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

Supervisors and location:

Teams: LIMOS / LMBP

Supervisor: Pierre Latouche, Professeur des Universités, Statistiques, Machine Learning et IA

Co-Supervisors:

1 Renaud Chicoisne, Maître de conférences, Recherche Opérationnelle

2 Rodolphe Le Riche, Directeur de Recherche CNRS, Processus Bayesiens

Location: LIMOS, INP-UCA, 1 rue de la Chebarde, 63178 Aubière, https://limos.fr

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

About UCA: Université Clermont Auvergne has a long history of contributions in operations research, probability, and statistics. UCA belongs to the top group of Universities in France which received the IDEX / I-SITE label from an international jury.

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. Closely related to computational statistics (CS), ML and AI algorithms are designed by optimizing their own, specific objective (loss) functions. Relying on probabilistic models, CS approaches have the key advantage of being interpretable and to take uncertainty into account. Starting from a statistical model with unknown parameters and complexity, inference techniques lead to functions to be optimized such as marginal likelihoods. In parallel, new developments are made in operations research (OR) where the goal is to focus specifically on the optimization task of well-established problems. This project starts with the idea that the use of recent OR algorithms to optimize CS objective functions will be a key ingredient to improve the performances of AI/ML/CS methods and their application to real world problems, such as traffic congestion prediction or frost event prediction in the agroindustry [5].

Objectives: This PhD project is in between ML and OR. 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 CS. Follows a detailed description of some ideas behind the PhD, that candidates are not expected to fully understand before the project.
In this PhD we aim first at addressing the problem of 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 to rely on modern discrete optimization techniques to optimize the type-II log-likelihood of the model. In [1-3], a simple relaxation along with an 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 - with no relaxation - in order to improve the selection of the variables even further. The starting point is the Bayesian linear regression model but extensions to the non-linear case with deep neural networks will also be investigated. 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. In the context of Bayesian logistic regression, where global and local variational techniques are used for variational inference, we also aim at addressing the selection of variables using related optimization techniques with cutting plane algorithms. Moreover, in an unsupervised setting, 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, operations research, machine learning and statistics. The candidate should also like programming.

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

Starting date : September 2022 (flexible)

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.