DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **13 octobre 2020**

Nom de famille et prénom de l'auteur : **Madame ETIEVANT Lola**

Titre de la thèse : *« Développements méthodologiques autour de l'inférence causale et de l'analyses de données en grande dimension »*

Résumé

L'identification des causes du cancer, mais aussi des mécanismes biologiques pouvant intervenir dans son développement, à partir de données observationnelles, est l'une des problématiques principales en épidémiologie du cancer. Les outils introduits récemment en inférence causale offrent un cadre formel pour répondre à de telles questions. En particulier, les variables contrefactuelles permettent de définir les effets causaux d'intérêts, et diverses conditions permettent de garantir qu'un effet causal donné soit estimable en pratique. Cependant, leur mise en application en épidemiologie du cancer présente un certain nombre d'enjeux ; l'objectif de cette thèse est d'en explorer quelques uns.

Tout d'abord, des réserves ont été émises concernant la pertinence des effets causaux estimés à partir de données observationnelles pour des expositions telles que l'obésité, pour laquelle il n'existe pas d'intervention « directe », mais seulement des interventions sur certaines de ses causes, comme l'activité physique ou l'alimentation. À cet effet, nous étudions comment l'effet d'une intervention hypothétique sur l'exposition d'intérêt est lié aux effets des interventions sur certaines de ses causes.

Ensuite, même si la plupart des modèles causaux d'intérêt en épidémiologie font intervenir des variables qui varient au cours du temps, ces dernières ne sont bien souvent observées qu'à un unique temps donné. De fait, il est assez usuel de travailler sous un modèle causal simplifié, qui néglige le caractère longitudinal de ces variables. Nous déterminons des conditions qui assurent que les quantités obtenues en travaillant sous de tels modèles soient liées à celles d'intérêt sous le vrai modèle longitudinal. Ces conditions, très restrictives, confirment ainsi que les quantités obtenues en travaillant sous des modèles causaux longitudinaux simplifiés doivent généralement être interprétées avec prudence.

Motivées par un projet sur les analyses en médiation en grande dimension, nous nous sommes intéressées à l'utilisation des modèles à variables latentes pour la réduction de dimension. Nous avons identifié un défaut dans plusieurs modèles proposés dans la littérature, notamment dans la

formulation probabiliste des moindres carrés partiels proposée par el Bouhaddani et al. (2018) Nous décrivons en détail le défaut sous leur modèle, et l'illustrons au moyen de simulations. Nos résultats suggèrent que les modèles à variables latentes doivent être développés avec précaution pour faire de la réduction de dimension, puisqu'ils peuvent en fait être trop simples lorsque les contraintes imposées sur les paramètres sont trop fortes.

Enfin, toujours motivées par le même projet, nous nous intéressons à la sélection du paramètre de régularisation dans les modèles de régression pénalisés. Plus précisément, nous considérons le lasso adaptatif, une extension du lasso qui utilise une version pondérée de la norme L_1 dans le terme de pénalité, où les poids sont obtenus à partir d'une estimation initiale du vecteur de paramètres. Nous montrons de manière empirique que la validation croisée « K-fold », bien que couramment employée, n'est pas adaptée à la calibration du paramètre de régularisation pour le lasso adaptatif. Une procédure alternative est proposée, et nous montrons sur des simulations qu'elle présente de meilleures performances que la validation croisée « K-fold ».

Référence

el Bouhaddani, S., Uh, H.-W., Hayward, C., Jongbloed, G., and Houwing, J. (2018). Probabilistic partial least squares model: identifiability, estimation and application. Journal of Multivariate Analysis, 167