DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **27 mars 2020**

Prénom et nom de famille de l'auteur : **Nadim BALLOUT**

Titre de la thèse : *« Approches penalisées pour les analyses en sous-groupes : Application en épidémiologie »*

Résumé

Dans un contexte où les cancers et l'insécurité routière font partie des principales causes de décès en France et dans le monde, chercher à en étudier les risques et aider à la prise en charge des malades et des victimes constituent un enjeu majeur de santé publique. Les études épidémiologiques mises en place pour répondre à ces besoins requièrent la disponibilité de nombreuses données qui deviennent de plus en plus détaillées et complexes. Certaines des méthodes statistiques utilisées classiquement ne satisfont pas pleinement aux exigences imposées par la taille et les caractéristiques de ces bases de données. L'objectif de ce travail de thèse est donc de développer des méthodes statistiques mieux adaptées, motivées par deux applications particulières : 1) la description des associations entre lésions chez les victimes d'accident de la route en fonction du type d'usager; et 2) l'étude du rôle de certains métabolites dans le développement du cancer du sein, en fonction du sous-type de cancer du sein. D'un point de vue méthodologique, les analyses stratifiées, ou en sous-groupes, constituent le coeur de nos recherches. En notant K le nombre de sous-groupes considérés, l'inférence statistique dans le contexte de ces analyses en sous-groupes revient en général à l'estimation de K vecteurs de paramètres, un vecteur par sous-groupe. Or, on s'attend généralement à une certaine homogénéité entre les K vrais vecteurs de paramètres. Nos méthodes reposent sur des pénalités de type fused lasso ou data shared lasso, et permettent de tirer profit de cette homogénéité pour réduire la complexité de la tâche d'apprentissage et améliorer la performance statistique de l'estimation. Par ailleurs, elles permettent l'identification des hétérogénéités éventuelles parmi les K vecteurs.
Dans le projet concernant la description des associations entre lésions chez les victimes d'accident de la route, nous nous sommes placés dans le cadre de l'estimation de modèles graphiques binaires stratifiés. Nous avons développé deux méthodes d'estimation basées chacune sur des régressions logistiques multiples en utilisant soit une pénalité de type fused lasso généralisé soit une pénalité de type data shared lasso. Dans le second projet, nous nous sommes placés dans le cadre général des études cas-témoins (appariées ou non), lorsque plusieurs sous-types de malades existent. Dans le cas des données appariées, nous avons étendu le data shared lasso au modèle de régression logistique conditionnelle, et avons montré la supériorité de l'approche par rapport à d'autres stratégies plus classiques. Dans le cas des données non-appariées, nous avons travaillé sous le modèle de régression logistique multinomial, pour lequel deux formulations pénalisées par la norme $L1$ ont été proposées dans la littérature. Nous montrons que l'une de ses formulations correspond en fait à la version data shared lasso de l'autre : nos résultats nous permettent ainsi de comparer formellement les deux formulations, et fournir des recommandations sur le choix de la formulation à utiliser en pratique. Globalement, nos résultats confirment que les méthodes tirant profit de l'homogénéité entre les K vecteurs, telles que celles reposant sur la pénalité data shared lasso, conduisent à des améliorations substantielles en termes d'efficacité d'estimation, lorsque cette homogénéité existe. Elles ciblent en effet une paramétrisation plus parcimonieuse lorsque des similarités existent entre les sous-groupes.

De plus, leur implémentation est relativement aisée, et en tout cas comparable à celle de méthodes plus classiques. Nous avons développé des codes permettant leur implémentation sous le logiciel R, qui sont accessible via la plateforme Github. Nous recommandons leur utilisation, en complément des approches plus classiques.