Journal of Machine Learning Research 25 (2024) 1-7 Submitted 7/23; Revised 11/23; Published 2/24

Causal-learn: Causal Discovery in Python

Yujia Zheng[1] YUJIAZH@CMU.EDU
Biwei Huang[2] BIH007@UCSD.EDU
Wei Chen[3] CHENWEIDELIGHT@GMAIL.COM
Joseph Ramsey[1] JDRAMSEY@ANDREW.CMU.EDU
Mingming Gong[4] MINGMING.GONG@UNIMELB.EDU.AU
Ruichu Cai[3] CAIRUICHU@GMAIL.COM
Shohei Shimizu[5,7] SHOHEI-SHIMIZU@BIWAKO.SHIGA-U.AC.JP
Peter Spirtes[1] PS7Z@ANDREW.CMU.EDU
Kun Zhang[1,6] KUNZ1@CMU.EDU
[1] *Carnegie Mellon University*
[2] *University of California, San Diego*
[3] *Guangdong University of Technology*
[4] *University of Melbourne*
[5] *Shiga University*
[6] *Mohamed bin Zayed University of Artificial Intelligence*
[7] *RIKEN*

Editor: Joaquin Vanschoren

Abstract

Causal discovery aims at revealing causal relations from observational data, which is a fundamental task in science and engineering. We describe *causal-learn*, an open-source Python library for causal discovery. This library focuses on bringing a comprehensive collection of causal discovery methods to both practitioners and researchers. It provides easy-to-use APIs for non-specialists, modular building blocks for developers, detailed documentation for learners, and comprehensive methods for all. Different from previous packages in R or Java, *causal-learn* is fully developed in Python, which could be more in tune with the recent preference shift in programming languages within related communities. The library is available at `https://github.com/py-why/causal-learn`.

Keywords: Causal Discovery, Python, Conditional Independence, Independence, Machine Learning

1. Introduction

A traditional way to uncover causal relationships is to resort to interventions or randomized experiments, which are often impractical due to their cost or logistical limitations. Hence, the importance of causal discovery, i.e., the process of revealing causal information through the analysis of purely observational data, has become increasingly apparent across diverse disciplines, including genomics, ecology, neuroscience, and epidemiology, among others (Glymour et al., 2019). For instance, in genomics, causal discovery has been instrumental in understanding the relationships between certain genes and diseases. Researchers might not have the resources to manipulate gene expressions, but they can analyze observational data, which are usually widely available, such as genomic databases, to uncover potential causal relationships. This can lead to breakthroughs in disease treatment and prevention strategies without the cost of traditional experimentation.

Current strategies for causal discovery can be broadly classified into constraint-based, score-based, functional causal models-based, and methods that recover latent variables. Constraint-based and score-based methods have been employed for causal discovery since the 1990s, using conditional independence relationships in data to uncover information about the underlying causal structure. Algorithms such as Peter-Clark (PC) (Spirtes et al., 2000) and Fast Causal Inference (FCI) (Spirtes et al., 1995) are popular, with PC assuming causal sufficiency and FCI handling latent confounders. In cases without latent confounders, score-based algorithms like the Greedy Equivalence Search (GES) (Chickering, 2002) aim to find the causal structure by optimizing a score function. These methods provide asymptotically correct results, accommodating various data distributions and functional relations but do not necessarily provide complete causal information as they usually output Markov equivalence classes of causal structures (graphs within the same Markov equivalence class have the same conditional independence relations among the variables).

On the other hand, algorithms based on Functional Causal Models (FCMs) have exhibited the ability to distinguish between different Directed Acyclic Graphs (DAGs) within the same equivalence class, thanks to additional assumptions on the data distribution beyond conditional independence relations. An FCM represents the effect variable as a function of the direct causes and a noise term; it renders causal direction identifiable due to the independence condition between the noise and cause: one can show that under appropriate assumptions on the functional model class and distributions of the involved variables, the estimated noise cannot be independent of the hypothetical cause in the reverse direction (Shimizu et al., 2006; Hoyer et al., 2008; Zhang and Hyvärinen, 2009). More recently, the Generalized Independent Noise condition (GIN) (Xie et al., 2020) has demonstrated its potential in learning hidden causal variables and their relations in the linear, non-Gaussian case. The identification of hidden variables as well as the causal structure among them is also a focus of causal representation learning (Schölkopf et al., 2021; Zheng and Zhang, 2024; Zhang et al., 2024).

To equip both practitioners and researchers with computational tools, several packages have been developed for or can be adapted for causal discovery. The Java library TETRAD (Glymour and Scheines, 1986; Scheines et al., 1998; Ramsey et al., 2018) contains a variety of well-tested causal discovery algorithms and has been continuously developed and maintained for over 40 years; R packages pcalg (Kalisch et al., 2012) and bnlearn (Scutari, 2010) also include some classical constraint-based and score-based methods such as PC and GES. However, these tools are based on Java or R, which may not align with the recent trend favoring Python in certain communities, particularly within machine learning. While there are Python wrappers available for these packages (e.g., py-tetrad (Andrew and Ramsey, 2023)/py-causal (Wongchokprasitti et al., 2019) for TETRAD, and Causal Discovery Toolbox (Kalainathan et al., 2020) for pcalg and bnlearn), they still rely on Java or R. This dependency can complicate deployment and does not cater directly to Python users seeking to develop their own methods based on an existing codebase. Thus, there is a pronounced need for a Python package that covers representative causal discovery algorithms across all primary categories.[1] Such a tool would significantly benefit a diverse range of users by providing access to both classical methods and the latest advancements in causal discovery.

In this paper, we describe *causal-learn*, an open-source python library for causal discovery. The library incorporates an extensive range of causal discovery algorithms, providing accessible APIs and thorough documentation to cater to a diversity of practical requirements and data assumptions. Moreover, it provides independent modules for specific functionalities, such as (conditional) independence tests, score functions, graph operations, and evaluation metrics, thereby facilitating custom needs and fostering the development of user-defined methods. An essential attribute of causal-learn is its full implementation in Python, eliminating dependencies on any other programming languages. As such, users are not required to have expertise in Java or R, enhancing the ease of integration within the enormous and growing Python ecosystem and promoting seamless utilization for a range of computational and scripting tasks. With causal-learn, modification and extensions based on the existing implementation of causal discovery methods also become plausible for developers and researchers who may not be familiar with Java or R. Additionally, it is convenient to apply *causal-learn* together with packages for inference (e.g., DoWhy (Sharma and Kiciman, 2020) and Ananke (Lee et al., 2023)) to conduct end-to-end causal pipelines.

2. Design

The design philosophy of causal-learn is centered around building an open-source, modular, easily extensible and embeddable Python platform for learning causality from data.

2.1 Methods

Causal-learn covers representative causal discovery methods across all major categories with official implementation of most algorithms, In addition, causal-learn also provides a variety of (conditional) independence tests and score functions as independent modules. All methods (version 0.1.3.8) are summarized in Table 1. Through the collective efforts of various teams and the contributions of the open-source community, *causal-learn* is always under active development to incorporate the most recent advancements in causal discovery.

2.2 Utilities

Causal-learn further offers a suite of utilities designed to streamline the assembly of causal analysis pipelines. The package features a comprehensive range of graph operations encompassing transformations among various graphical objects integral to causal discovery. These include Directed Acyclic Graphs (DAGs), Completed Partially Directed Acyclic Graphs (CPDAGs), Partially Directed Acyclic Graphs (PDAGs), and Partially Ancestral Graphs (PAGs). Additionally, metrics including precision and recall for arrow directions or adjacency matrices, along with the Structural Hamming Distance (Acid and de Campos, 2003), have also been included for ease of evaluation.

2.3 Demos, APIs, and benchmark datasets

The *causal-learn* package also contains extensive usage examples of all search methods, (conditional) independence tests, score functions, and utilities (`https://github.com/py-why/`

[1] LiNGAM (Ikeuchi et al., 2023) focuses on LiNGAM-based methods, of which many implementations are included in *causal-learn*; gCastle (Zhang et al., 2021) focuses on gradient-based DAG structure learning.

Table 1: Methods in *causal-learn* (version 0.1.3.8).

Categories	**Methods**
Constraint-based causal discovery	PC (Spirtes et al., 2000), MV-PC (Tu et al., 2019), FCI (Spirtes et al., 1995), CD-NOD (Huang et al., 2020)
Score-based causal discovery	GES (Chickering, 2002), A* (Yuan and Malone, 2013), Dynamic Programming (Silander and Myllymäki, 2006), GRaSP (Lam et al., 2022)
Function-based causal discovery	ANM (Hoyer et al., 2008), PNL (Zhang and Hyvärinen, 2009), LiNGAM (Shimizu et al., 2006), DirectLiNGAM (Shimizu et al., 2011), VAR-LiNGAM (Hyvärinen et al., 2010), RCD (Maeda and Shimizu, 2020), CAM-UV (Maeda and Shimizu, 2021)
Causal representation learning	GIN (Xie et al., 2020)
(Conditional) Independence tests	Fisher-z test (Fisher et al., 1921), Missing-value Fisher-z test, Chi-Square test, Kernel-based conditional independence (KCI) test and independence test (Zhang et al., 2011), G-Square test (Tsamardinos et al., 2006)
Score functions	BIC (Schwarz, 1978), BDeu (Buntine, 1991), Generalized Score (Huang et al., 2018)

`causal-learn/tree/main/tests`). For instance, causal discovery using PC is as simple as:

```
cg = pc(data) # apply PC with default parameters
```

Detailed documentation including all APIs and data structures is available at `https://causal-learn.readthedocs.io/en/latest`. It also includes a collection of well-tested benchmark datasets–since ground-truth causal relations are often unknown for real data, evaluation of causal discovery methods has been notoriously known to be hard, and we hope the availability of such benchmark datasets can help alleviate this issue and inspire the collection of more real-world datasets with (at least partially) known causal relations. Functions to import these datasets have also been included in the library.

3. Conclusion

The *causal-learn* library serves as a comprehensive toolset for causal discovery, significantly advancing the field of causal analysis and its applications in domains such as machine learning. It provides a robust platform for not only applying causal analysis techniques but also for facilitating the development of novel or enhanced algorithms. This is achieved by providing an infrastructure fully in Python that allows users to efficiently modify, extend, and tailor existing implementations, contribute new ones, and maintain high-quality standards. Given the current demand for causal learning and the rapid progress in this field, coupled with the active development and contribution from our team and the community, the *causal-learn* library is poised to bring causality into an indispensable component across diverse disciplines.

Acknowledgments

We are grateful for the collective efforts of all open-source contributors who continue to foster the growth of *causal-learn*. Especially, we would like to thank Yuequn Liu, Zhiyi Huang, Feng Xie, Haoyue Dai, Erdun Gao, Aoqi Zuo, Takashi Nicholas Maeda, Takashi Ikeuchi, Madelyn Glymour, Ruibo Tu, Wai-Yin Lam, Ignavier Ng, Bryan Andrews, Yewen Fan, and Xiangchen Song. The work of WC is supported in part by the Natural Science Foundation of China (62206064). The work of MG is supported in part by ARC DE210101624. The work of RC is supported in part by the National Key R&D Program of China (2021ZD0111501) and the National Science Fund for Excellent Young Scholars (62122022). The work of SS is supported by ONR N00014-20-1-2501 and JST CREST JPMJCR22D2. This project is partially supported by NSF Grant 2229881, the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Apple Inc., KDDI Research Inc., Quris AI, and Infinite Brain Technology.

References

Silvia Acid and Luis M de Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445–490, 2003.

Bryan Andrew and Joseph Ramsey. py-tetrad, 2023. URL `https://github.com/cmu-phil/py-tetrad`.

Wray Buntine. Theory refinement on bayesian networks. In *Uncertainty proceedings 1991*, pages 52–60. Elsevier, 1991.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Ronald Aylmer Fisher et al. 014: On the" probable error" of a coefficient of correlation deduced from a small sample. 1921.

Clark Glymour and Richard Scheines. Causal modeling with the tetrad program. *Synthese*, 68:37–63, 1986.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, Bernhard Schölkopf, et al. Nonlinear causal discovery with additive noise models. In *NIPS*, volume 21, pages 689–696. Citeseer, 2008.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1551–1560, 2018.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89):1–53, 2020.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Takashi Ikeuchi, Mayumi Ide, Yan Zeng, Takashi Nicholas Maeda, and Shohei Shimizu. Python package for causal discovery based on lingam. *vol*, 24:1–8, 2023.

Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. Causal discovery toolbox: Uncovering causal relationships in python. *The Journal of Machine Learning Research*, 21(1): 1406–1410, 2020.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012. doi: 10.18637/jss.v047.i11.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence*, pages 1052–1062. PMLR, 2022.

Jaron JR Lee, Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Ananke: A python package for causal inference using graphical models. *arXiv preprint arXiv:2301.11477*, 2023.

Takashi Nicholas Maeda and Shohei Shimizu. Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pages 735–745. PMLR, 2020.

Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Uncertainty in Artificial Intelligence*, pages 97–106. PMLR, 2021.

Joseph D Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A Barnes, and Clark Glymour. Tetrad—a toolbox for causal discovery. In *8th international workshop on climate informatics*, page 29, 2018.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03.

Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

Tomi Silander and Petri Myllymäki. A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 445–452, 2006.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506, 1995.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search.* MIT press, 2000.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78, 2006.

Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770. PMLR, 2019.

Chirayu (Kong) Wongchokprasitti, Harry Hochheiser, Jeremy Espino, Eamonn Maguire, Bryan Andrews, Michael Davis, and Chris Inskip. bd2kccd/py-causal v1.2.1, December 2019. URL `https://doi.org/10.5281/zenodo.3592985`.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *NeurIPS*, 2020.

Changhe Yuan and Brandon Malone. Learning optimal bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48:23–65, 2013.

K Zhang and A Hyvärinen. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655. AUAI Press, 2009.

Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. gcastle: A python toolbox for causal discovery. *arXiv preprint arXiv:2111.15155*, 2021.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ICA beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36, 2024.