Ph.D. Thesis offer
Advanced discrete optimization meets machine learning

Supervisors and location:

- ***Team:*** LIMOS (Clermont Auvergne INP) & LMBP (Université de Clermont-Auvergne)
- ***Supervisor:*** Pierre Latouche, Professeur des Universités, Statistiques, Machine Learning.
- ***Co-Supervisors:*** Renaud Chicoisne, Maître de conférences, Recherche Opérationnelle.
  Rodolphe Le Riche, Directeur de recherche CNRS, Processus Gaussiens.
- ***Location:*** LIMOS, INP-UCA, 1 rue de la Chebarde, 63178 Aubière

Contact : renaud.chicoisne@uca.fr, pierre.latouche@math.cnrs.fr, rodolphe.le_riche@uca.fr

Context and project: In the last 15 years, there has been a particularly strong interest in the development of machine learning (ML) and artificial intelligence (AI) algorithms: In fact, for a large set of tasks such as regression, classification, recommendation, and clustering, ML/AI approaches have been shown to outperform alternative strategies.

In particular, sparse regression models aim to estimate parameters to approximate the data at hand, while involving only a limited number of the most relevant parameters, which can be done by maximizing a likelihood function with a hopefully low number of relevant parameters. The latter can be modeled with nonlinear objective functions, while the former gives birth to integer variables in the resulting optimization problem.

In many real-life applications, these optimization problems are not solved exactly and instead, approximate solutions produced by heuristics and relaxations are used. While these approximate solutions can be enough in several contexts, the aim of this Ph.D. Thesis is to solve the sparse regression problems exactly or with a provable degree of approximation using advanced optimization algorithms.

This project starts with the idea that the use of recent Operations Research (OR) algorithms to optimize ML objective functions will be a key ingredient to improve the performances of AI/ML methods and their application to real world problems, such as traffic congestion prediction or frost event prediction in the agro-industry [5].

Objectives: This Ph.D. project is in OR applied to ML. Broadly speaking, at the core of this research field is the will to rely on strong optimization techniques to leverage existing and relevant models for ML/AI. Follows a detailed description of some ideas behind the Ph.D., that candidates are not expected to fully understand before the project.
In this PhD we aim first at addressing the problem of sparse Bayesian variable selection for high-dimensional linear and nonlinear regression. Thus, we will consider a generative model that uses a spike-and-slab-like prior distribution obtained by multiplying a deterministic binary vector, which conveys the sparsity of the problem, with a random Gaussian parameter vector.

Then, the originality of this work is to propose modern discrete optimization techniques to optimize the type-II log-likelihood of the model. In [1-3], a simple relaxation along with an Expectation-Maximization (EM) algorithm were employed for inference. Competitive results on simulated and real data were obtained in comparison with state-of-the-art techniques.

We plan here to address the optimization task directly - without relaxation - in order to improve the selection of the variables even further. The resulting maximum-likelihood problem can be cast as a high-dimensional, mixed-integer nonlinear optimization program that can be solved via a tailored cutting plane algorithm combined with piecewise linear approximation techniques. We aim at addressing the selection

of the input variables along with the dimension of the latent factors, in the context of variational auto-encoders [4]. Such a problem has received strong attention in the last couple of years with no satisfying solutions.

The algorithms developed in this PhD project are meant to process large scale data sets from 1) traffic/congestion information from Santiago de Chile (with the partner Univ. de Chile) to predict short-term congestion and 2) soil and weather information provided by our partner Instacrops to predict frost events accurately [5].

Candidate background :

the candidate should hold a Master degree or equivalent with good skills in applied mathematics, in relation to optimization and operations research. Being knowledgeable in machine learning and/or statistics is a plus. The candidate should also like programming and be willing to learn CPLEX.

Salary : **Nationwide standard French Ph.D. student income (See LPR)**

References:

[1] Latouche, P., Mattei, P. A., Bouveyron, C., & Chiquet, J. (2016). Combining a relaxed EM algorithm with Occam's razor for Bayesian variable selection in high-dimensional regression. *Journal of Multivariate Analysis*, *146*, 177-190.

[2] Bouveyron, C., Latouche, P., & Mattei, P. A. (2018). Bayesian variable selection for globally sparse probabilistic PCA. *Electronic Journal of Statistics*, *12*(2), 3036-3070.

[3] Bouveyron, C., Latouche, P., & Mattei, P. A. (2020). Exact dimensionality selection for Bayesian PCA. *Scandinavian Journal of Statistics*, *47*(1), 196-211.

[4] *K*ingma, D. P., & Welling, M. (2014, April). Stochastic gradient VB and the variational auto-encoder. In Second International Conference on Learning Representations, ICLR (Vol. 19, p. 121).

[5] Diedrichs, A. L., Bromberg, F., Dujovne, D., Brun-Laguna, K., & Watteyne, T. (2018). Prediction of frost events using machine learning and IoT sensing devices. *IEEE Internet of Things Journal*, *5*(6), 4589-4597.