LiNGAM with latent confounding factors using optimization

Technical Report

https://github.com/Sanaxen/Statistical_analysis
Research & Development Group
-bccatfalcon@gmail.com

October 24, 2021

Abstract

In recent years, as mechanisms such as **IoT**, **DX**, and **AI** have been introduced, the need for these technologies has increased. More than that, however, there is a need to identify the root causes of problem occurrences and elucidate their mechanisms. One of the major obstacles is the assumption that there are no unobserved potential common causes. The assumption that there are no unobserved potential common causes is a major obstacle for LiNGAM [5] to produce results. Although there are several previous studies, this method proposes a way to solve this problem by using known methods together.

Keywords Causal search · Unobserved confounding · non-Gaussianity

1 Introduction

It is well known that statistical causal search (LiNGAM) cannot estimate the correct causal structure when there are unobserved latent common variables. This is not necessarily a problem with LiNGAM itself, since the absence of unobserved latent common variables is a condition for applying LiNGAM, but in practice, knowing all variables and having them as observed data is a major problem in practice. LiNGAM is represented by the following equation. The matrix B represents the causal structure between the data. Since the data distribution is assumed to be non-Gaussian, if the data distribution is not Gaussian, and moreover linear, the causal structure can be represented by this equation.

$$\mathbf{x_i} = \sum_{\mathbf{k(j)<k(i)}} \mathbf{B_{ij}} \mathbf{x_j} + \varepsilon_\mathbf{i} \tag{1}$$

This assumes that the residuals $\varepsilon_\mathbf{i}$ are independent of each other. This assumes that there are no unobserved latent common causes in the data. The model can be extended to the case where there are unobserved latent common causes as follows.

$$\mathbf{x_i} = \mu_\mathbf{i} + \sum_{\mathbf{k(j)<k(i)}} \mathbf{B_{ij}} \mathbf{x_j} + \sum_{\mathbf{l=1\cdots L}} \lambda_{\mathbf{il}} \mathbf{f_l} + \varepsilon_\mathbf{i} \tag{2}$$

The unobserved potential common factor il needs to be determined in some way. However, since the number of unobserved potential common factors, L, is also unknown to begin with, the number of potential common factors, L, must also be determined, which will be very difficult.

Representative Previous Research LvLiNGAM (Latent variable LiNGAM) ([2]), Pairewise LvLiNGAM algorithm([1]) ([7]) and LiNGAM Mixture Model ([4]) , RCD([3]) have been proposed.

2 Overview of the proposed method

$$x = \boldsymbol{B}^{(c)}(\boldsymbol{x} - \boldsymbol{\mu}^{(c)}) + \boldsymbol{\mu}^{(c)} + \boldsymbol{e}^{(c)} \tag{3}$$

([4])

$$x_j = \sum_{k(j)<k(i)} B_{i,j}(x_j - \mu_j) + e_i + \mu_i \tag{4}$$

Problem to be solved The basic idea is to attribute the entire problem to the optimization problem, and to find the cause-effect relationship of the whole rather than from the parts.

The optimization conditions and parameters are

- The regression model is correct (residuals are minimized).
- The distribution parameters of μ are obtained so that the amount of mutual information in the residuals is minimized as much as possible.
- If there is prior knowledge of the causal relationship, add a penalty F due to the prior knowledge.
- Use a simple and easy to maintain algorithm

The above two conditions are to satisfy the LiNGAM requirement. It is a requirement of LiNGAM that the independence of residuals is equivalent to the absence of unobserved common causes. And it means that the data can be explained as a linear model.

It can be solved as a constrained multi-objective optimization problem with prior knowledge penalty, where the distribution parameters of μ are obtained so that the residuals of each variable do not fall into local solutions independently, while increasing the accuracy of the regression model under this condition.

2.1 Optimization

Optimization minimizes the amount of residuals and mutual information. This is a multi-objective optimization calculation. Determine the distribution parameters of μ so that the amount of mutual information between residuals is minimized as much as possible.

$$\mu = \arg\min\left[\max\left(\max\left(\mathbf{e_i}\right), \max\left(\iint p\left(\mathbf{e_i}, \mathbf{e_j}\right)\log\left(\frac{p\left(\mathbf{e_i}, \mathbf{e_j}\right)}{p\left(\mathbf{e_i}\right)p\left(\mathbf{e_j}\right)}\right)d\mathbf{e_i}d\mathbf{e_j}\right)\right) + \theta\ \mathbf{penaltyF}\right] \tag{5}$$

penaltyF is a function that is zero if the prior knowledge of the causal structure is satisfied, but can be greater than zero otherwise. θ is a positive, sufficiently large value.[1]

$$\mu \sim Generalized\_Gaussian(\beta, \rho, \widetilde{x}) \equiv \frac{\beta^{\frac{1}{2}}}{2\Gamma(1+\frac{1}{\rho})}\exp\left(-\beta^{\frac{1}{2}}|x-\widetilde{x}|^{\rho}\right) \tag{6}$$

[1] $p(e_i)$ is the probability density of e_i.

$Generalized\_Gaussian(\beta, \rho, \tilde{x})$

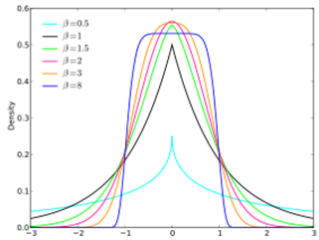

Figure 1: distribution

While improving the accuracy of the regression model, determine the parameters that determine the distribution of $\mu\ (\beta, \rho, \widetilde{x})$ so that the residuals of each variable are independent, but do not fall into local solutions. The solution is obtained as a multi-objective optimization problem.

2.1.1 Loss computation

Perform multi-objective optimization with the expanded Tchebyshev-scalarization function or the weighted average scalarization function.

$$\textbf{loss}=max(w1\ independ,\ w2\ residual)+\varepsilon(w1\ independ+\ \ w2\ residual) \tag{7}$$

Expanded Tchebyshev-scalarization function

$$\textbf{loss}=w1\ independ+\ \ w2\ residual \tag{8}$$

Weighted average scalarization function

$$w=\sqrt{best\_independ+best\_residual}$$

$$w1=\frac{best\_independ}{w}$$

$$w2=\frac{best\_residual}{w}$$

$$independ=\max_{i<j}\left(\iint p\left(\mathbf{e_i},\mathbf{e_j}\right)\log\left(\frac{p\left(\mathbf{e_i},\mathbf{e_j}\right)}{p\left(\mathbf{e_i}\right)p\left(\mathbf{e_j}\right)}\right)d\mathbf{e_i}d\mathbf{e_j}\right) \tag{9}$$

$$residual=\max_{i}(\mathbf{e_i}) \tag{10}$$

$best\_independ = maximum\ amount\ of\ mutual\ information\ between\ residuals\ at\ the\ best\ value\ of\ loss$
$best\_residual = the\ maximum\ value\ of\ residuals\ at\ the\ best\ value\ of\ loss$
$\varepsilon = 0.0001$

This optimization minimizes the residuals and the amount of mutual information, but in most cases the magnitudes of the maximum absolute values of the residuals and the amount of mutual information are quite different. However, in most cases, the magnitude of the maximum absolute value of the residuals and the magnitude of the mutual information are quite different. If the calculation of multi-objective optimization is performed as it is, only one of them may be minimized numerically. In this experiment, we found that both can be minimized by normalizing the maximum absolute value of the residuals used in the loss calculation and the maximum absolute value of the mutual information. Normalization means to set the maximum absolute value of the residuals and the maximum absolute value of the mutual information calculated in the first evaluation of the optimization calculation to 1.0. In our experiments, we found that it was not necessary to normalize the mutual information.

3 Experiment

The data used in the experiment is artificial data and unobserved conditions are created by excluding common cause variables. The experiment generates the μ from a generalized Gaussian distribution **??** and estimates the causal structure with LiNGAM. From this structure, we can calculate the residuals and mutual information for each variable. Using this, we can search for μ until the residuals and mutual information are minimized.

In our experiments, we used the **Probabilistic Meta-Algorithm (SA)** to perform the calculations.

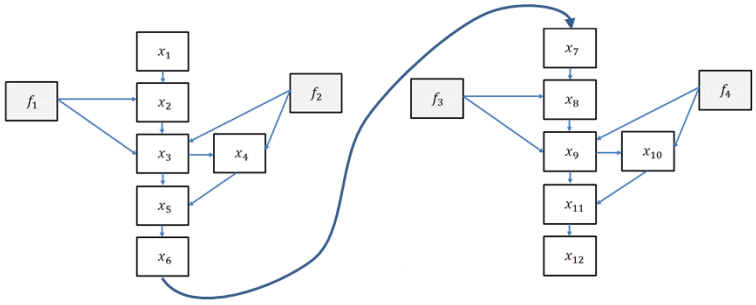

Figure 2: example causal structure

For example, we created 19 cases of data as shown in 2. f_1, f_2, f_3, f_4 are the confounding variables. By not including these variables in the data, we generate data with unobserved common causes.

error term e_i was set by a uniformly distributed random number.

$x_1 = e_1$
$x_2 = 0.4x_1 + 0.8f_1 + e_2$
$x_3 = 0.3x_2 + 0.7f_1 + 0.7f_2 + e_3$
$x_4 = 0.2x_3 + 0.8f_2 + e_4$
$x_5 = 0.5x_3 + 0.5x_4 + e_5$
$x_6 = 0.5x_5 + e_6$
$x_7 = 0.5x_6 + e_7$
$x_8 = 0.4x_7 + 0.8f_3 + e_8$
$x_9 = 0.3x_8 + 0.7f_3 + 0.7f_4 + e_9$
$x_{10} = 0.2x_9 + 0.8f_4 + e_{10}$
$x_{11} = 0.5x_9 + 0.5x_{10} + e_{11}$
$x_{12} = 0.5x_{11} + e_{12}$

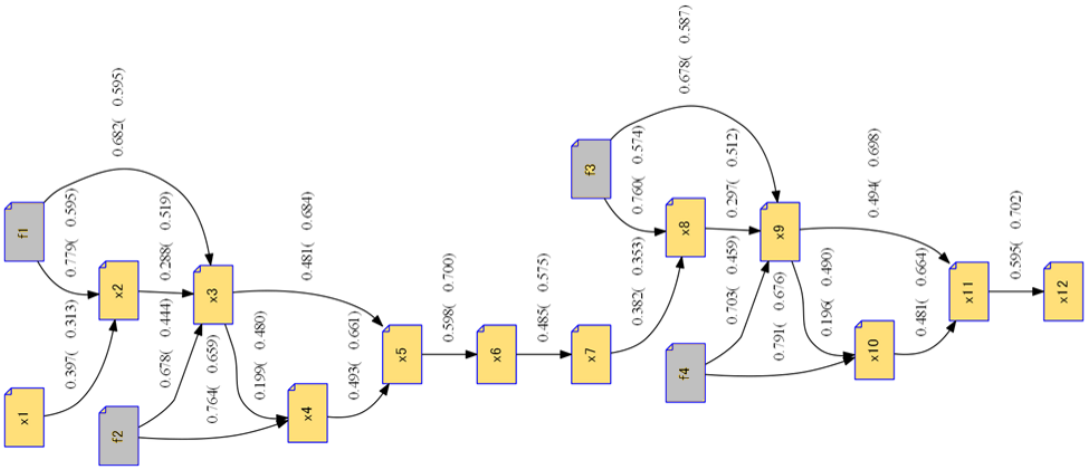

Figure 3: example

3.1 Experimental results

In LiNGAM, we got 30% of the total causal direction wrong.(4)

Our proposed method correctly estimated the direction of all causal relationships(5)

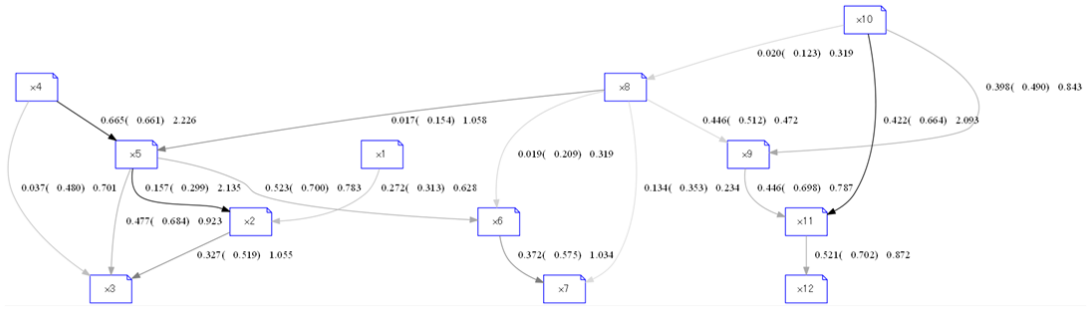

Figure 4: ICA-LiNGAM

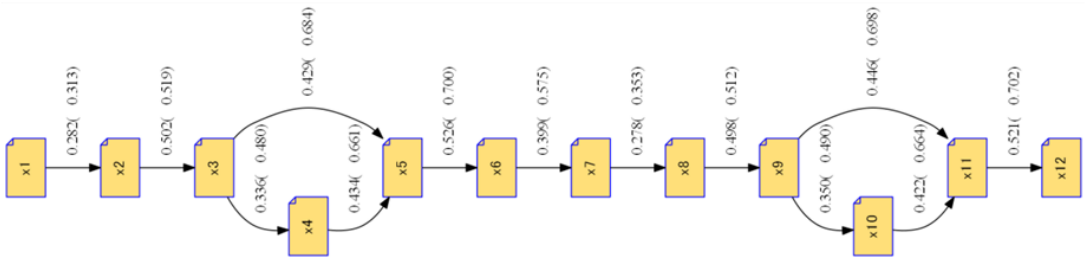

Figure 5: Our approach

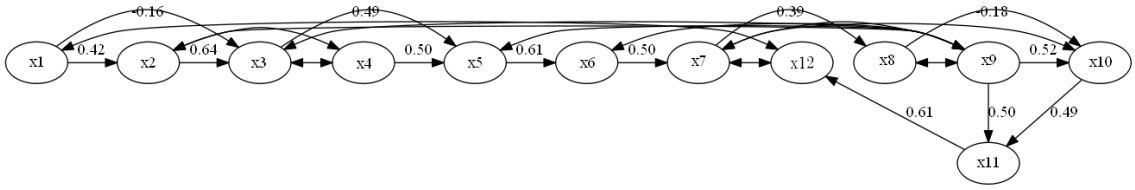

Figure 6: RCD([3])

results of the experiment for 19 cases.

Table 1: 19 cases

	ICA-LiNGAM	Our	Direct-LiNGAM
Average	84%	97%	85%
Data without unobserved common cause	96%	100%	100%
Data with unobserved common cause	75%	95%	74%

ICA-LiNGAM ([5]) , Direct-LiNGAM ([6]) , Even in the absence of unobserved data, the independence of the residuals and the residuals is corrected by optimization to create a more accurate linear model.7 The

amount of data is a uniform 10000 lines of data. The value of the accuracy rating is expressed as the ratio of the number of edges estimated in the correct direction for all edges of the estimated causal structure.

$$Accuracy = \frac{number\ of\ edges\ estimated\ in\ the\ correct\ direction}{number\ of\ all\ edges\ in\ the\ causal\ structure}$$

Figure 7: Optimization calculation

I estimated the wrong result with one piece of data.

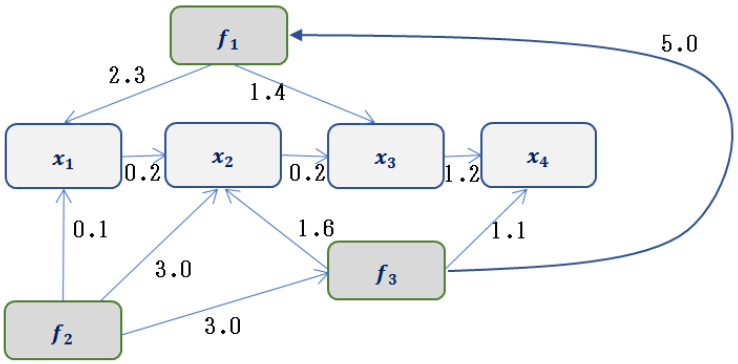

Figure 8: Example of calculating an incorrect result

The causal structure of the case was as shown in the figure.8
$f_2 = e_{f_2}$
$f_3 = 3.0f_2 + e_{f_3}$
$f_1 = 5.0f_3 + e_{f_1}$
$x_1 = 2.3f_1 + 0.1f_2 + e_1$
$x_2 = 0.2x_1 + 3.0f_2 + 1.6f_3 + e_2$
$x_3 = 0.2x_2 + 1.4f_1 + e_3$
$x_4 = 1.2x_3 + 1.1f_3 + e_4$

This data did not allow us to infer the correct causal structure. We will compare it with RCD([3]) in the next section.

3.2 Comparison with RCD([3])

Our approach had the wrong result. So we tried to apply RCD.

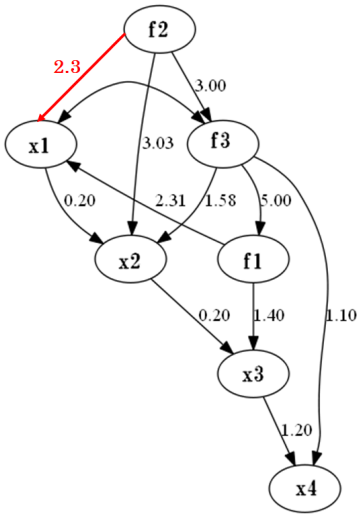

Figure 9: results for RCD

Experimental results for RCD in the absence of unobserved common cause data. RCD did not produce any red arrows 9. A bi-directional path has been added to $x_1 and f_3$. This implies that there is an unobserved common cause between $x_1 and f_3$.

RCD estimates that there is an unobserved common cause between x1 and f3 that does not actually exist, but there is no unobserved common cause in fact.

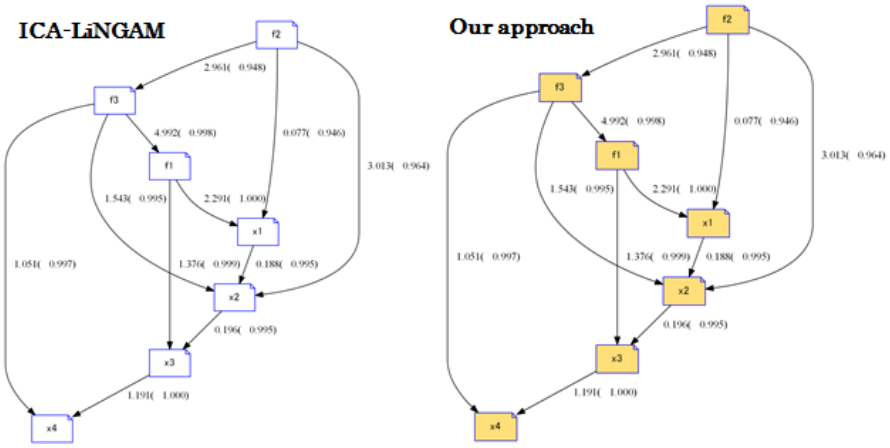

Figure 10: ICA-LiNGAM & Our approach

Neither ICA-LiNGAM nor our proposed method results in the existence of wrong paths. However, since we have removed redundant paths, we cannot make an equivalent comparison, but we can see that neither of them estimates the wrong path.

Experimental results of RCD in the presence of unobserved common cause data. Data with $f_1, f_2, and f_3$ deleted

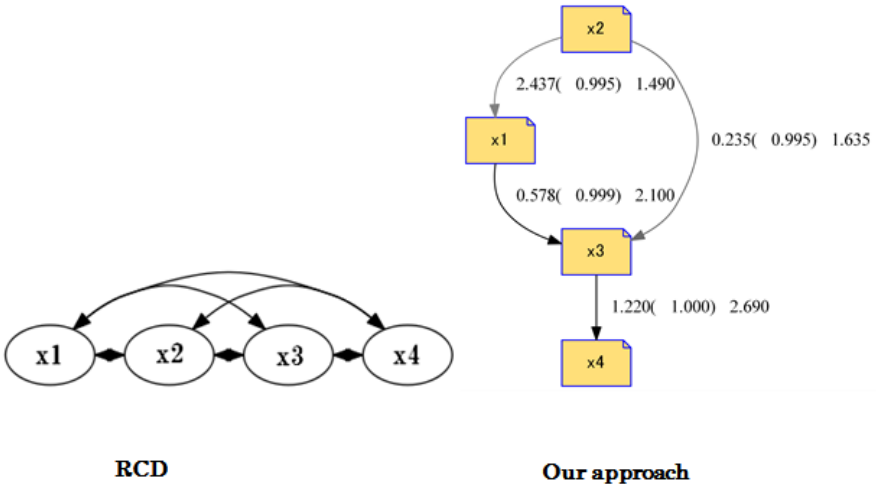

Figure 11: RCD & Our approach

Our method showed the opposite direction betweenx_1, x_2 and x_1, x_3. RCD infers that there is an unobserved common cause between all variables.

Experiments have shown that when unobserved common causes are very large relative to other observed variables, they can contain erroneous results. Therefore, we conducted an experiment with data adjusted so that the common cause of unobserved variables is not relatively large.

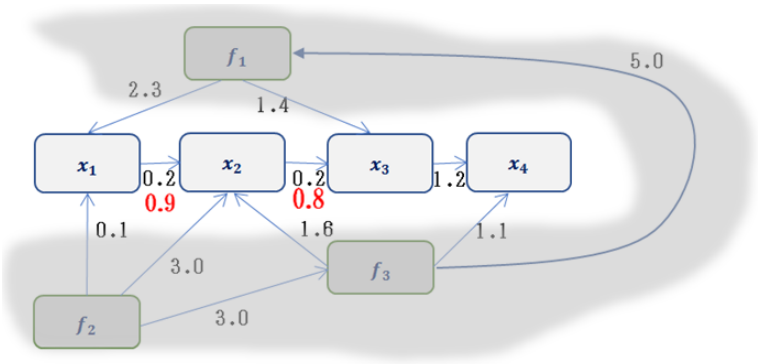

Figure 12: data

Note that the structure of the causal graph is exactly the same

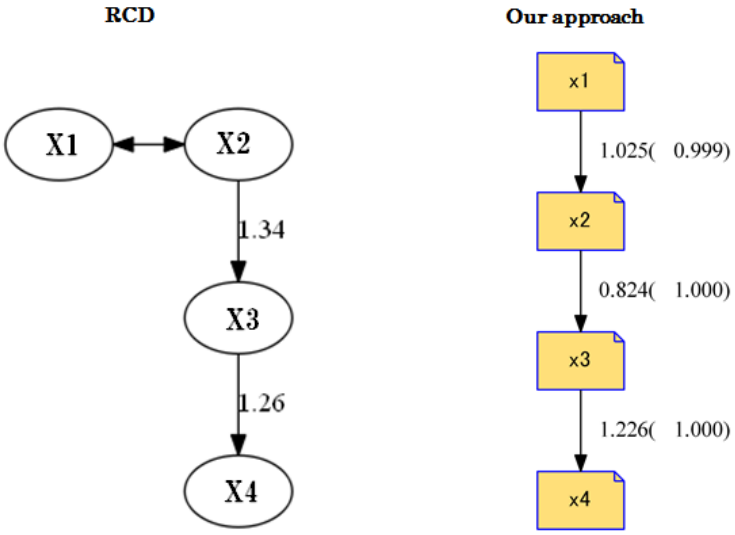

Figure 13: RCD & Our approach

In RCD, there is an unobserved common variable between $x_1 and x_2$, and the proper relationship is estimated for $x_2, x_3, and x_4$.

In our method, there is no wrong path. The coefficient values for the causal relationships also seem to be fine.

4 Discussion

The advantage of this method is that it can be combined with existing computational techniques to build a practical and maintainable system in a short period of time, and there are almost no parameters that need to be adjusted. The only parameter that needs to be adjusted is the maximum number of optimization iterations, which needs to be set by the user, but in most cases it converges after about 10000 iterations.

We also found a couple of problems. First, the computation time is slow. In particular, with 30 variables and 10000 rows of data, it can take up to 7 hours to finish the calculation. We believe that this can be overcome by making good use of GPU acceleration and parallelization. The biggest challenge is that if the contribution of the unobserved potential common variable is very large compared to the other variables, it may calculate the wrong path. 8 1

Another problem is that LiNGAM assumes continuous data, so it cannot be applied to discontinuous sensor data. We used artificial data in this experiment, but we would like to apply it to real-world data to clarify the scope of application and improve it if possible.

5 Acknowledgments

The style file used to create the Tex of this paper can be found at https://github.com/kourgeorge/arxiv-style. We would like to thank George Kour for making available the style file used to create the text for this paper.

RCD was calculated using the following method. https://github.com/cdt15/lingam was used. We thank the developers T. Ikeuchi, G. Haraoka, M. Ide, W. Kurebayashi, and S. Shimizu.

References

[1] Doris Entner and Patrik O Hoyer. Discovering unconfounded causal relationships using linear non-gaussian models. In *JSAI International Symposium on Artificial Intelligence*, pages 181–195. Springer, 2010.

[2] Patrik O Hoyer, Shohei Shimizu, Antti J 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.

[3] 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.

[4] Shohei Shimizu and Kenneth Bollen. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *J. Mach. Learn. Res.*, 15(1):2629–2652, 2014.

[5] 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.

[6] 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. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

[7] 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.

A Statistical Causality

A.1 LiNGAM Model (Linear Non-Gaussian Acyclic Model)

Given n observed variables $x = \{x_1, x_2, \cdots, x_n\}$, estimates which variables affect which variables and by how much (causality).

The $\epsilon = \{\epsilon_1, \epsilon_2, \cdots, \epsilon_n\}$ are unobserved variables that may have an impact on the observed variables, such as noise and other factors that cannot be directly observed.

Assumption 1

$\epsilon = \{\epsilon_1, \epsilon_2, \cdots, \epsilon_n\}$ are independent of each other and follow a non-Gaussian continuous distribution.

Note: Assuming independence means that for each observed variable, there is no unobserved common cause.

Assumption 2

Linearity

$$\mathbf{x} = \mathrm{Bx} + \varepsilon$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Hereafter, the above matrix B will be referred to as the acyclic directed graph matrix.

Assumption 3

We assume acyclicity in the causality of each observed variable. This means that when we consider a causal graph, no matter which variable we start from and follow the relationship, we cannot come back to the original variable, so

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ B_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ B_{n1} & B_{n2} & \cdots & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$x_1 = \epsilon_1$$

$$x_2 = B_{21} x_1 + \epsilon_2$$

$$x_i = B_{i1} x_1 + B_{i2} x_2 + B_{i2} x_3 + \cdots + B_{\mathrm{ik}} x_k + \epsilon_i \qquad k < i$$

$$x_n = B_{n1} x_1 + B_{n2} x_2 + \cdots + \epsilon_n$$

$$x_1 = \epsilon_1$$

$$x_2 = B_{21} x_1 + \epsilon_2$$

The non-zero components represent a measure of how much each observable affects the observables on the left-hand side.

A.2 Calculation procedure for ICA-LiNGAM

Calculating the Acyclic Directed Graph Matrix

irst, we need to find the $\epsilon = \{\epsilon_1, \epsilon_2, \cdots, \epsilon_n\}$ that is not directly observed.

$$\mathbf{x} = \mathrm{Bx} + \varepsilon$$

.

$$\varepsilon = \mathbf{x} - \mathbf{Bx} \rightarrow \varepsilon = (\mathbf{I} - \mathbf{B})\mathbf{x}$$

$$\mathbf{W} \equiv (\mathbf{I} - \mathbf{B})$$

$$\varepsilon = \mathbf{Wx}$$

.

Estimate W here (can be calculated using FastICA, an estimation method for independent component analysis).

If we know the appropriate substitution matrix P and the appropriate matrix D of the scaling transformation, we can obtain the restoration matrix

$$\mathbf{W}_{\mathrm{ICA}}$$

that can be calculated using the FastICA method of independent component analysis estimation.

$$\mathbf{W} \cong \mathbf{D^{-1}PW}_{\mathrm{ICA}}$$

.

The substitution matrix P rearranges the rows such that the absolute value of the diagonal component of the restoration matrix $\mathbf{W}_{\mathrm{ICA}}$ is maximized.

The scale transformation matrix D is $D = diag(\mathbf{PW}_{\mathrm{ICA}})$

$$\mathrm{diag}\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nn} \end{pmatrix} \equiv \begin{pmatrix} X_{11} & 0 & \cdots & 0 \\ 0 & X_{22} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & X_{nn} \end{pmatrix}$$

$$\mathbf{B} = \mathbf{I} - \mathbf{W} \cong \mathbf{I} - \mathbf{D^{-1}PW}_{\mathrm{ICA}}$$

This does not necessarily mean that the acyclic directed graph matrix B computed with this guarantees acyclicity. In other words, redundant causal relations appear, so we remove them. The proposed method is as follows.

Forcibly replace $n(n+1)/2$ components of B with zero components in the order of decreasing absolute value.

Check if each row of B can be rearranged into a lower triangular matrix, and if not, replace the next smallest absolute value component with zero.