Ph.D. offer - Institut 3IA Côte d'Azur
Université Côte d'Azur & INRIA
Deep Latent Variable Models for the Analysis of Massive Heterogenous Data

Advisor and location:

- Team: Maasai project-team, Université Côte d'Azur & Inria
- Advisor: Pr. Charles Bouveyron
- Localization: Maasai project-team, Centre Inria d'Université Côte d'Azur, 2004 Route des Lucioles, 06902 Sophia-Antipolis

Context and project: In all aspects of everyday life, there is a massive digitalization of systems that is increasingly important. One of the consequences of this phenomenon is the massive production of data, especially heterogenous data made of several data types (continuous, texts, images, times series, networks, …). For example, social and communication networks allow users to interact through text, images, networks, … A similar situation can be encountered in the context of medical data, where the data types may be even more large. It is therefore of strong interest to be able to analyze those massive heterogenous data using information extracted from all data types, in particular in the context of unsupervised learning.

The purpose of this Ph.D position, within the Institut 3IA Côte d'Azur (Univ. Côte d'Azur & INRIA), will be focused on **the development and the understanding of deep latent variables models for unsupervised learning with massive heterogenous data**. Although deep learning methods and their statistical extensions, the deep latent variables models (DLVM) [1], allowed clear advances in Artificial Intelligence in the last 5 years, they clearly suffer from an overall weak knowledge of their theoretical foundations and behavior, in particular in the context of unsupervised learning. These issues are indeed barriers to the wide acceptation of the use of AI in sensitive applications, such as medicine, transport, or defense. On the technical side, we aim at combining statistical latent variable models with deep learning algorithms to justify existing results and allow a better understanding of their performances and their limitations. In particular, in the context of unsupervised learning with deep latent variable models, the question of dealing with massive heterogenous data and the question of model selection (choosing the number of clusters, the dimension of the latent spaces, the architecture selection, …) are almost unexplored at the moment. The preliminary results we obtained in the past years have shown that DLVM models can be extended with success to at least two different types of data (network and texts, text and images, …) but the extension to several data types is still difficult in the sound generative context. A first goal of this Ph.D. will be to propose a generative DLVM model specifically designed for massive heterogenous data. Regarding the problem of model selection in this context, some preliminary studies we performed have highlighted the surprising fact that the evidence lower bound of the fitted model may be used as a model selection criterion in some extent. This strongly suggests revisiting the study of these latent variable models with a Bayesian point of view and to understand how this evidence lower bound integrate implicit priors on the latent variables. Having a clear understanding of this point will offer an elegant and powerful tool for picking the appropriate model (latent dimensions, network architecture, network sparsity, …) for the data at hand.

The proposed methodologies will be then applied to real-world situations in either Medicine (Pharmacovigilance, omics-based clinical discovery, …) or Digital Humanities (History, Archeology, …).

Expected skills: The candidate should have a graduate degree (Master 2 degree). Him/her scholar background should include:

- statistical/machine learning, statistical inference, clustering, classification
- deep learning, variational auto-encoder, back-propagation,
- knowledge of R (main programming language), Python and C++.

Application: Application files should contain a resumé, an application letter and grade records of the 2 last years (M1 & M2).
Applications should be sent by email to charles.bouveyron@inria.fr.

References:

- D. Kingma and M. Welling, Auto-encoding variational bayes, stat, 1050:1, 2014.
- Jiang, Z., Zheng, Y., Tan, H., Tang, B., and Zhou, H, Variational deep embedding: An unsupervised and generative approach to clustering, In International Joint Conferences on Artificial Intelligence Organization, 2017.
- C. Bouveyron, P. Latouche and R. Zreik, The Stochastic Topic Block Model for the Clustering of Networks with Textual Edges, Statistics and Computing, vol. 28(1), pp. 11-31, 2017
- C. Bouveyron, M. Corneli, P. Latouche and D. Liang, *Clustering by Deep Latent Position Model with Graph Convolutional Network*, Advances in Data Analysis and Classification, in press, 2024
- C. Bouveyron, M. Corneli and G. Marchello, *A Deep Dynamic Latent Block Model for the Co-clustering of Zero-Inflated Data Matrices*, Journal of Computational and Graphical Statistics, in press, 2024
- A. Destere, G. Marchello, D. Merino, N. Ben Othman, A. Gérard, T. Lavrut, De. Viard, F. Rocher, M. Corneli, C. Bouveyron and M. Drici, *An artificial intelligence algorithm for co-clustering to help in Pharmacovigilance before and during the COVID-19 pandemic*, British Journal of Clinical Pharmacology, in press, 2024
- R. Boutin, C. Bouveyron and P. Latouche, *Embedded Topics in the Stochastic Block Model*, Statistics and Computing, vol. 33(5), pp. 1-20, 2023