DSSV 2023 Data Science, Statistics & Visualisation
EC 2023 European Conference on Data Analysis

TSLiNGAM: DirectLiNGAM under heavy tails

Sarah Leyder[1], Jakob Raymaekers[2], and Tim Verdonck[1]

[1]Department of Mathematics, University of Antwerp, Belgium
[2]Department of Quantitative Economics, Maastricht University, The Netherlands

One of the established approaches to causal discovery consists of the direct acyclic graphs (DAGs) framework. DAGs represent the variables of interest as nodes in a graph where directed edges between nodes stand for causal relations. These DAGs are then typically complemented by structural causal models (SCMs) used to describe the functional dependence of an effect on its causes. Possible identifiability of the SCM given data depends on assumptions made on the noise variables and on restrictions made on the functional classes in the SCM. For instance, in Shimizu et al. [1], the functional class is restricted to linear functions and the disturbances have to be non-Gaussian and mutually independent. This model is known as the linear, non-Gaussian, acyclic model (LiNGAM).

In our work, we focus on DirectLiNGAM (Shimizu et al. [2]), a direct method to obtain the causal LiNGAM model by using linear regression as a tool to remove the effect of a variable. For this, ordinary least squares (OLS) is used to estimate the slope. It is known that OLS has favorable properties in the presence of normal errors, however, in the LiNGAM model, the overarching assumption is that the error variables are non-Gaussian. Hence in this context, the optimality properties of OLS disappear. Therefore a notable improvement to the DirectLiNGAM algorithm can be made by plugging in a more appropriate slope estimator.

In this talk we suggest two adjustments to the DirectLiNGAM method by applying a simple plug-in strategy, justified by a detailed theoretical underpinning. First, we increase the efficiency of the causal algorithm by using a better suited slope estimator instead of the classical OLS. Second, we study the use of a different independence measure to speed up the computational time. For these adjustments, we study the Theil-Sen estimator of Theil and Sen [3], the repeated median of Siegel [4] and the distance correlation of Székely et al. [5]. It turns out that our method, called TSLiNGAM, performs significantly better on heavy-tailed data and discovers the right causal order on smaller sample sizes.

Keywords: Causal discovery, Structural causal models, LiNGAM, Efficiency, Heavy-tailed data

[1] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery.," *Journal of Machine Learning Research*, vol. 7, 2006.
[2] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, "Directlingam: A direct method for learning a linear non-gaussian structural equation model," *Journal of Machine Learning Research*, vol. 12, 2011.
[3] P. K. Sen, "Estimates of the regression coefficient based on kendall's tau," *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1379–1389, 1968.
[4] A. F. Siegel, "Robust regression using repeated medians," *Biometrika*, vol. 69, no. 1, pp. 242–244, 1982.
[5] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769 – 2794, 2007.