mathematical/statistical sciences or Machine learning with strong skills in statistical inference and in programming with `R` or `C`, and `Matlab`.
Expected starting date: feb/mar/aprl 2019, for 4 to 6 months
Application deadline: you can apply while the position is not indicated as filled
Salary: 550 € net (all taxes included) per month.
PhD Director: Faïcel Chamroukhi (Principal Investigator of SMILES): http://math.unicaen.fr/∼chamroukhi/
Institution: University of Caen - The Lab of Mathematics Nicolas Oresme - UMR CNRS, France.
How to apply: Please send your application file (CV+transcripts of the master programm and possible recommendation letters) in **A SINGLE .pdf FILE** to chamroukhi@unicaen.fr.

References

[1] Chamroukhi, F. and Biernacki, C. (2017). Model-Based Co-Clustering of Multivariate Functional Data. In *ISI 2017 - 61st World Statistics Congress*, Marrakech, Morocco.

[2] Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. (2009). Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602.

[3] Slimen, Y. B., Allio, S., and Jacques, J. (2018). Model-based co-clustering for functional data. *Neurocomputing*, 291:97 – 108.