UNIVERSITE CLAUDE BERNARD – LYON I

DIPLÔME NATIONAL DE DOCTORAT (Arrêté du 25 mai 2016)

Date de la soutenance : 13 décembre 2016

Nom de famille et prénom de l'auteur : **Ghislain DURIF**

Titre de la thèse : « Analyse multivariée de données de séquençage à haut débit. »

Résumé de la thèse

L'analyse statistique de données de séquençage à haut débit (NGS) pose des challenges computationnelles concernant la modélisation et l'inférence. Les technologies à haut débit permettent maintenant d'enregistrer l'expression de milliers de gènes tout en considérant un nombre croissant d'individus, par exemple des centaines de cellules individuelles. Malgré cette augmentation du nombre d'observations, les données de génomiques sont toujours caractérisées par leur grande dimension. Les orientations de recherche qui seront explorées dans ce manuscrit portent sur des méthodes de réductions de dimension hybrides qui sont basées sur des approches de compression (représentation des données dans un espace de faible dimension) et de sélection de variables. Des développements sont menés concernant: i) la régression Partial Least Squares parcimonieuse dans le contexte de la classification supervisée, et ii) les méthodes de factorisation parcimonieuse de matrices dans le contexte de l'exploration de données non supervisée. Dans chaque situation, notre principal objectif sera de se concentrer sur les problématiques de reconstruction et de visualisation des structures complexes organisant les données.

Dans cette optique, nous abordons des défis particuliers quant au développement de méthodes pour l'analyse de données en grande dimension. En effet, les questions de dimensionnalité interfèrent directement avec les procédures d'optimisation. Dans une première partie, nous développerons une approche de type PLS parcimonieuse, basée sur une pénalité adaptative, dans le contexte de la régression logistique, c-a-d pour prédire le label d'une réponse discrète. Cette approche sera par exemple utilisée pour des problèmes de prédiction (devenir de patients ou type cellulaire de cellules uniques) à partir de profils d'expression de gènes. La principale problématique dans ces circonstances est de prendre en compte la réponse pour écarter les variables non intéressantes. Nous mettrons en avant le lien direct qu'il existe entre la dérivation des algorithmes et la fiabilité des résultats.

Dans une seconde partie, motivés par des questions relatives à l'analyse de données "single-cell", nous considérerons des méthodes de factorisation parcimonieuse de matrices pour des données de comptages. Nous proposerons une approche à base de modèles statistiques qui est très flexible et qui prend en compte la sur-dispersion et l'amplification des zéros ("zero-inflation") lesquelles caractérisent les données "single-cell". Notre méthode de factorisation de matrices est fondée sur un modèle hiérarchique pour lequel nous dérivons une procédure d'estimation basée sur l'inférence variationnelle. Dans ce schéma, nous considérons une procédure de sélection de variables basée sur un modèle "spike-and-slab" approprié pour les données de comptage. L'intérêt d'une telle méthode pour la reconstruction, la visualisation et le clustering de données est illustré par des simulations et par la présentation de résultats préliminaires concernant une étude en cours sur des données "single-cell". Par ailleurs, toutes les méthodes proposées sont implémentées dans deux packages R: "plsgenomics" et "CMF".