PhD position – INSERM U1136 CIPHOD team, Paris

Identifiability and optimal adjustability of total and direct effects using abstract graphs with applications to Epidemiology

Project summary: Accessing to causal graphs is pivotal for estimating causal effects [Greenland et al., 1999, Savitz and Wellenius, 2016]. These graphs represent qualitative cause-and-effect relations between exposures, health outcomes, and other variables. When there is no hidden confounding, causal graphs are directed acyclic graphs (DAGs) where each vertex is supposed to be statistically independent of its non-descendant given its parents—this is known as the causal Markov condition. Usually, in applications where the data depends on time, vertices in the causal graph correspond to different points in time of different time series, i.e., a causal relation between two vertices is a causal relation between two time series with a specific lag (the lag is equal or greater than zero). This said, in many applications, it is difficult for a practitioner to provide or validate this type of graph because it is difficult to determine the lag between a cause and an effect. It is of course easier for an expert to propose or validate an abstraction of the causal graph which represents a less detailed version (a macro view) and can even be cyclic. One example of abstraction is the summary causal graph [Assaad et al., 2022], where each vertex represents an entire time series and no temporal information is involved. Whether a causal graph or an abstraction is used, these tools can guide study design, statistical modeling, and the identification of potential biases or confounding factors, strengthening the validity of epidemiological investigations. In this project, we want to develop new causal inference tools especially based on abstractions and temporal data that can provide a deeper understanding of the mechanisms underlying health outcomes.

In this project, we will mainly be interested in two types of causal effects: the total effect which represents the overall impact of the one variable on another variable, considering all possible pathways, including both direct and indirect effects [Robins and Greenland, 1992, Pearl, 2001]; and the direct effect which represents the specific impact of one variable on another without considering the influence mediated through other variables in the system. The identifiability of causal effect (total or direct) from causal graphs received much attention in the last three decades [Robins, 1986, Greenland and Robins, 1986, Robins and Greenland, 1992, Greenland et al., 1999, Pearl, 1993, 1995, Spirtes et al., 2000, Didelez et al., 2006, Eichler and Didelez, 2007, Eichler and Vanessa, 2009, Perković et al., 2018] especially for non temporal data and for temporal data without instantaneous relations. In case of identifiability, many graphical tools (e.g., back-door, front-door, and single-door criterions and do-calculus) list different adjustment sets that allow for estimating causal effects. Furthermore, recently, there has been much progress to find among all adjustment sets, the optimal adjustment set [Witte et al., 2020, Henckel et al., 2022, Smucler et al., 2021, Runge, 2021] which is is characterized as the one with minimal asymptotic estimation variance. This adjustment set can help to choose the best model to link an exposure to a beneficial or harmful effect on health.

However, the identifiability as well as the optimal adjustability problem received less attention for abstractions [Beckers and Halpern, 2019, Beckers et al., 2020] and even less so for abstractions and temporal data with instantaneous relations (e.g., summary causal graph [Assaad et al., 2022]). Although, in epidemiology, opting for abstractions over causal graphs offers advantages, e.g., when investigating the impact of air pollution on respiratory diseases using temporal data, a causal graph that considers every temporal interaction among variables like pollutants, weather conditions, and individual behaviors can lead to excessive complexity. Abstractions reduce this complexity, which makes them easier to build, more computationally efficient, more interpretable, and accessible, which means they can be valuable for communicating findings to both policymakers and the general public. Additionally, many epidemiological studies require both instantaneous and delayed relations, e.g, in infectious disease research, it is essential to account for both direct person-to-person transmission (instantaneous) and the incubation period (delayed). Recently, we have solved the identifiability problem for total and direct effects [Assaad et al., 2024, Ferreira and Assaad, 2024] in summary causal graphs assuming there is no hidden confounding. However, finding all possible adjustment sets is still an open problem in summary causal graphs as well as accounting for hidden confounding. In addition, even if the problem of finding the optimal adjustment set for total effects has been addressed in recent years, addressing the problem of optimal adjustment set for direct effects is still neglected.

Therefore, the objectives of this project are the following:

- Extend the identifiability results for the total effect and the direct effect given in Ferreira and Assaad [2024], Assaad et al. [2024] to account for hidden confounding and optionally to different abstractions ;
- In case of identifiability, propose a graphical method that lists all adjustment sets ;
- Propose a graphical method to find the optimal adjustment set ;
- Implement these developed methods in epidemiological applications in collaboration with other teams within IPLESP.

Lab location and description: The Pierre Louis Institute of Epidemiology and Public Health (co-accredited by Inserm and Sorbonne University) is located at the Sorbonne University Faculty of Medicine - Hôpital Saint Antoine in Paris. It is composed of six teams, in addition to the recently established CIPHOD "Causal Inference in Public Health using large Observational health Databases" team. The general research objectives of CIPHOD are to put forth novel theoretical findings and develop innovative methodologies in the realm of causal inference, with a focus on their applicability and utility for epidemiologists.

Contract: The thesis contract will start in September 2024, for a duration of 36 months and after registration with the ED393 Doctoral School Pierre Louis of Public Health: Epidemiology and biomedical information sciences. The monthly doctoral allowance will be €2,131 gross, subject to annual revaluation.

The doctoral student will co-supervised by Dr. Charles Assaad (CIPHOD team) and Pr. Fabrice Carrat (CLEPIVIR team). The student will also collaborate with other teams in IPLESP and will have access to the resources and infrastructure available at INSERM/IPLESP and Sorbonne Université.

Candidat profile: Highly motivated candidate with an M2 degree and strong background in probability, machine learning, and causal inference, along with a keen interest in epidemiology. Proficiency in programming is also required.

Contact: Candidates are requested to send their CV (including a list of publications, research experiences, and references) along with a motivation letter to Charles Assaad (charles.assaad@inserm.fr) by June 13 2024. For additional details, please reach out to the same email address.

References

C. K. Assaad, E. Devijver, and E. Gaussier. Survey and evaluation of causal discovery methods for time series. *J. Artif. Int. Res.*, 73, apr 2022. doi: 10.1613/jair.1.13428.

C. K. Assaad, A. Meynaoui, E. Devijver, E. Gaussier, and G. Gossler. Identifiability in abstractions of time series causal graphs. submitted, 2024.

S. Beckers and J. Y. Halpern. Abstracting causal models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33012678.

S. Beckers, F. Eberhardt, and J. Y. Halpern. Approximate causal abstractions. In R. P. Adams and V. Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 606–615. PMLR, 22–25 Jul 2020.

V. Didelez, A. P. Dawid, and S. Geneletti. Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, page 138–146, Arlington, Virginia, USA, 2006. AUAI Press. ISBN 0974903922.

M. Eichler and V. Didelez. Causal reasoning in graphical time series models. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, page 109–116, Arlington, Virginia, USA, 2007. AUAI Press. ISBN 0974903930.

M. Eichler and D. Vanessa. On granger-causality and the effect of interventions in time series. *Lifetime Data Analysis*, 16, 02 2009. doi: 10.1007/s10985-009-9143-3.

S. Ferreira and C. K. Assaad. Identifiability of direct effects from summary causal graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, 2024.

S. Greenland and J. Robins. Identifiability, Exchangeability, and Epidemiological Confounding. *International Journal of Epidemiology*, 15(3):413–419, 09 1986. ISSN 0300-5771. doi: 10.1093/ije/15.3.413.

S. Greenland, J. Pearl, and J. Robins. Causal diagrams for epidemiologic research. *Epidemiology (Cambridge, Mass.)*, 10(1):37—48, January 1999. ISSN 1044-3983. doi: 10.1097/00001648-199901000-00005.

L. Henckel, E. Perković, and M. H. Maathuis. Graphical Criteria for Efficient Total Effect Estimation Via Adjustment in Causal Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2): 579–599, 03 2022. ISSN 1369-7412. doi: 10.1111/rssb.12451.

J. Pearl. [Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention. *Statistical Science*, 8(3):266 – 269, 1993. doi: 10.1214/ss/1177010894.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444.

J. Pearl. Direct and indirect effects. *Probabilistic and Causal Inference*, 2001.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18 (220):1–62, 2018.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986. ISSN 0270-0255. doi: https://doi.org/10.1016/0270-0255(86)90088-6.

J. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2): 143–155, 1992. ISSN 10443983.

J. Runge. Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15762–15773. Curran Associates, Inc., 2021.

D. A. Savitz and G. A. Wellenius. Causal Diagrams for Epidemiologic Inference. In *Interpreting Epidemiologic Evidence: Connecting Research to Applications*. Oxford University Press, 08 2016. ISBN 9780190243777. doi: 10.1093/acprof:oso/9780190243777.003.0003.

E. Smucler, F. Sapienza, and A. Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika*, 109(1):49–65, 03 2021. ISSN 1464-3510. doi: 10.1093/biomet/asab018.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

J. Witte, L. Henckel, M. H. Maathuis, and V. Didelez. On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020.