Statistical learning of latent variable models for complex data analysis

Faicel Chamroukhi (LMNO)

Résumé : Learning for complex data analysis is a central topic of modern statistics and is related in particular to the filed of statistical inference. In this talk, I will present latent data models and inference algorithms to learn from heterogeneous data, including temporal, functional, spatial, and high-dimensional multivariate data. The focus will be on the methodology and applications of mixture models and their extensions, and on maximum likelihood inference via expectation-maximization (EM) algorithms and Bayesian inference via Markov Chain Monte Carlo (MCMC) sampling techniques. I will first present hidden process regression models for non-stationary temporal data modeling and segmentation. The problem of modeling heterogeneity of data issued from possibly skewed and heavy-tailed distributions will be considered by proposing dedicated mixtures of experts. Then, I will consider the problem of statistical modeling when the basic unit of information is a curve, that is, the framework of functional data analysis, where I will present hierarchical dynamical mixtures for simultaneous curve clustering and segmentation. The problem of modeling and clustering of heterogeneous surfaces is also considered by using Bayesian spatial spline regression mixtures with mixed effects. From a massive data analysis prospective, I will consider the Bayesian non-parametric framework, and present Dirichlet process parsimonious mixtures for model-based cluster analysis of high-dimensional multivariate data. The presented models will be illustrated on real-world applications. Finally, I will briefly describe current work on co-clustering of high-dimensional functional data, and open perspectives on large scale collaborative mixtures for density estimation and clustering of massive data.