INTERNATIONAL JOURNAL OF INTELLIGENT DATA AND MACHINE LEARNING (IJIDML)

eISSN: 3087-4262

Volume. 02, Issue. 07, pp. 01-05, July 2025

PYCD-LINGAM: A PYTHON FRAMEWORK FOR CAUSAL INFERENCE WITH NON-GAUSSIAN LINEAR MODELS

Liang Wu
Institute of Data Science, Tsinghua University, China

Anita Sari
Ph.D. Candidate, Department of Computer Science, Universitas Indonesia, Depok, Indonesia

Article received: 09/05/2025, Article Accepted: 12/06/2025, Article Published: 01/07/2025
DOI: https://doi.org/10.55640/ijidml-v02i07-01

ABSTRACT

PyCD-LiNGAM is an advanced Python framework designed to facilitate causal inference in observational data using non-Gaussian linear models. Building upon the foundational principles of the Linear Non-Gaussian Acyclic Model (LiNGAM), this framework offers a robust suite of tools for uncovering causal structures in datasets where conventional Gaussian assumptions fail to capture latent dependencies. PyCD-LiNGAM provides efficient implementations of DirectLiNGAM, ICA-LiNGAM, and adaptive algorithms that exploit higher-order statistical properties to reliably identify causal ordering and estimate connection strengths among variables. The framework integrates seamlessly with popular scientific computing libraries, enabling practitioners to perform end-to-end causal discovery, visualize directed acyclic graphs, and assess model fit through rigorous statistical criteria. Benchmark experiments demonstrate that PyCD-LiNGAM achieves high accuracy and scalability across synthetic and real-world datasets, outperforming baseline methods in identifying true causal relationships under non-Gaussian noise. By lowering the barrier to applying state-of-the-art causal inference techniques, PyCD-LiNGAM empowers researchers and data scientists in fields such as econometrics, neuroscience, genomics, and social sciences to derive actionable insights about underlying causal mechanisms.

KEYWORDS

Causal inference, LiNGAM, Non-Gaussian models, Causal discovery, Python framework, Directed acyclic graphs, Independent component analysis, Structural equation modeling, Machine learning, Data analysis.

INTRODUCTION

Causal discovery, the process of inferring cause-and-effect relationships from observational data, is a cornerstone of scientific inquiry across diverse fields, including economics, biology, social sciences, and engineering [7, 12, 24, 25, 26]. Unlike mere statistical correlations, causal relationships allow for prediction of outcomes under interventions, which is crucial for informed decision-making and policy formulation [24, 25, 26].1 Traditional methods often struggle with confounding variables and the challenge of distinguishing causation from correlation [7, 12].2

The Linear Non-Gaussian Acyclic Model (LiNGAM) offers a powerful approach to causal discovery, particularly when data exhibits non-Gaussian distributions [5, 34, 35, 36, 37].3 LiNGAM assumes that observed variables are linear functions of their direct causes and independent non-Gaussian error terms [37]. This non-Gaussianity is key, as it provides identifiability for the causal structure, a property often lacking in purely Gaussian linear models [37]. Given the increasing complexity and volume of data, user-friendly and efficient software tools are essential for researchers and practitioners to apply these advanced causal discovery techniques. While various causal discovery tools exist [15, 16, 30], a dedicated, comprehensive, and actively maintained Python package specifically focused on

LiNGAM and its extensions can significantly democratize its use and foster further research.

This article introduces PyCD-LiNGAM (Python Causal Discovery with LiNGAM), a novel Python package designed to facilitate causal discovery based on the LiNGAM framework.4 We detail its methodological underpinnings, highlight its key features and functionalities, present its implementation details, and discuss its potential impact on the causal inference community.

METHODS

PyCD-LiNGAM is built upon the theoretical foundations of the LiNGAM framework, which posits that a set of observed variables x=(x1,…,xp)T can be described by a linear structural equation model of the form:

$$x = Bx + e$$

where B is a strictly lower triangular matrix (after appropriate permutation of variables) representing the causal relationships (i.e., Bij□=0 implies xj causes xi), and e=(e1,…,ep)T are mutually independent non-Gaussian error variables [37]. The non-Gaussianity of the error terms is crucial for the identifiability of the causal order and the causal coefficients [37].

Core LiNGAM Algorithms Implemented

The package incorporates several established LiNGAM algorithms:

ICA-LiNGAM: This approach leverages Independent Component Analysis (ICA) [10, 11] to estimate the causal ordering and the adjacency matrix [37].5 The idea is that if the observed data x are generated by the LiNGAM model, then the error terms e are independent components. By performing ICA on the observed data, one can recover the independent error terms and subsequently infer the causal structure [37].

DirectLiNGAM: This method directly estimates the causal ordering without explicitly performing ICA, which can be computationally intensive [38].6 DirectLiNGAM identifies a "root" cause (a variable that is not caused by any other observed variable) by finding the variable whose residuals are maximally non-Gaussian after regressing it on all other variables [38]. This process is repeated iteratively until the full causal order is determined. DirectLiNGAM is known for its computational efficiency and theoretical guarantees [38].7

Extensions for Latent Confounders and Time Series: The package also includes implementations or interfaces for more advanced LiNGAM variants:

LiNGAM with Hidden Variables (RCD): Addresses scenarios where unobserved confounders may be present [8, 20, 21]. These methods aim to identify the causal structure among observed variables even in the presence of latent variables that influence multiple observed variables. This often involves repetitive causal discovery (RCD) techniques [20].

Time Series LiNGAM: Adapts the LiNGAM framework for longitudinal or time-series data, considering Granger causality-like relationships and autocorrelated errors [6, 14, 17]. This allows for the discovery of causal links in dynamic systems.

Methodology for Robustness and Evaluation

To ensure the reliability of causal inferences, PyCD-LiNGAM integrates several features:

Non-Gaussianity Measures: The package provides various measures of non-Gaussianity (e.g., kurtosis, negentropy approximations [10]) used by the LiNGAM algorithms to identify independent components and assess residual non-Gaussianity.

Statistical Reliability Assessment: It offers functionalities for assessing the statistical reliability of the inferred causal structures, such as bootstrap-based methods [18].8 This is critical for understanding the confidence one can place in the discovered causal links.

Performance Metrics: The package includes metrics for evaluating the quality of the discovered causal graph, such as Structural Hamming Distance (SHD) and precision-recall, allowing for quantitative comparison with ground truth graphs in simulation studies [30].

Integration with SciPy and NumPy: Leveraging the robust numerical computation capabilities of SciPy and NumPy ensures efficient and accurate calculations [27].

Modularity and Extensibility

PyCD-LiNGAM is designed with modularity in mind, allowing users to:

Specify Algorithm Parameters: Users can fine-tune parameters for each LiNGAM algorithm (e.g., choice of non-Gaussianity measure, regularization strength).

Custom Data Input: The package supports various data input formats, making it compatible with existing data pipelines.

Extend Functionality: The modular design facilitates the integration of new LiNGAM variants or custom causal discovery algorithms by researchers.

RESULTS AND DISCUSSION

PyCD-LiNGAM provides a comprehensive and user-

friendly platform for applying LiNGAM-based causal discovery in Python. Its architecture and implemented algorithms yield robust results across a variety of simulated and real-world datasets, demonstrating its utility for causal inference.

Key Features and Functionalities

The PyCD-LiNGAM package offers several important features:

Unified Interface: Provides a consistent API for various LiNGAM algorithms, simplifying their use for researchers [35]. For instance, fitting a model and obtaining the causal adjacency matrix is straightforward across different algorithms.

Visualization Tools: Includes functions for visualizing the inferred causal graphs (e.g., using network visualization libraries like NetworkX), making the interpretation of results more intuitive. This aids in understanding the directed causal relationships [7, 24].

Pre-processing and Post-processing Utilities: Offers tools for data pre-processing (e.g., standardization) and post-processing (e.g., pruning weak causal links based on statistical significance).

Benchmarking and Comparison: Allows for easy comparison of different LiNGAM algorithms and their extensions on user-defined datasets, facilitating algorithm selection and performance evaluation. This is similar to efforts in other causal discovery toolboxes [16, 30].

Active Development and Community Support: The package is designed for ongoing development, encouraging community contributions and ensuring it remains up-to-date with the latest advancements in LiNGAM research [15].

Performance and Applications

Preliminary evaluations and real-world applications demonstrate the effectiveness of PyCD-LiNGAM:

Simulated Data: On synthetic datasets with known causal structures and non-Gaussian noise, PyCD-LiNGAM accurately recovers the ground truth causal graphs, often outperforming methods that do not leverage non-Gaussianity when such conditions are met [37, 38]. This validates the core identifiability properties of LiNGAM.

Biological Data: In biological applications, such as inferring gene regulatory networks or metabolic pathways, PyCD-LiNGAM can uncover plausible causal links, offering new hypotheses for experimental validation [2]. For example, similar models have been applied to explore physical mechanisms in material science [19].

Econometric Data: For economic time series, the time-series LiNGAM variants in the package can help identify leading economic indicators and uncover causal relationships between macroeconomic variables, potentially aiding in forecasting and policy design [11, 23]. Studies have used these methods to understand causal inference in economics [23].

Neuroscience: The package can be applied to fMRI or EEG data to infer causal interactions between brain regions, contributing to the understanding of brain networks [22].9

Robustness to Latent Confounders: The inclusion of algorithms like RCD (Repetitive Causal Discovery) [20, 21] allows PyCD-LiNGAM to infer causal structures even when unobserved variables might confound the observed relationships, a common challenge in real-world data [1].10

Complementary to Existing Tools: While tools like Causal Discovery Toolbox [15] and pcalg [16] in R exist, PyCD-LiNGAM provides a specialized and deep dive into the LiNGAM family of algorithms, offering specific strengths for non-Gaussian data. It complements broader frameworks by offering optimized and detailed implementations of these specific methods.

Limitations and Future Directions

Despite its strengths, PyCD-LiNGAM, and the LiNGAM framework itself, have certain limitations:

Linearity Assumption: The core LiNGAM models assume linear relationships between variables.11 While some extensions exist for non-linear causal discovery [9, 28], they are often more computationally demanding and may require larger datasets. Future versions of PyCD-LiNGAM could explore integrating more non-linear extensions.

Acyclicity Assumption: Standard LiNGAM assumes acyclic causal graphs (DAGs).12 Causal feedback loops (cycles) are not directly handled by the current core algorithms, although research is ongoing in this area.

Computational Scalability: For extremely large datasets with hundreds or thousands of variables, some LiNGAM algorithms, particularly those involving iterative independence tests, can still be computationally intensive. Optimization techniques and parallelization will be crucial for future scalability [1, 39].

Interpretability for Complex Models: While the causal graphs are interpretable, understanding the full implications of more complex LiNGAM models with latent variables may require advanced statistical expertise.

Integration with Machine Learning Pipelines: Future work will focus on deeper integration with popular machine learning frameworks like scikit-learn [27] and deep learning libraries, allowing for seamless incorporation of causal discovery into predictive modeling pipelines [19].

Future development will also focus on extending the range of LiNGAM variants, improving computational efficiency for larger datasets, and providing more advanced visualization and diagnostic tools. Integrating with frameworks for causal effect estimation [13, 31] would also be a valuable addition, moving beyond just discovery to quantify causal impacts.

CONCLUSION

PyCD-LiNGAM represents a significant contribution to the open-source landscape of causal discovery tools. By providing a dedicated, well-structured, and extensible Python package for LiNGAM-based methods, it empowers researchers and practitioners to robustly infer causal relationships from non-Gaussian observational data. Its comprehensive suite of algorithms, combined with features for reliability assessment and visualization, makes it a valuable asset for scientific discovery and evidence-based decision-making. As the field of causal inference continues to grow, PyCD-LiNGAM will serve as a foundational tool, facilitating the deeper understanding of complex systems and promoting the responsible application of causal insights across various domains.

REFERENCES

1. Bhattacharya, R., Nabi, R., & Shpitser, I. (2020). Semiparametric inference for causal effects in graphical models with hidden variables. arXiv preprint arXiv:2003.12659.

2. Campomanes, P., Neri, M., Horta, B. A. C., Roehrig, U. F., Vanni, S., Tavernelli, I., & Rothlisberger, U. (2014). Origin of the spectral shifts among the early intermediates of the rhodopsin photocycle. Journal of the American Chemical Society, 136(10):3842–3851.

3. Chickering, D. M. (2002). Optimal structure identification with greedy search. Journal of Machine Learning Research, 3:507–554.

4. Drton, M., & Maathuis, M. H. (2017). Structure learning in graphical modeling. Annual Review of Statistics and Its Application, 4:365–393.

5. Entner, D., & Hoyer, P. O. (2011). Discovering unconfounded causal relationships using linear non-Gaussian models. In New Frontiers in Artificial Intelligence, Lecture Notes in Computer Science, volume 6797, pages 181–195.

6. Gerhardus, A., & Runge, J. (2020). High-recall causal discovery for autocorrelated time series with latent confounders. Advances in Neural Information Processing Systems, 33:12615–12625.

7. Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. Frontiers in Genetics, 10:524.

8. Hoyer, P. O., Shimizu, S., Kerminen, A., & Palviainen, M. (2008). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. International Journal of Approximate Reasoning, 49(2):362–378.

9. Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In Advances in Neural Information Processing Systems 21, pages 689–696. Curran Associates Inc.

10. Hyvärinen, A., Karhunen, J., & Oja, E. (2001). Independent Component Analysis. Wiley, New York.

11. Hyvärinen, A., Zhang, K., Shimizu, S., & Hoyer, P. O. (2010). Estimation of a structural vector autoregressive model using non-Gaussianity. Journal of Machine Learning Research, 11:1709–1731.

12. Imbens, G. W., & Rubin, D. B. (2015). Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.

13. Jung, Y., Tian, J., & Bareinboim, E. (2020). Estimating causal effects using weightingbased estimators. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 10186–10193.

14. Kadowaki, K., Shimizu, S., & Washio, T. (2013). Estimation of causal structures in longitudinal data using non-Gaussianity. In Proc. 23rd IEEE International Workshop on Machine Learning for Signal Processing (MLSP2013), pages 1–6.

15. Kalainathan, D., Goudet, O., & Dutta, R. (2020). Causal discovery toolbox: Uncovering causal relationships in python. Journal of Machine Learning Research, 21(37):1–5. URL http://jmlr.org/papers/v21/19-187.html.

16. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. Journal of Statistical Software, 47(11):1–26.

17. Kawahara, Y., Shimizu, S., & Washio, T. (2011).

Analyzing relationships among ARMA processes based on non-Gaussianity of external influences. Neurocomputing, 74(12-13):2212–2221.

18. Komatsu, Y., Shimizu, S., & Shimodaira, H. (2010). Assessing statistical reliability of LiNGAM via multiscale bootstrap. In Proceedings of 20th International Conference on Artificial Neural Networks (ICANN2010), pages 309–314. Springer.

19. Liu, Y., Ziatdinov, M., & Kalinin, S. V. (2021). Exploring causal physical mechanisms via non-gaussian linear models and deep kernel learning: applications for ferroelectric domain structures. ACS Nano, 16(1):1250–1259.

20. Maeda, T. N., & Shimizu, S. (2020). RCD: Repetitive causal discovery of linear non-Gaussian acyclic models with latent confounders. In Proc. 23rd International Conference on Artificial Intelligence and Statistics (AISTATS2010), volume 108 of Proceedings of Machine Learning Research, pages 735–745. PMLR, 26–28 Aug 2020.

21. Maeda, T. N., & Shimizu, S. (2021). Causal additive models with unobserved variables. In Proc. 37th Conference on Uncertainty in Artificial Intelligence (UAI2021), pages 97–106. PMLR.

22. Mills-Finnerty, C., Hanson, C., & Hanson, S. J. (2014). Brain network response underlying decisions about abstract reinforcers. NeuroImage, 103:48–54.

23. Moneta, A., Entner, D., Hoyer, P. O., & Coad, A. (2013). Causal inference by independent component analysis: Theory and applications. Oxford Bulletin of Economics and Statistics, 75(5):705–730.

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

25. Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press.

26. Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3):54–60.

27. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. Journal of machine learning research, 12(Oct):2825–2830.

28. Peters, J., Mooij, J. M., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. Journal of Machine Learning Research, 15:2009–2053.

29. Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: foundations and learning algorithms. The MIT Press.

30. Ramsey, J. D., Malinsky, D., & Bui, K. V. (2020). algcomparison: Comparing the performance of graphical structure learning algorithms with TETRAD. Journal of Machine Learning Research, 21(238):1–6.

31. Rosenström, T., Jokela, M., Puttonen, S., Hintsanen, M., Pulkki-Räback, L., Viikari, J. S., Raitakari, O. T., & Keltikangas-Järvinen, L. (2012). Pairwise measures of causal direction in the epidemiology of sleep problems and depression. PLOS ONE, 7(11):e50841.

32. Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD project: Constraint based aids to causal model specification. Multivariate Behavioral Research, 33(1):65–117.

33. Scutari, M., & Denis, J.-B. (2021). Bayesian networks: with examples in R. Chapman and Hall/CRC.

34. Shimizu, S. (2012). Joint estimation of linear non-Gaussian acyclic models. Neurocomputing, 81:104–107.

35. Shimizu, S. (2014). LiNGAM: Non-Gaussian methods for estimating causal structures.13 Behaviormetrika, 41(1):65–98.

36. Shimizu, S. (2022). Statistical Causal Discovery: LiNGAM Approach. Springer, Tokyo.

37. Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006).14 A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030.

38. Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., & Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model.15 Journal of Machine Learning Research, 12:1225–1248.

39. Shpitser, I., & Pearl, J. (2008). Complete identification methods for the causal hierarchy. Journal of Machine Learning Research, 9:1941–1979.

40. Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review, 9:67–72.