Statistique en grande dimension

- Pierre Bellec (Rutgers University)

  Title: Phase transitions and mean-field asymptotics in high-dimensional statistics

  Abstract: The talk will survey a number of phase transitions appearing in high-dimensional statistics with random design, when the number of samples is comparable to the ambient dimension. The talk will start with early results in this line of research, with the celebrated phase transition for exact recovery in sparse compressed sensing, followed by more recent results of a similar nature including perfect recovery in robust regression, separability and existence of minimizers. We will review several typical mean-field asymptotics results that describe the behavior of regularized estimators in the same proportional regime, and describe recent challenges and advances towards estimating the corresponding mean-field parameters for statistical inference. Time permitting, we will describe recent works in this area involving causal inference, subsampling/bagging and iterative algorithms.

- Jaouad Mourtada (ENSAE)

  Title: Estimation of discrete distributions in relative entropy, and the deviations of the missing mass

  Abstract: We consider the problem of estimating a distribution on a finite set (alphabet) using an i.i.d. sample. More precisely, we are interested in estimators achieving a small relative entropy (Kullback-Leibler divergence) with respect to the true distribution, with high probability over the sampling of the dataset. While the empirical distribution (or MLE) is arguably the most natural estimator, it turns out to be inadequate for this problem, since it may underestimate the frequency of certain classes, or even miss them altogether. A simple correction due to Laplace consists in smoothing the empirical distribution by adding 1 to the count of each class. The resulting estimator is known to be minimax-optimal in expectation, yet its tail behavior is not fully understood.

  In this talk, we will describe optimal high-probability guarantees of the Laplace estimator, as well as the best achievable high-probability guarantees (by any estimator) in a minimax sense. These rates are slightly larger than suggested by asymptotics, and exhibit a separation between confidence-independent and confidence-dependent estimators.

  Finally, we will present a modified estimator, which unlike the Laplace estimator adapts to the "effective support size" of the true distribution. If time permits, we will also discuss the question of bounding the "missing mass" from the sample, which plays an essential role in the analysis.

- Madalina Olteanu (Paris Dauphine)

  Title: Feature selection for unsupervised learning

Abstract: Assessing the underlying structure of a dataset is often done by training a clustering procedure on the features describing the data. In practice, while the data may be described by numerous features, only a minority of them may be actually informative with regard to the structure. Furthermore, redundant features may also bias the clustering, whether one speaks of redundancy in the informative or the uninformative features. This presentation will review several approaches for sparse clustering, and focus on a recent algorithm designed for mixed data (made of numerical and categorical features). We will also see how the ideas developed for sparse clustering may be transposed in a multivariate change-point detection framework. The performances and the interpretability of the methods will be illustrated on several real-life data sets.

- Guillaume Braun (RIKEN Japan)

  Title: Clustering de graphes bipartis en grande dimension : limites statistiques et algorithmes efficaces

  Abstract: Le clustering des graphes bipartis est une tâche fondamentale en analyse de réseaux. En grande dimension, lorsque le nombre de lignes et de colonnes de la matrice d'adjacence associée au graphe sont d'ordres de grandeur différents, les méthodes existantes adaptées de celles utilisées pour les graphes symétriques produisent souvent des résultats sous-optimaux. Compte tenu du nombre croissant d'applications impliquant des graphes bipartis, il est crucial de concevoir des algorithmes efficaces et optimaux adaptés à ce cadre.

  Dans la première partie de cette présentation, nous examinerons les limites statistiques du problème en utilisant le Modèle à Blocs Stochastiques Biparti (BiSBM), un modèle de graphe aléatoire populaire utilisé comme référence pour le clustering des graphes bipartites. Dans la deuxième partie, nous démontrerons l'optimalité d'une méthode spectrale simple capable d'estimer la partition latente des lignes du graphe biparti lorsque c'est statistiquement possible.