Pénalité "Data-driven" en sélection de modèles

Matthieu Lerasle (Université de São Paulo)

La sélection de modèles a connu un essor considérable ces dernières années, pour l'étude de problèmes de statistique non paramétrique, comme l'estimation de la densité, de la régression, ou encore des problèmes de classification. Elle a fait émerger un grand nombre de questions; il est clair qu'une bonne procédure de sélection doit en pratique être complètement explicite, simple à implémenter, rapide à calculer et fonctionner quel que soit le nombre de données. En théorie, elle doit être définie et satisfaire de bonnes propriétés asymptotiques de manière aussi générale que possible. Ces deux enjeux, bien que fondamentaux, restent difficiles à concilier.
Il existe aujourd'hui un nombre formidable de procédures de sélections, répondant toutes plus ou moins à ces objectifs. Les méthodes "déterministes", comme celles définies par des pénalités proportionnelles à la dimension des modèles, les pénalités L^1 ou les procédures d'agrégation sont plus simples et plus rapides à calculer, elles permettent de couvrir de grandes collections de modèles. Elles sont toutefois généralement trop conservatives pour être aymptotiquement optimales ou s'adapter à des problèmes complexes (bruit hétéroscédastique, condition de marge, données mélangeantes...). À l'inverse, les procédures "data-driven" offrent l'avantage de bien apprendre la structure des données et d'obtenir, pour des problèmes de sélection relativement simples, des résultats remarquables comme des inégalités oracles asymptotiquement optimales. Elles sont toutefois plus lourdes à implémenter; cette gêne pouvant être acceptable pour certaines procédures (heuristique de la pente avec complexité déterministe, validation croisée V-fold) ou complètement rédibitoires (validation croisée "leave-p-out").
L'objectif de l'exposé est de montrer les avantages et les limites de certaines pénalités "data-driven" dans un cadre relativement général d'estimation de densité. Nous verrons comment calibrer ces différentes pénalités lorsque les observations sont indépendantes, ou mélangeantes. Pour ces différentes procédures, nous montrerons que l'estimateur sélectionné vérifie une inégalité oracle optimale. Nous discuterons aussi de l'heuristique de la pente que nous justifierons là aussi pour des processus non nécessairement indépendants.