Question of Interest

Plenty of open challenges

Causality on discrete data

Problems:

- Less explored than continuous data
- Independence tests on contingency tables are sensitive to dimensionality, sparsity, etc.

[1](Peters, Janzing, Schölkopf. 2011) [2](Budhathoki. 2020)

Causality on discrete data

Problems:

- Less explored than continuous data
- Independence tests on contingency tables are sensitive to dimensionality, sparsity, etc.

Some solutions:

- Additive noise model in the ring domain[1]
- Shannon entropy-based score[2]

[1](Peters, Janzing, Schölkopf. 2011) [2](Budhathoki. 2020)

Causality on mixed-type data

Problems:

- Independence tests designed for either discrete or continuous data only
- Amplitude inequivalence (categorical domain size vs. continuous range)

Causality on mixed-type data

Problems:

- Independence tests designed for either discrete or continuous data only
- Amplitude inequivalence (categorical domain size vs. continuous range)

Some solutions:

- Independence test-based methods: converting data types into the same type, e.g., using discretization or one-hot encoding (information loss, variable structure destruction…)[1]
- Information-theoretic causal discovery: comparing the gain bit normalized with by the original description length[2]

[1](Zeng et al. 2022) [2](Marx and Vreeken. 2019)

Non-identical sets of variables

Problem:

- Related to missing data problem
- And often comes with heterogeneous data problem

A	B	D
…	…	…
…	…	…
…	…	…

A	E
…	…
…	…
…	…

A	B	C
…	…	…
…	…	…
…	…	…

[1](Kyono et al. 2021) [2](Danks. 2002) [3](Huang et al. 2020)

Non-identical sets of variables

A	B	D
…	…	…
…	…	…
…	…	…

Problem:

- Related to missing data problem
- And often comes with heterogeneous data problem

A	E
…	…
…	…
…	…

A	B	C
…	…	…
…	…	…
…	…	…

Some solutions:

- Data imputation[1]
- Learning partial graphs and combining them (many indeterminacies left)[2]
- Likelihood maximization while enforcing sparsity[3]

[1](Kyono et al. 2021) [2](Danks. 2002) [3](Huang et al. 2020)

Causal Representation Learning

Problems:

- The data we have is often not the data we want to infer causality on (e.g. satellite images)
- Instead, the (high-dimensional) data is created as an emission of a (low-dimensional) causal structure

Causal Representation Learning

Problems:

- The data we have is often not the data we want to infer causality on (e.g. satellite images)
- Instead, the (high-dimensional) data is created as an emission of a (low-dimensional) causal structure

Some solutions:

- Given enough interventional data, we can learn the underlying causal structure
- Invariance-based approaches can give us (partial) identifiability
- Multi-modal data can be incredibly potent to obtain more thorough identifiability of the underlying causal model

Cyclic Causal Graphs

Problems:

- In many cases, data (at the scale we can measure it) does not follow an acyclic causal model
- Instead, data which is best modeled via a cyclic causal graph (e.g. equilibrium data of a biological process)

Cyclic Causal Graphs

Problems:

- In many cases, data (at the scale we can measure it) does not follow an acyclic causal model
- Instead, data which is best modeled via a cyclic causal graph (e.g. equilibrium data of a biological process)

Some solutions:

- Explicitly try to model the cyclic causal structure (this is usually very hard)
- Leverage equivalences between latent variable models and cyclic causal models
- Fit explicit time-based models which entail the right correlation structure in the equilibrium state

Text and Causality

Problems:

- For many things, RCT data is impossible or radically incomplete
- Instead, since time immemorial, people have made decisions based on a Google search “how to X reddit”
- Now in the era of LLMs, can we leverage these to obtain (average or individual) treatment effects of various interventions, using only those self-reports in text form?

Text and Causality

Problems:

- For many things, RCT data is impossible or radically incomplete
- Instead, since time immemorial, people have made decisions based on a Google search “how to X reddit”
- Now in the era of LLMs, can we leverage these to obtain (average or individual) treatment effects of various interventions, using only those self-reports in text form?

Some solutions:

- How big are the resulting biases? How can we deal with those?
- When people start asking ChatGPT instead of reddit, how much better or worse is this?
- How do we even begin to measure this?

Conclusion

- Causal discovery from observational data is fundamental for decision making, trustworthy AI, etc.

- Most technics make a lot of assumtions (sufficiency, faithfulness…) that do no always hold in the real-world

- That is not inevitable, and those challenges can be opportunities

Tutorial website

Find slides and references on tutorial website: https://eda.rg.cispa.io/events/cdrw25sdm/

Thank you!

- Causal discovery from observational data is fundamental for decision making, trustworthy AI, etc.
- Most technics make a lot of assumtions (sufficiency, faithfulness…) that do no always hold in the real-world
- That is not inevitable, and those challenges can be opportunities

Tutorial website

Find slides and references on tutorial website: https://eda.rg.cispa.io/events/cdrw25sdm/

Bibliographic references (Intro and basics)

- Chickering, David Maxwell (2002). “Optimal structure identification with greedy search”. In: Journal of machine learning research 3.Nov, pp. 507–554.
- Hoyer, Patrik et al. (2008). “Nonlinear causal discovery with additive noise models”. In: Advances in neural information processing systems 21.
- Janzing, Dominik and Bernhard Schölkopf (2010). “Causal inference using the algorithmic Markov condition”. In: IEEE Transactions on Information Theory 56.10, pp. 5168–5194.
- Mian, Osman A, Alexander Marx, and Jilles Vreeken (2021). “Discovering fully oriented causal networks”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. 10, pp. 8975–8982.
- Pearl, Judea (2009). Causality. Cambridge university press.
- Shimizu, Shohei et al. (2006). “A linear non-Gaussian acyclic model for causal discovery.” In: Journal of Machine Learning Research 7.10.
- Spirtes, Peter, Clark N Glymour, and Richard Scheines (2000). Causation, prediction, and search. MIT press. Zheng, Xun et al. (2018). “Dags with no tears: Continuous optimization for structure learning”. In: Advances in neural information processing systems 31.

Bibliographic references (Confounding)

- Dong, Xinshuai et al. (2024). "A Versatile Causal Discovery Framework to Allow Causally-Related Hidden Variables". In: International Conference on Learning Representations. Poster. url: https://openreview.net/forum?id=FhQSGhBlqv.
- Kaltenpoth, David and Jilles Vreeken (July 2023a). "Causal Discovery with Hidden Confounders using the Algorithmic Markov Condition". In: Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence. Vol. 216. Proceedings of Machine Learning Research. PMLR, pp. 1016–1026. url: https://proceedings.mlr.press/v216/kaltenpoth23a.html.
- — (2023c). "Nonlinear Causal Discovery with Latent Confounders". In: Proceedings of the 40th International Conference on Machine Learning. Vol. 202. Proceedings of Machine Learning Research, pp. 15639–15654. url: https://proceedings.mlr.press/v202/kaltenpoth23a.html.
- Messerli, Franz H (2012). "Chocolate consumption, cognitive function, and Nobel laureates". In: N Engl J Med 367.16, pp. 1562–4.
- Prinz, Aloys Leo (2020). "Chocolate consumption and noble laureates". In: Social Sciences & Humanities Open 2.1, p. 100082.
- Xie, Feng et al. (2022). "Identification of Linear Non-Gaussian Latent Hierarchical Structure". In: Proceedings of the 39th International Conference on Machine Learning. Vol. 162. Proceedings of Machine Learning Research, pp. 24370–24387. url: https://proceedings.mlr.press/v162/xie22a.html.

Bibliographic references (Measurement error)

- Cook, James R. and Leonard A. Stefanski (1994). “Simulation–Extrapolation Estimation in Parametric Measurement Error Models”. In: Journal of the American Statistical Association 89.428, pp. 1314–1328. doi: 10.1080/01621459.1994.10476871.
- Dai, Haoyue, Peter Spirtes, and Kun Zhang (2022). “Independence Testing-Based Approach to Causal Discovery under Measurement Error and Linear Non-Gaussian Models”. In: Advances in Neural Information Processing Systems 35. url: https://arxiv.org/abs/2210.11021.
- Freedman, Laurence S. et al. (2008). “A Comparison of Regression Calibration, Moment Reconstruction and Imputation for Adjusting for Covariate Measurement Error in Regression”. In: Statistics in Medicine 27.25, pp. 5195– 5216. doi: 10.1002/sim.3361. url: https://pubmed.ncbi.nlm.nih.gov/18680172/.
- Stefanski, Leonard A. and Raymond J. Carroll (1987). “Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models”. In: Biometrika 74.4, pp. 703–716. url: https://academic.oup.com/biomet/article-pdf/74/4/703/786279/74-4-703.pdf.
- Van Huffel, Sabine and Joos Vandewalle (2007). “Overview of Total Least Squares Methods”. In: Signal Processing 87.10, pp. 2283–2302. doi: 10.1016/j.sigpro.2007.04.004.

Bibliographic references (Selection bias)

- Huang, Jiayuan et al. (2007). "Correcting Sample Selection Bias by Unlabeled Data". In: Advances in Neural Information Processing Systems 19, pp. 601– 608. url: https://papers.nips.cc/paper/3075-correcting-sample-selection-bias-by-unlabeled-data.
- Kaltenpoth, David and Jilles Vreeken (2023b). "Identifying Selection Bias from Observational Data". In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. 7, pp. 8177–8185. doi: 10.1609/aaai.v37i7.25987. url: https://doi.org/10.1609/aaai.v37i7.25987.
- Lin, Fengming, Xiaolei Fang, and Zheming Gao (2022). "Distributionally Robust Optimization: A Review on Theory and Applications". In: Numerical Algebra, Control and Optimization 12.1, pp. 19–54. doi: 10.3934/naco.2021057. url: https://doi.org/10.3934/naco.2021057.
- Manski, Charles F. (1990). "Nonparametric Bounds on Treatment Effects". In: American Economic Review 80.2, pp. 319–323. url: https://ideas.repec.org/a/aea/aecrev/v80y1990i2p319-23.html.
- Sugiyama, Masashi and Shinichi Nakajima (2007). "Covariate Shift Adaptation by Importance Weighted Cross Validation". In: Journal of Machine Learning Research 8, pp. 985–1005. url: https://www.jmlr.org/papers/v8/sugiyama07a.html.

Bibliographic references (Invariance)

- Arjovsky, Martin et al. (2020). “Invariant Risk Minimization”. In: arXiv preprint arXiv:1907.02893. url: https://arxiv.org/abs/1907.02893.
- Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen (2016). “Causal Inference Using Invariant Prediction”. In: Journal of the Royal Statistical Society: Series B 78.5, pp. 947–1012. doi: 10.1111/rssb.12167. url: https://arxiv.org/abs/1501.01332.
- Rojas-Carulla, Mauricio et al. (2018). “Invariant Models for Causal Transfer Learning”. In: Proceedings of the 34th International Conference on Machine Learning. url: https://arxiv.org/abs/1507.05333.
- Zhang, Kun et al. (2020). “Causal Discovery from Heterogeneous/Nonstationary Data”. In: Journal of Machine Learning Research 21.89, pp. 1–36. url: https://www.jmlr.org/papers/volume21/19-232/19-232.pdf.

Bibliographic references (Sparse mechanism shift)

- Mameche, Sarah, David Kaltenpoth, and Jilles Vreeken (2023). “Learning Causal Models under Independent Changes”. In: Advances in Neural Information Processing Systems 36. url: https://eda.rg.cispa.io/pubs/2023/linc-mameche,kaltenpoth,vreeken.pdf.
- Müller, Jens et al. (2020). “Learning Robust Models Using the Principle of Independent Causal Mechanisms”. In: arXiv preprint arXiv:2010.07167. url: https://arxiv.org/abs/2010.07167.
- Rothenhäusler, Dominik, Peter Bühlmann, Nicolai Meinshausen, and Jonas Peters (2021). “Anchor Regression: Heterogeneous Data Meet Causality”. In: Journal of the Royal Statistical Society: Series B 83.2, pp. 215–246. doi: 10.1111/rssb.12398. url: https://doi.org/10.1111/rssb.12398.
- Woodruff, Maverick, David Kaltenpoth, and Jilles Vreeken (2022). “Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis”. In: Advances in Neural Information Processing Systems 35. url: https://arxiv.org/abs/2206.02013.

Bibliographic references (Contextual confounders)

- Mameche, Sarah, David Kaltenpoth, and Jilles Vreeken (2024). “Identifying Confounding from Causal Mechanism Shifts”. In: Proceedings of the 41st International Conference on Machine Learning. Vol. 238. Proceedings of Machine Learning Research, pp. 12345–12356. url: https://proceedings.mlr.press/v238/mameche24a.html.
- Mooij, Joris M., Sara Magliacane, and Tom Claassen (2020). “Joint Causal Inference from Multiple Contexts”. In: Journal of Machine Learning Research 21.99, pp. 1–108. url: https://jmlr.org/papers/v21/17-123.html.
- Rothenhäusler, Dominik, Peter Bühlmann, and Nicolai Meinshausen (2019). “Causal Dantzig: Fast Inference in Linear Structural Equation Models with Hidden Variables under Additive Interventions”. In: Annals of Statistics 47.3, pp. 1718–1758. doi: 10.1214/18-AOS1732. url: https://arxiv.org/abs/1706.06159.

Bibliographic references (Time series)

- Huang, Biwei et al. (2020). “Causal discovery from heterogeneous/nonstationary data”. In: Journal of Machine Learning Research 21.89, pp. 1–53.
- Mameche, Sarah et al. (2025). “SPACETIME: Causal Discovery from Non-Stationary Time Series”. In: arXiv preprint arXiv:2501.10235.

Bibliographic references (Open challenges)

- Budhathoki, Kailash (2020). “Causal inference on discrete data”. PhD thesis. Danks, David (2002). “Learning the causal structure of overlapping variable sets”. In: International Conference on Discovery Science. Springer, pp. 178–191.
- Dhawan, Nikita et al. (2024). “End-to-end causal effect estimation from unstructured natural language data”. In: arXiv preprint arXiv:2407.07018. Kyono, Trent et al. (2021). “Miracle: Causally-aware imputation via learning missing data mechanisms”. In: Advances in Neural Information Processing Systems 34, pp. 23806–23817.
- Marx, Alexander and Jilles Vreeken (2019). “Causal inference on multivariate and mixed-type data”. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part II 18. Springer, pp. 655–671.
- Mooij, Joris and Tom Heskes (2013). “Cyclic causal discovery from continuous equilibrium data”. In: arXiv preprint arXiv:1309.6849.
- Schölkopf, Bernhard et al. (2021). “Toward causal representation learning”. In: Proceedings of the IEEE 109.5, pp. 612–634.
- Zeng, Yan et al. (2022). “Causal discovery for linear mixed data”. In: Conference on Causal Learning and Reasoning. PMLR, pp. 994–1009.