Journal of Machine Learning Research 24 (2023) 1-8 Submitted 3/21; Revised 9/22; Published 1/23

Python package for causal discovery based on LiNGAM

Takashi Ikeuchi IKEUCHI@SCREEN.CO.JP
Mayumi Ide IDE@SCREEN.CO.JP
SCREEN Advanced System Solutions Co., Ltd., Japan
Yan Zeng YANAZENG013@GMAIL.COM
Department of Computer Science and Technology, Tsinghua University, China
Takashi Nicholas Maeda TN.MAEDA@MAIL.DENDAI.AC.JP
School of System Design and Technology, Tokyo Denki University, Japan
Center for Advanced Intelligence Project, RIKEN, Japan
Shohei Shimizu SHOHEI-SHIMIZU@BIWAKO.SHIGA-U.AC.JP
Faculty of Data Science, Shiga University, Japan
Center for Advanced Intelligence Project, RIKEN, Japan

Editor: Andreas Mueller

Abstract

Causal discovery is a methodology for learning causal graphs from data, and LiNGAM is a well-known model for causal discovery. This paper describes an open-source Python package for causal discovery based on LiNGAM. The package implements various LiNGAM methods under different settings like time series cases, multiple-group cases, mixed data cases, and hidden common cause cases, in addition to evaluation of statistical reliability and model assumptions. The source code is freely available under the MIT license at `https://github.com/cdt15/lingam`.

Keywords: Causal structure learning, statistical reliability, model evaluation

1. Introduction

Statistical causal inference learns causal quantities from data (Imbens and Rubin, 2015; Pearl, 2000). A common procedure for this is as follows: Users first specify the causal quantity to be estimated, e.g., the intervention effect of a variable on another variable. Second, they draw the causal graph based on background knowledge. Then, they derive variables (if any) that should be used to identify the quantity of interest based on graphical criteria, such as back-door and front-door criterion, and their generalizations (Pearl, 1995; Shpitser and Pearl, 2008; Bhattacharya et al., 2020; Jung et al., 2020).

A fundamental step of the aforementioned procedure is to draw the causal graph based on background knowledge. However, it is often the case that background knowledge is not enough to draw the causal graph. Causal discovery (Spirtes et al., 1993; Pearl, 2019) is a methodology for inferring causal graphs in data-driven ways; it aims to help users draw causal graphs by combining data with prior knowledge.

A classic approach for causal discovery is to use conditional independence of variables for inferring the underlying causal graph (Spirtes et al., 1993; Pearl, 2000). This approach, in principle, does not make specific assumptions on the functional forms of the causal relations of variables or distributions of variables; it only infers a set of equivalent models and is not able to estimate causal directions for most cases.

In contrast, a recent approach (Shimizu, 2014; Zhang and Hyvärinen, 2016; Shimizu, 2022) makes some assumptions on the functional forms or/and distributions of variables to address this limitation. The linear non-Gaussian acyclic model (Shimizu et al., 2006), abbreviated as LiNGAM, is the most well-known example, where the error variables assumedly follow non-Gaussian continuous distributions, but at most one error variable may be Gaussian. The assumption of non-Gaussian errors enables examining the independence of error variables, unlike that of Gaussian errors. This LiNGAM approach achieves better identification results and is capable of uniquely estimating causal directions in much more cases than the classic approach based on conditional independence. This feature of identifiability has attracted much attention of the research community (Drton and Maathuis, 2017; Glymour et al., 2019; Peters et al., 2017; Shimizu, 2014, 2022) and has led to numerous applications of the methodology, for example, in epidemiology (Rosenström et al., 2012), economics (Moneta et al., 2013), neuroscience (Mills-Finnerty et al., 2014), and materials science (Campomanes et al., 2014; Liu et al., 2021). See `https://www.shimizulab.org/lingam/lingampapers` for a list of papers on the methodology and its applications.

Representative causal discovery packages are TETRAD (Scheines et al., 1998; Ramsey et al., 2020), pcalg (Kalisch et al., 2012) and bnlearn (Scutari and Denis, 2021). These packages are rich in classic methods based on conditional independence including constraint-based methods such as PC and FCI (Spirtes and Glymour, 1991; Spirtes et al., 1995) and greedy score-based methods such as GES (Chickering, 2002) and NOTEARS (Zheng et al., 2018), whereas TETRAD and pcalg only provide a basic method for the LiNGAM approach (Shimizu et al., 2006) based on independent component analysis (ICA) (Hyvärinen et al., 2001). Causal Discovery Toolbox (Kalainathan et al., 2020) gives a Python front end to perform methods of pcalg and bnlearn written in R, but only offers the basic ICA-based method for LiNGAM (Shimizu et al., 2006). Further, their implementation of a nonlinear causal discovery method based on a similar idea of LiNGAM (Hoyer et al., 2009) is limited to two variable cases. Tigramite (`https://github.com/jakobrunge/tigramite`) offers time series causal discovery methods based on conditional independence (Gerhardus and Runge, 2020), but does not exploit additional information on the functional forms of the causal relations of variables or distributions of variables, unlike LiNGAM-type methods.

Thus, in this paper, we present a Python package for performing various LiNGAM-type methods including time series cases (Hyvärinen et al., 2010; Kawahara et al., 2011), multiple-group cases (Shimizu, 2012; Kadowaki et al., 2013), mixed data cases (Zeng et al., 2022), hidden common cause cases (Maeda and Shimizu, 2020; Zeng et al., 2021), and (multivariate) nonlinear cases (Peters et al., 2014). The package covers most of the major LiNGAM-type methods already used in application papers and relevant extensions. Users can choose suitable methods depending on what they assume based on their background knowledge. We plan to extend it further and encourage others to join the development. Moreover, the package offers additional functionalities including evaluation of statistical reliability based on bootstrapping (Komatsu et al., 2010) and model evaluation based on the magnitude of error independence (Entner and Hoyer, 2011; Tashiro et al., 2014).

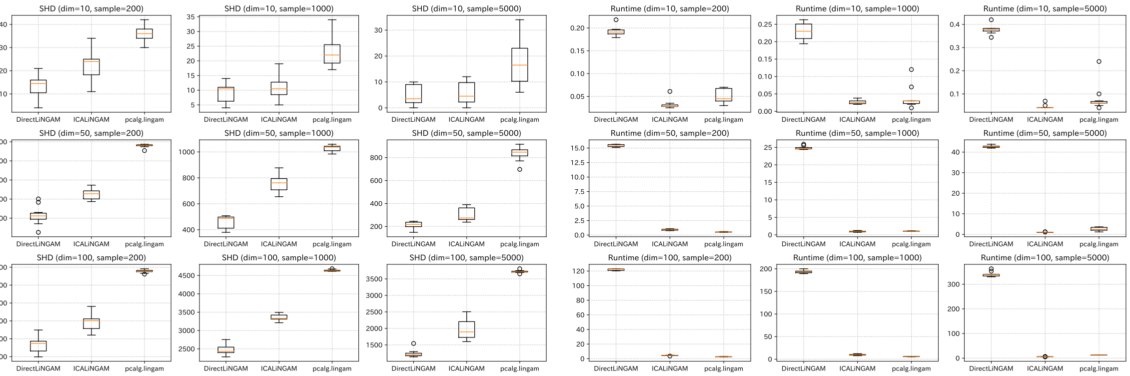

Figure 1: Left: SHD (Structural Hamming Distance), Right: Runtime.

2. Available models and estimation algorithms

This section gives a brief description of each of the LiNGAM methods available in this package. See the online documentation (`https://lingam.readthedocs.io/en/latest/`) for more details.

The most basic LiNGAM model (Shimizu et al., 2006) assumes that their causal relations are acyclic, with no hidden common causes, linearity, and non-Gaussian errors. The basic LiNGAM model can be estimated in different ways. This package implements two major algorithms: the original LiNGAM discovery algorithm based on ICA (Shimizu et al., 2006) and a direct method called DirectLiNGAM (Shimizu et al., 2011). The package further offers utilities to compute total effects between observed variables and their direct effects, drawing causal graphs using Graphviz, computing bootstrap probabilities of directed paths and edges, and incorporating prior knowledge on topological causal orders in the estimation by DirectLiNGAM. It further enables model evaluation by examining the independence of errors. Two extensions of the basic LiNGAM models are available in our package for multi-group analysis. First, Shimizu (2012) jointly estimates multiple LiNGAMs using multiple datasets from multiple sources by constraining their topological causal orders to be identical. This would enable a more accurate estimation of the LiNGAMs than estimating them separately, given the prior knowledge that they share is a topological causal order. Second, Kadowaki et al. (2013) consider performing causal discovery on paired samples and propose an estimation method for learning causal structures in longitudinal data that collects samples over time. Their algorithm can analyze causal structures, including topological causal orders, that may change over time.

We compared the accuracy and runtime of our implementation of the ICA-based LiNGAM algorithm with those of an existing package, pcalg, for different numbers of variables. We also tested our implementation of DirectLiNGAM for comparison. The python code used to generate artificial data in our experiments is available at `https://github.com/cdt15/lingam/blob/master/examples/data/GenerateDatasets.ipynb`. Fig. 1 shows that our implementation of DirectLiNGAM was more accurate than our and pcalg implementations of ICA-based LiNGAM. Our implementation of ICA-based LiNGAM was faster than its pcalg version, wheareas DirectLiNGAM was slower than our and pcalg implementations of ICA-based LiNGAM.

The package further offers two time series extensions of the basic LiNGAM: VAR-LiNGAM (Hyvärinen et al., 2010) combines it with vector autoregressive models (VAR), and VARMA-LiNGAM (Kawahara et al., 2011) does the same with vector autoregressive moving average models (VARMA). The package also provides a mixed date extension of the basic LiNGAM: Linear Mixed (LiM) causal discovery algorithm extends LiNGAM to handle the mixed data that consists of both continuous and discrete variables (Zeng et al., 2022). Further, our package can perform a nonlinear causal discovery RESIT (Peters et al., 2014) assuming a nonlinear additive noise model with acyclicity and no hidden common causes (Hoyer et al., 2009). Users can use a nonlinear regression from those implemented in scikit-learn (Pedregosa et al., 2011).

Another important extension is LiNGAM with hidden common causes or latent factors (Hoyer et al., 2008; Zeng et al., 2021). We implemented the RCD algorithm (Maeda and Shimizu, 2020) and CAM-UV algorithm (Maeda and Shimizu, 2021). The RCD algorithm allows the existence of hidden common causes and outputs a causal graph, where a bi-directed arc indicates the pair of variables that have the same hidden common causes and a directed arrow indicates the causal direction of a pair of observed variables that are not affected by the same hidden common causes. CAM-UV is its nonlinear variant and assumes the structural equations additive in the observed variables and errors. We also implemented the Multi-Domain LiNGAM algorithm for latent factors (MD-LiNA) (Zeng et al., 2021). Given the observed measurement data, MD-LiNA allows to locate the latent factors in addition to uncovering the causal structure between such latent factors of interests.

3. Design, API and future development

To facilitate application by machine learning users, we designed the model with the fit function, similar to scikit-learn. The standard flow is to call the fit method shown below to build the model after the model instance is created.

```
model = lingam.DirectLiNGAM()
model.fit(X)
```

After model building, the graph is returned as an adjacency matrix.

We plan to continue further development of the package. We also encourage others to join the project. It is easy to extend the model by following the Contribution Guide in the online documentation and further referring to other models. The Contribution Guide includes the code style and the ways to check the format, write the documentation, perform unit tests, and create a pull request. The package would benefit applied researchers and practitioners in enjoying the results of recent developments in causal discovery and deriving better causal conclusions based on domain knowledge and data.

Acknowledgments

We thank support from JSPS Grant-in-Aid for Scientific Research (C) #20K11708, ONR N00014-20-1-2501, and China Postdoctoral Science Foundation (2022M711812).

References

Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

Pablo Campomanes, Marilisa Neri, Bruno A.C. Horta, Ute F. Roehrig, Stefano Vanni, Ivano Tavernelli, and Ursula Rothlisberger. Origin of the spectral shifts among the early intermediates of the rhodopsin photocycle. *Journal of the American Chemical Society*, 136(10):3842–3851, 2014.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Doris Entner and Patrik O. Hoyer. 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, 2011.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems*, 33: 12615–12625, 2020.

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, Shohei Shimizu, Antti Kerminen, and Markus. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

Patrik O. Hoyer, Dominik Janzing, Joris Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, pages 689–696. Curran Associates Inc., 2009.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley, New York, 2001.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregressive model using non-Gaussianity. *Journal of Machine Learning Research*, 11:1709–1731, 2010.

Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating causal effects using weighting-based estimators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10186–10193, 2020.

Kento Kadowaki, Shohei Shimizu, and Takashi Washio. 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, 2013.

Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. Causal discovery toolbox: Uncovering causal relationships in python. *Journal of Machine Learning Research*, 21(37):1–5, 2020. URL `http://jmlr.org/papers/v21/19-187.html`.

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.

Yoshinobu Kawahara, Shohei Shimizu, and Takashi Washio. Analyzing relationships among ARMA processes based on non-Gaussianity of external influences. *Neurocomputing*, 74 (12-13):2212–2221, 2011.

Yusuke Komatsu, Shohei Shimizu, and Hidetoshi Shimodaira. Assessing statistical reliability of LiNGAM via multiscale bootstrap. In *Proceedings of 20th International Confereznce on Artificial Neural Networks (ICANN2010)*, pages 309–314. Springer, 2010.

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

Takashi Nicholas Maeda and Shohei Shimizu. 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.

Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Proc. 37th Conference on Uncertainty in Artificial Intelligence (UAI2021)*, pages 97–106. PMLR, 2021.

Colleen Mills-Finnerty, Catherine Hanson, and Stephen Jose Hanson. Brain network response underlying decisions about abstract reinforcers. *NeuroImage*, 103:48–54, 2014.

Alessio Moneta, Doris Entner, Patrik O. Hoyer, and Alex Coad. Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, 75(5):705–730, 2013.

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

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Joseph D. Ramsey, Daniel Malinsky, and Kevin V. Bui. algcomparison: Comparing the performance of graphical structure learning algorithms with TETRAD. *Journal of Machine Learning Research*, 21(238):1–6, 2020.

Tom Rosenström, Markus Jokela, Sampsa Puttonen, Mirka Hintsanen, Laura Pulkki-Råback, Jorma S Viikari, Olli T Raitakari, and Liisa Keltikangas-Järvinen. Pairwise measures of causal direction in the epidemiology of sleep problems and depression. *PLOS ONE*, 7(11):e50841, 2012.

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.

Marco Scutari and Jean-Baptiste Denis. *Bayesian networks: with examples in R*. Chapman and Hall/CRC, 2021.

Shohei Shimizu. Joint estimation of linear non-Gaussian acyclic models. *Neurocomputing*, 81:104–107, 2012.

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

Shohei Shimizu. *Statistical Causal Discovery: LiNGAM Approach*. Springer, Tokyo, 2022.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7: 2003–2030, 2006.

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

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:67–72, 1991.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Springer Verlag, 1993. (2nd ed. MIT Press 2000).

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proc. 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI1995)*, pages 491–506, 1995.

Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. ParceLiNGAM: A causal ordering method robust against latent confounders. *Neural Computation*, 26(1): 57–83, 2014.

Yan Zeng, Shohei Shimizu, Ruichu Cai, Feng Xie, Michio Yamamoto, and Zhifeng Hao. Causal discovery with multi-domain LiNGAM for latent factors. In *Proc. 30th International Joint Conference on Artificial Intelligence (IJCAI2021)*, 2021.

Yan Zeng, Shohei Shimizu, Hidetoshi Matsui, and Fuchun Sun. Causal discovery for linear mixed data. In *Proceedings of the First Conference on Causal Learning and Reasoning (CLeaR2022)*, volume 177 of *Proceedings of Machine Learning Research*, pages 994–1009. PMLR, 11–13 Apr 2022.

Kun Zhang and Aapo Hyvärinen. Nonlinear functional causal models for distinguishing causes form effect. In *Statistics and Causality: Methods for Applied Empirical Research.* Wiley & Sons, 2016.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.