Séminaire de Probabilités et Statistique

Mercredi 25 avril à 14h00

Laboratoire Dieudonné

Salle de Conférences

Nicolas Verzelen

(INRA Montpellier)

Variable clustering : optimal bounds and a convex approach

The problem of variable clustering is that of grouping similar components of a p-dimensional vector $X = (X_1, \ldots, X_p)$, and estimating these groups from n independent copies of X. Although K-means is a natural strategy for this problem, I will explain why it cannot lead to perfect cluster recovery. Then, I will introduce a correction that can be viewed as a penalized convex relaxation of K-means. The clusters estimated by this method are shown to recover the partition G at a minimax optimal cluster separation rate. Along the way, I will discuss some connections with graph clustering problems.