Modèle de Cox et données en très grande dimension

Dr Philippe Bastien, L'Oréal

Résumé

L'objectif de cette présentation est de proposer des solutions dans le cadre de l'analyse de données censurées lorsque l'on dispose de très nombreuses variables explicatives potentiellement très corrélées avec souvent un nombre beaucoup plus faible d'individus.

Il existe deux principales alternatives en terme de régularisation qui sont d'un côté une approche par réduction de la dimensionnalité et de l'autre une approche par pénalisation L1 de la vraisemblance partielle.

La première approche a été proposée par Bastien et Tenenhaus (modèle PLS-Cox, 2001) dans le cadre de la régression PLS généralisée. La seconde approche a été proposée par Gui et Li (2005) avec la méthode "LARS-Lasso".

Cette dernière utilise la connexion entre les méthodes LAR et LASSO (Efron et al., 2004) pour rendre utilisable dans le cas de données de très grande dimension la méthode de sélection de variables "Lasso" développée et adaptée par Tibshirani (1995, 1997) dans le cadre du modèle de Cox.

Dans le cadre de la régression PLS on s'intéressera particulièrement aux méthodes à noyaux qui dans le cas linéaire peuvent être utilisées pour réduire la taille des matrices des descripteurs diminuant ainsi considérablement les temps de calculs et qui dans le cas plus général permettent d'accéder à des approches non linéaires en utilisant les outils simples et efficaces de l'algèbre linéaire.

Des applications sur des données d'expression de gènes illustreront les méthodes présentées.