

Développement de méthodes statistiques en causalité

Stage de recherche M2

Sujet

Le sujet du stage porte sur l'estimation causale en statistique avec des applications en épidémiologie. Deux directions sont possibles : la première s'intéresse à la causalité en présence d'une exposition multiple et la deuxième, à la causalité dans un cadre de données censurées à droite. Le choix du sujet s'effectuera selon les affinités du/de la candidat.e. Il est à noter que nous recherchons en priorité un candidat qui souhaite poursuivre en thèse, sur les thèmes abordés pendant le stage. Les sujets sont décrits en détail ci-dessous.

Causalité en présence d'une exposition multiple

Dans les cours introductifs de causalité, le premier type d'effet causal à être présenté est en général l'“effet causal moyen” (ATE) pour une exposition binaire A et une réponse continue Y . Dans ce cas, l'ATE est défini comme la différence entre les valeurs attendues de la réponse sous l'exposition et en absence de l'exposition : $ATE = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. Pour certaines applications, il est intéressant de considérer l'effet d'une exposition vectorielle A dans un espace euclidien de dimension d . Dans ce contexte, le paramètre cible $\mathbb{E}[Y(a_1, \dots, a_d)]$ définit une fonction “dose-réponse” de $\mathbb{R}^d$ vers $\mathbb{R}$. Pour identifier ce paramètre, on peut utiliser une approche basée sur les scores de propension multivariés. La difficulté consiste alors à estimer des fonctions de densités conditionnelles multivariées. Une méthode simple consiste à modéliser ces densités par régression linéaire multivariée [1], [2]. Nous souhaiterions explorer d'autres approches pour estimer ces densités multivariées conditionnelles à l'aide de méthodes non paramétriques ou paramétrique mais flexibles. Dans cette deuxième classe de méthodes, on trouve les approches basées sur des réseaux de neurones. Par exemple, Shi et al. [3] décrivent une architecture pour l'estimation du score de propension pour une exposition réelle. Il serait intéressant d'étudier la généralisation à une exposition multivariée. Une autre piste intéressante est celle des flots génératifs (voir [4]). Le stage pourrait porter sur l'étude d'une, ou plusieurs de ces approches, avec une validation par simulations. Nous aurons aussi la possibilité d'appliquer ces méthodes à des données réelles issues d'une cohorte française pour étudier les effets d'un mélange de polluants organiques persistants sur un événement de santé.

Causalité en analyse de survie

En analyse de survie causale, on s'intéresse à l'effet causal d'une exposition sur le temps avant un événement d'intérêt T en présence de données censurées. Le correspondant de l'ATE dans ce contexte est donné par la différence $\mathbb{P}(T(1) > t) - \mathbb{P}(T(0) > t)$, ou son intégrale par rapport à t sur un intervalle limité. Les résultats d'identification de l'ATE sur données non-censurées peuvent s'étendre à ce cadre, en considérant des poids qui prennent en compte non seulement la probabilité conditionnelle du traitement mais aussi la fonction de survie conditionnelle de la censure [5]. Cette approche donne un estimateur consistant si les modèles utilisés pour estimer les probabilités conditionnelles définissant les poids sont bien spécifiés. Des estimateurs doublement robustes qui sont consistants sous des hypothèses moins restrictives ont également été proposés [6]. Ces estimateurs, appelés estimateurs “d'apprentissage ciblé” sont issus de la théorie de l'efficacité semi-paramétrique (voir par exemple Schuler et van der Laan [7]). La théorie mathématique sous-jacente à ces estimateurs est particulièrement riche et élégante mais demande un peu d'investissement. Le stage pourrait porter sur un travail bibliographique pour comprendre et présenter de façon accessible ces approches et, si possible, sur une validation par simulations. Il sera aussi possible d'appliquer les méthodes étudiées à des données de vie réelle pour étudier l'effet d'un traitement hormonal sur le cancer du sein.

Encadrants

- Olivier Bouaziz, professeur des universités, Laboratoire Paul Painlevé (UMR CNRS 8524), Université de Lille, olivier.bou-aziz@univ-lille.fr
- Vittorio Perduca, maître de conférences, Laboratoire MAP5 (UMR CNRS 8145), Université de Paris, vittorio.perduca@u-paris.fr

References

- [1] Christian Fong, Chad Hazlett, and Kosuke Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.
- [2] Justin R Williams and Catherine M Crespi. Causal inference for multiple continuous exposures via the multivariate generalized propensity score. *arXiv preprint arXiv:2008.13767*, 2020.
- [3] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.
- [4] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.
- [5] Chao Cheng, Fan Li, Laine E Thomas, and Fan Li. Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. *American journal of epidemiology*, 191(6):1140–1151, 2022.
- [6] Helene CW Rytgaard, Frank Eriksson, and Mark J van der Laan. Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*, 79(4):3038–3049, 2023.
- [7] Alejandro Schuler. Introduction to modern causal inference. <https://alejandroschuler.github.io/mci/introduction-to-modern-causal-inference.html>.

Development of Statistical Methods in Causal Inference

M2 Research Internship

Topic

The internship focuses on causal estimation in statistics, with applications in epidemiology. Two directions are possible: the first concerns causal inference in the presence of multiple exposures, and the second deals with causal inference in the context of right-censored data. The specific topic will be chosen according to the candidate's interests. It should be noted that we are primarily seeking a candidate who intends to pursue a PhD on themes related to this internship. The topics are described in detail below.

Causality in the presence of multiple exposures

In introductory courses on causal inference, the first type of causal effect generally introduced is the *average treatment effect* (ATE) for a binary exposure A and a continuous outcome Y . In this case, the ATE is defined as the difference between the expected values of the outcome under exposure and under no exposure: $ATE = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. For certain applications, it is of interest to consider the effect of a vector-valued exposure A in a Euclidean space of dimension d . In this context, the target parameter $\mathbb{E}[Y(a_1, \dots, a_d)]$ defines a “dose-response” function from $\mathbb{R}^d$ to $\mathbb{R}$. To identify this parameter, one can use an approach based on multivariate propensity scores. The main difficulty then lies in estimating conditional multivariate density functions. A simple approach consists in modeling these densities through multivariate linear regression [1, 2]. We would like to explore alternative approaches for estimating these conditional multivariate densities using either nonparametric methods or flexible parametric models. Within this second class of methods, approaches based on neural networks are of particular interest. For instance, Shi et al. [3] describe an architecture for estimating the propensity score for a continuous exposure. It would be interesting to study the generalization of such methods to multivariate exposures. Another promising direction involves generative flows (see [4]). The internship could focus on studying one or several of these approaches, with validation through simulations. We will also have the opportunity to apply these methods to real data from a French cohort to investigate the effects of mixtures of persistent organic pollutants on a health outcome.

Causality in survival analysis

In causal survival analysis, one is interested in the causal effect of an exposure on the time-to-event T in the presence of censored data. The analogue of the ATE in this context is given by the difference $\mathbb{P}(T(1) > t) - \mathbb{P}(T(0) > t)$, or by its integral with respect to t over a bounded interval. Identification results for the ATE in uncensored settings can be extended to this framework by considering weights that account not only for the conditional probability of treatment but also for the conditional survival function of censoring [5]. This approach yields a consistent estimator when the models used to estimate the conditional probabilities defining the weights are correctly specified. Doubly robust estimators, which remain consistent under weaker assumptions, have also been proposed [6]. These estimators, known as “targeted learning estimators,” stem from the theory of semiparametric efficiency (see, for example, Schuler and van der Laan [7]). The mathematical theory underlying these estimators is particularly rich and elegant, though it requires some investment to master. The internship could involve a bibliographic study aimed at understanding and presenting these approaches in an accessible way, and, if possible, a simulation-based validation. It will also be possible to apply the studied methods to real-world data to investigate the effect of hormone therapy on breast cancer.

Supervisors

- Olivier Bouaziz, Professor, Laboratoire Paul Painlevé (UMR CNRS 8524), Université de Lille, olivier.bou-aziz@univ-lille.fr
- Vittorio Perduca, Associate Professor, Laboratoire MAP5 (UMR CNRS 8145), Université de Paris, vittorio.perduca@u-paris.fr

References

- [1] Christian Fong, Chad Hazlett, and Kosuke Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.
- [2] Justin R Williams and Catherine M Crespi. Causal inference for multiple continuous exposures via the multivariate generalized propensity score. *arXiv preprint arXiv:2008.13767*, 2020.
- [3] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.
- [4] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.
- [5] Chao Cheng, Fan Li, Laine E Thomas, and Fan Li. Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. *American journal of epidemiology*, 191(6):1140–1151, 2022.
- [6] Helene CW Rytgaard, Frank Eriksson, and Mark J van der Laan. Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*, 79(4):3038–3049, 2023.
- [7] Alejandro Schuler. Introduction to modern causal inference. <https://alejandroschuler.github.io/mci/introduction-to-modern-causal-inference.html>.