Postdoc/PhD-research Eng positions in Data Science. Project AStERiCs:
Statistical learning for large-scale unsupervised data representation and classification

[French version here: https://chamroukhi.users.lmno.cnrs.fr/ASTERICS/AStERiCs-PostDocs-IGRs-fr.pdf]

Starting date: Earlier 2018
How to apply : send your complete CV (including a full publication list and if possible a link to your PhD thesis) to faicel.chamroukhi@unicaen.fr and indicate the number of the position for which you apply
International applications are welcome

Position 1 : Postdoc (duration18 months):

The objective of this postdoc research is to develop statistical models for unsupervised representation and classification of high-dimensional continuous data with possibly infinite dimension, and to develop optimized inference algorithms.

We will mainly focus on sparse mixtures and the case where the space of the latent variable can be of infinite dimension. The inference of such models in a large-scale context (very large dimension) requires the control of the optimization problem when carrying out the maximum likelihood estimation and suggests new strategies of regularization (unsupervised framework). We can rely on penalized log-likelihood criteria where the penalty should take into account the missing data (problem of unsupervised feature selection) and its possible organization in hierarchy (structured regularization). These problems of regularization in an unsupervised context (clustering and feature selection) are quite recent (Devijver (2015a, b)., G. Celeux, et al. (2011), Ruan et al. (2011), Witten & Tibshirani, (2010)), particularly regarding functional data (Devijver (2015b)). We will also be interested in the extension of these models to regularized mixtures of experts (an ongoing work).

- ***Required Profile*** :
  - PhD in statistics/statistical leaning
  - An experience in learning representations and classification of massive data
  - Skills in working on real-world data and applications
  - Programming skills in Matlab/R/Python
- **Hosting lab**: The lab of mathematics Nicolas Oresme, UMR CNRS
- **Salary**: ~ 2700€ gross (~ 2250€ nets) per month

Position 2 : Postdoc (duration18 months):

The objective of this postdoc research is to develop statistical models for unsupervised representation and classification of high-dimensional continuous data with possibly infinite dimension, and to develop optimized inference algorithms.

Providing accurate and flexible answers to multiform classification problems, finite mixture distributions have become today an extremely studied tool and used successfully in various disciplines (genetics, image processing, astronomy ...). We will rely on discrete mixture models and we can start by extending the work of Karlis and Meligkotsidou (2007) and Shi and Valdez (2014) to multivariate and overdispersed data. Models of non-or semi-parametric mixtures will also be considered based on kernel methods as in Benaglia et al. (2009). We will examine how to calibrate the kernel smoothing parameter (the window) by recent selection methods as in Goldenshluger and Lepski (2011) or Lacour et al (2017). These methods have been shown to outperform conventional cross-validation criteria in terms of computation time (e.g., Chagny and Roche 2015), which is crucial in this context of large-scale data. The results can also be extended to the sequential modeling of data by Markovian models. The developed algorithms will be applied to the processing of genomic data (of the RNA-Seq type) and in particular the differential analysis of genes.

- ***Required Profile*** :

- PhD in statistics/statistical leaning
- An experience in learning representations and classification of massive data
- Skills in working on real-world data and applications
- Programming skills in Matlab/R/Python

- **Hosting lab**: The lab of mathematics Raphael Salem, UMR CNRS
- **Salary**: ~ 2700€ gross (~ 2250€ nets) per month

Position 3 : Phd Research-Engineer (duration 18 months) :

One of the major objectives of the project is to create a public scientific and technical platform dedicated to unsupervised statistical learning from large-scale data (BigData). This platform will propose a complete architecture (pre-treatment, representation, classification, visualization) and will consider the following aspects of the data: high-dimension, big volume, heterogeneity. It will propose original algorithms for the analysis of heterogeneous data of different types (continuous, longitudinal / functional, discrete) with high-performance computing (the CRIANN can provide us with high performance distributed computing resources).

The candidate will participate in collaboration with the rest of the project staff to the creation of this platform. The first step consists in prototyping of algorithms already developed at the LMNO lab, their integration into this platform on various real applications (environmental monitoring, time series, genomic sequences, etc.). These algorithms are unsupervised classification algorithms based on latent variable models for different types of high-dimensional data. The second step is to contribute to the integration of the algorithms developed during the project. This second task is more fundamental and will focus in particular on distributed clustering using mixture models. For that we will rely on the bootstrap theory to infer a latent variable model (eg mixture model) from a big volume of data, for which parallel computing is a natural way to proceed especially for batch mode data processing. The issues to be addressed in this context are mainly i) obtaining guarantees and new aggregation strategies for local estimators, i.e how to obtain an ``optimal'' estimator as an aggregation of several estimators constructed from bootstrap samples, and ii) deal with the problem of model selection which in this distributed case consists of aggregating model selection criteria, constructed from small subsamples to have pseudo-criteria of large samples.
The main technical missions are: (i) Prototyping unsupervised learning algorithms (ii) High performance distributed cloud computing (iii) Web integration and interfacing with the platform.

- ***Required Profile*** :
  - PhD in statistics/statistical leaning or computer science with specialization in machine learning
  - Experience in latent data models and complex data analysis
  - Experience in software development
  - Skills in working on real-world data and applications
  - Programming skills in Matlab/R/Python
  - Skills on big-data platforms (Hadoop/Spark, MapReduce), Cloud computing, OLAP, web technologies
- **Hosting lab**: The lab of mathematics Nicolas Oresme, UMR CNRS
- **Salary**: ~ 2700€ gross (~ 2250€ nets) per month

Position 4 : Phd Research-Engineer (duration 14 months) :

One of the major objectives of the project is to create a public scientific and technical platform dedicated to unsupervised statistical learning from large-scale data (BigData). This platform will propose a complete architecture (pre-treatment, representation, classification, visualization) and will consider the following aspects of the data: high-dimension, big volume, heterogeneity. It will propose original algorithms for the analysis of heterogeneous data of different types (continuous, longitudinal / functional, discrete) with high-performance computing (the CRIANN can provide us with high performance distributed computing resources).

The candidate will participate in collaboration with the rest of the project staff in the creation of this platform. The first step concerns prototyping of algorithms already developed at the LMRS lab, their

integration into this platform on various real applications. These are unsupervised classification algorithms based on latent variable models for discrete (eg genomic) data and functional data analysis by non-parametric methods. The second step is to participate in integrating the algorithms that will be developed during the project. This second task will be carried out in collaboration mainly with the LMNO lab on distributed regularized mixture models with environmental applications / genomic sequences.

The main technical missions are: (i) Prototyping unsupervised learning algorithms (ii) High performance distributed cloud computing (iii) Web integration and interfacing with the platform.

- ***Required Profile*** :
  - PhD in statistics/statistical leaning or computer science with specialization in machine learning
  - Experience in latent data models and complex data analysis
  - Experience in software development
  - Skills in working on real-world data and applications
  - Programming skills in Matlab/R/Python
  - Skills on big-data platforms (Hadoop/Spark, MapReduce), Cloud computing, OLAP, web technologies
- **Hosting lab**: The lab of mathematics Raphael Salem, UMR CNRS
- **Salary**: ~ 2700€ gross (~ 2250€ nets) per month

Some references related to the project:

F. Chamroukhi, (2017) “Skew *t* mixture of experts”, *Neurocomputing*, V266, pp. 390-408.

F. Chamroukhi, (2016) “Robust Mixture of Experts modeling using the *t* distribution”, *Neural Networks*, V79, pp 20-36.

F. Chamroukhi, (2016) “Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation”, *Journal of Classification*, 33(3):374–411.

F. Chamroukhi, (2016) “Unsupervised learning of regression mixture models with unknown number of components”, *Journal of Statistical Computation and Simulation*, V86.12, pp. 2308-2334.

F. Chamroukhi, H. Glotin & A. Samé (2013) “Model-based functional mixture discriminant analysis with hidden process regression for curve classification”, *Neurocomputing*, 112:153-163.

F. Chamroukhi, S. Mohammed, D. Trabelsi, L. Oukhellou, Y. Amirat, (2013) “Joint segmentation of multivariate time series with hidden process regression for human activity recognition”, *Neurocomputing*, 120: 633-644.

J.-L Starck, F. Murtagh, M.J. Fadili, (2016) “*Sparse Image and Signal Processing: Wavelets and Related Geometric Multiscale Analysis*”, (2nd Edition, ed.), Cambridge University Press, Cambridge, ISBN 9781107088061.

C. Chesneau, M.J. Fadili, B. Maillot, (2015) “Adaptive estimation of an additive regression function from weakly dependent data”, *J. of Multivariate Analysis*, V.133.1, pp. 77-94.

C. Bérard, M. Seifert, T. Mary-Huard, M-L. Martin-Magniette, (2013) “MultiChIPmixHMM : an R package for ChIP-chip data analysis modeling spatial dependencies and multiple replicates”. *BMC Bioinformatics*, 14 :271.

S. Volant, C. Bérard, M-L. Martin-Magniette, S. Robin, (2014) “Hidden Markov Models with mixture as emission distribution”. *Statistics and Computing*, 24(4):493–504.

G. Chagny, Roche, A. (2015) “Adaptive estimation in the functional nonparametric regression model”, *J. of Multiv. analysis*. V146, pp. 105-118.

G. Chagny, C. Lacour, (2015) “Optimal adaptive estimation of the relative density”, *TEST* 24(3) : 605-631.

G. Chagny (2015) “Adaptive warped kernel estimators”, *Scandinavian Journal of statistics*, 42(2) : 336-360.

A. Channarond, J.-J. Daudin, S. Robin, (2012) “Classification and estimation in the Stochastic Blockmodel based on the empirical degrees”, *Electronic Journal of Statistics*, 6 : 2574-2601

G. Chagny, A.Roche, (2014) “Adaptive and minimax estimation of the cumulative distribution function given a functional covariate”. *Electronic Journal of Statistics*, 8 : 2352-2404.

A. Patel, T. Nguyen, R. Baraniuk (2016) “A Probabilistic Framework for Deep Learning”. In NIPS, Barcelona.

A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan (2014) “A scalable bootstrap for massive data”. *JRSS B*, 76(4):795-816.

D. Witten, and R. Tibshirani, (2010) “A framework for feature selection in clustering”. *Journal of the American Statistical Association*, 105(490):713-726.

L. Ruan, M. Yuan., H. Zou (2011) “Regularized parameter estimation in high-dimensional Gaussian mixture models”. *Neural Computation*, 23:1605-1622.

G. Celeux, M.-L. Martin-Magniette, C. Maugis, A.E. Raftery, (2011) Letter to the editor: “A framework for feature selection in clustering”. *Journal of the American Statistical Association*, 106:383.

E. Devijver (2015a) “An l1-oracle inequality for the Lasso in finite mixture of multivariate Gaussian regression models”.

*ESAIM:PS*19. 649-670.

E. Devijver (2015b) “Finite mixture regression: a sparse variable selection by model selection for clustering”, *Electronic Journal of Statistics* 9(2), pp. 2642-2674.

D. Karlis, L. Meligkotsidou, (2007) "Finite mixtures of multivariate Poisson distributions with application". *Journal of Statistical Planning and Inference*, 137(6), pp. 1942-1960.

P. Shi, E.A. Valdez, E. A. (2014) "Multivariate negative binomial models for insurance claim counts". *Insurance: Maths. & Economics*, 55, pp. 18-29.

T. Benaglia, Chauveau, D. Hunter, D. R. (2009) "An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures", *Journal of Computational and Graphical Statistics*, 18(2), pp. 505-526.

A. Goldenshluger, 0. Lepski, (2011) "Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality". *The Annals of Statistics*, 39(3), pp. 1608-1632.

C. Lacour, P. Massart, and V. Rivoirard, (2016). Estimator selection: a new method with applications to kernel density estimation.