PhD position in Data Science & Artificial Intelligence

Latent Data Models for Large-Scale Clustering

Statistical **M**odeling and **I**nference for unsupervised **L**earning at larg**E**-**S**cale (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.

The PhD research will be guided by the two following directions:

Proven large-scale model-based clustering via distributed 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 propose to approximate the overall density of the heterogeneous data by mixtures thanks to their universal approximation property. Indeed, we can consider the global density of the data as a mixture density, thanks to the property of denseness of mixtures, as well as their performance in clustering. 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 [11], 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, notably through a mixture of experts modeling [7, 15]. Mixture of experts are based on the principle of divide and conquer and are dense (e.g., see [16, 2]) and could therefore be adapted for distributed mixture modeling and aggregation. The practical question of distributed computing can be performed within standard frameworks such as MapReduce or Hadoop, etc. Attention will simultaneously be given to the problem of high-dimension namely by regularization in latent data models, which is less common than the well-known supervised variable-selection problem in supervised learning. We intend to investigate in particular high-dimenional mixture and mixture of experts (e.g., see [8, 10, 9]) as it opens modeling issues and computational challenges related to inference, following our very recent results [5, 4, 15].

Scalable mixture models will therefore be the general guideline of this research. This will involve estimating mixtures for various types of data 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 are in selection and estimation. 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 θ for a parametric mixture model) which will be built respectively from small sub-samples $(\mathscr{D}_1,\ldots,\mathscr{D}_B)$ on parallel computers, to have pseudo-criteria of large samples. Then, these selection criteria like BIC, AIC, ICL [1] will be transposed to a more general framework of the selection of "number of clusters or blocks, number of latent dimensions, complexity of the correlation matrix", with the aim of deriving possibly analytic criteria, while the latest work on the criteria cited concern the selection of the number of blocks) by approximate criteria [12, 13, 14].

Large-scale LDM and inference for functional data: This research direction aims at developing models and algorithms for large-scale data arised in functions with hidden dynamical structures. It will be focused on the field of "Functional Data Analysis (FDA)". Functional Data (FD) are discretised values of very high dimension observed from possibly infinite dimensional smooth curves or surfaces. In addition to their big volume and their very high dimension, FD are in general very structured and the underlying structure is always hidden. Thus, the third complexity of analyzing FD resides in their latent complex structure that has to be retrieved in an unsupervised way. The scientific achievement expected from this part of the project consists in developing latent data models and dedicated learning algorithms from large-scale functional data with complex hidden structure.

We also further consider the problem of high-dimensional functional data, 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 [3]). In our proposal for dealing with large-scale functional data, we associate both model-based co-clustering with 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 [3, 17]. 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 [3]. A regression model with a hidden process (e.g [6]) 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 should have a master degree in mathematical/statistical sciences, Machine learning, statistical signal processing, or a closely related area, and strong skills in statistical inference and in programming with Matlab and/or R and/or Python. The PhD thesis as well all the research reports/papers will be written in English. So strong skills in English writing/speaking are needed. Expected skills include unsupervised learning, model-based clustering, and distributed large-scale algorithms computing. International applications are welcome to join our international team (there is no any required French skills). For candidates who wish to learn French, free courses are offered by the university/the doctoral school to foreign students.

Expected starting date: nov-2018 - Jan 2019 (for 36 months)
Application deadline: you can apply while the position is not indicated as filled
Salary: 1900 € gross per month (Attractive slaray!). We also offer to the PhD student the possibility to teach courses in statistics/data analysis (up to 64 hours a year) for undergraduate students at the department of Mathematics & Computer Science.
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 last three academic years) in **A SINGLE .pdf FILE** to chamroukhi@unicaen.fr.

References

[1] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

[2] Chamroukhi, F. (2016). Robust mixture of experts modeling using the t distribution. *Neural Networks*, 79:20–36.

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

[4] Chamroukhi, F. and Huynh, B. (2018a). Regularised maximum-likelihood inference of mixture-of-experts models. In *The International Joint Conference on Neural Networks (IJCNN)*, Rio.

[5] Chamroukhi, F. and Huynh, B. T. (2018b). Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. *submitted to the journal of the French Statistical Society SFDS*.

[6] 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.

[7] 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.

[8] Khalili, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4):519–539.

[9] Khalili, A. (2011). An overview of the new feature selection methods in finite mixture of regression models. *Journal of The Iranian Statistical Society*, 10(2):201–235.

[10] Khalili, A., Chen, J., and Lin, S. (2011). Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. *Biostatistics*, 12(1):156–172.

[11] 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.

[12] Lomet, A. (2012). *Sélection de modèle pour la classification croisée de données continues*. Ph.D. thesis, Université de Technologie de Compiègne.

[13] Lomet, A., Govaert, G., and Grandvalet, Y. (2012a). An approximation of the integrated classification likelihood for the latent block model. In *ICDM Workshops*, pages 147–153.

[14] Lomet, A., Govaert, G., and Grandvalet, Y. (2012b). Model selection in block clustering by the integrated classification likelihood. In *20th International Conference on Computational Statistics*, pages 519–530.

[15] Nguyen, H. D. and Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pages e1246–n/a.

[16] Nguyen, H. D., Lloyd-Jones, L. R., and McLachlan, G. J. (2016). A universal approximation theorem for mixture-of-experts models. *Neural Computation*, 28(12):2585–2593.

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