SOUTENANCE DE THESE

Jean-Michel BECU

Unité de Recherche : UMR 7253 Laboratoire Heudiasyc

soutiendra sa thèse de Doctorat

sur le sujet :

Contrôle des fausses découvertes lors de la sélection de variables en grande dimension

Résumé

Dans le cadre de la régression, de nombreuses études s'intéressent au problème dit de la grande dimension, où le nombre de variables explicatives mesurées sur chaque échantillon est beaucoup plus grand que le nombre d'échantillons. Ces problèmes sont aujourd'hui fréquents dans le cadre des analyses de données en biologie. Si la sélection de variables est une question classique, les méthodes usuelles ne s'appliquent pas dans le cadre de la grande dimension, où ce problème devient très complexe. Il est alors très difficile d'identifier les variables explicatives associées à la variable expliquée.

Dans ce manuscrit, nous présentons la transposition de tests statistiques classiques à la grande dimension. Ces tests sont construits sur des estimateurs des coefficients de régression produits par des approches de régressions linéaires pénalisées, applicables dans le cadre de la grande dimension. L'objectif principal des tests que nous proposons consiste à contrôler le taux de fausses découvertes, c'est à dire le taux de variables sélectionnées à tort dans l'ensemble des variables supposées pertinentes. Cet objectif permet une interprétation intuitive des inférences réalisées, avec des ambitions réalistes pour les applications en biologie, où les sources d'aléa sont généralement nombreuses. La première contribution de ce manuscrit répond à un problème de quantification de l'incertitude sur les coefficients de régression réalisée sur la base de la régression ridge, qui pénalise les coefficients de régression par leur norme l2, dans le cadre de la grande dimension. Nous proposons un test statistique basé sur le rééchantillonage; ce test, bien qu'approximatif, a toujours permis de contrôler le taux d'erreur dans nos simulations balayant un large spectre de conditions expérimentales.

La seconde contribution de ce manuscrit porte sur une approche de sélection en deux étapes : une première étape de criblage des variables, basée sur la régression parcimonieuse Lasso précède l'étape de sélection proprement dite, où la pertinence des variables pré-sélectionnées est testée. En basant les tests sur l'estimateur de la régression ridge adaptive, dont la pénalité est construite à partir des coefficients de régression du Lasso, nous obtenons des gains

importants en terme de sensibilité, c'est à dire sur la capacité à retrouver les variables effectivement associées à la variable expliquée.

Une dernière contribution consiste à transposer cette approche en deux étapes à la sélection de groupes de variables, c'est à dire de groupes comportant au moins une variable pertinente. Ce problème de sélection, moins ambitieux, apparait être plus pertinent quand l'aléa est important. Le Sparse-Group-Lasso est alors utilisé pour la pré-sélection des groupes de variables. La modification des pénalités spécifiques de l'adaptive ridge ainsi que celle de la procédure de test retourne directement un résultat sur la pertinence des groupes.