PhD position – INSERM U1136 CIPHOD team, Paris

Causal representation in a high dimensional setting with applications to epidemiology

Project summary: Access to causal graphs is essential for estimating causal effects [Greenland et al., 1999, Savitz and Wellenius, 2016]. These graphs represent qualitative cause-and-effect relationships between exposures, health outcomes, and other variables. When there is no hidden confounding factor, causal graphs are directed acyclic graphs (DAGs), where each node is independent of its nondescendants conditionally on its parents (causal Markov condition). However, in many applications, it is challenging for a practitioner to provide such a graph.

In certain cases, it is feasible to discover the causal graph from data under specific assumptions [Spirtes et al., 2000, Assaad et al., 2022]. However, the field of epidemiology, particularly with the increasing use of medico-administrative databases such as the EDS of AP-HP or the National Health Data System (SNDS), presents unique challenges. These include the high dimensionality of data, including temporal aspects; the reliance on variables that often serve only as proxies for the variables of real interest. These factors increase the risk of failure in accurately discovering causal graphs from data. In this context, causal representation emerges as a promising research approach [Schölkopf et al., 2021]. This approach aims to identify confounding factors among observed variables and to detect causal relationships between these hidden factors. However, the assumptions required for causal representation are generally more stringent than those for causal graph discovery from data [Yao et al., 2022, Sturma et al., 2023], raising questions about their applicability in the healthcare sector.

Additionally, distinguishing between proxy and non-proxy variables is not straightforward. Therefore, integrating a large language model into the causal representation framework could be beneficial. Such a model could retrieve information about proxies from existing epidemiological studies, potentially enhancing the accuracy and reliability of causal analysis in these complex settings.

Therefore, the objectives of this project are the following:

- Investigate the applicability of causal representation methods to discover causal relationships between hidden variables that cause observed proxy variables using medico-administrative databases;
- If existing methods are not applicable, propose a new causal representation algorithm suited for medico-administrative databases;
- Propose a pipeline that starts with using LLMs to find proxy variables from epidemiological studies and then apply causal representation algorithms on these variables.

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 be co-supervised by Dr. Charles Assaad (CIPHOD team) and Pr. Pierre-Yves Boëlle (SUMO 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) to Charles Assaad (charles.assaad@inserm.fr) by October 15 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.

A. Aït-Bachir, C. K. Assaad, C. de Bignicourt, E. Devijver, S. Ferreira, E. Gaussier, H. Mohanna, and L. Zan. Case Studies of Causal Discovery from IT Monitoring Time Series. In *The 39th European Conference on Uncertainty in Artificial Intelligence (UAI)*, Workshop on Causal Inference for Time Series Data, 2023.

S. Biswas, L. Corti, S. Buijsman, and J. Yang. Chime: Causal human-in-the-loop model explanations. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):27–39, Oct. 2022. doi: 10.1609/hcomp.v10i1.21985.

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.

B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. SchÃ¶lkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Z. Jin, J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*, 2023.

A. Meyer-Vitali and W. Mulder. Causing Intended Effects in Collaborative Decision-Making. In *CEUR Workshop Proceedings*, HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI, 2023.

O. Mian, D. Kaltenpoth, M. Kamp, and J. Vreeken. Nothing but regrets — privacy-preserving federated causal discovery. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8263–8278. PMLR, 25–27 Apr 2023.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

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.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, May 2021. ISSN 1558-2256. doi: 10.1109/JPROC.2021.3058954.

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

N. Sturma, C. Squires, M. Drton, and C. Uhler. Unpaired multi-domain causal representation learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34465–34492. Curran Associates, Inc., 2023.

W. Yao, G. Chen, and K. Zhang. Temporally disentangled representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26492–26503. Curran Associates, Inc., 2022.

M. Zečević, M. Willig, D. S. Dhami, and K. Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.