最適化による潜在的交絡因子を考慮したLiNGAM

Technical Report

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

February 2, 2022

Abstract

近年、製造業などで**IoT**や**DX**、**AI**などの仕組みが導入され始めたことでこれらの技術の必要性が製造業以外の様々な分野でも高まってきています。 しかし、こういった仕組みの導入に伴い問題発生の根本原因を特定し、そのメカニズムを解明することが求められています。この課題解決に不可欠な技術が因果探索とよばれる技術ですがその代表とも言えるLiNGAM[5]を適用する上で大きな障害の一つが、「観測されていない潜在的な共通原因は存在しない」という仮定です。 観察されていない潜在的な共通原因がないという仮定は、LiNGAM[5] が結果を出す上での大きな障害となっているためこういった制約を無くすべく先行研究はいくつかありますが、本手法は複雑なアルゴリズムを使わずに既知の手法を併用することでこの問題を解決する方法を提案しています。

Keywords Causal search · Unobserved confounding · non-Gaussianity

1 Introduction

統計的因果探索（LiNGAM）では、観測されていない潜在的な共通変数がある場合、正しい因果構造を推定できないことがよく知られている。未観測の潜在的共通変数がないことがLiNGAM適用の条件なので、これは必ずしもLiNGAM自体の問題ではありませんが、実際にはすべての変数を知っていて、それを観測データとして持つことは実用上大きな問題となります。 LiNGAMは以下の式1で表されます。行列Bはデータ間の因果構造を表しています。データの分布はガウス分布ではないと仮定されているので、データの分布がガウス分布ではなく、さらには線形であれば、因果構造は次の式で表すことができる。

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

さらに残差ε_iが互いに独立であることを仮定しています。 これは、データに観測されていない潜在的な共通原因がないことを仮定しています。このモデルは、観測されていない潜在的な共通原因がある場合には、次のように拡張することができます。

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

観測されていない潜在的共通因子 $\lambda_{il} f_l$ を何らかの方法で決定する必要があります。しかし、観測されていない潜在的共通因子の数Lもそもそも未知であるため、潜在的共通因子の数Lも決定しなければならず、B_{ij}を求めるのは非常に困難です。

代表的な先行研究

LvLiNGAM (Latent variable LiNGAM) ([2]), Pairewise LvLiNGAM algorithm([1]) ([7]) and LiNGAM Mixture Model ([4]) , RCD([3])

2 Overview of the proposed method

2.1 解決すべき問題

LiNGAM Mixture Model ([4])では式2の$\sum_{l=1\cdots L}\lambda_{il}f_l$ をμに押し込むことで陽に交絡因子を指定する事を回避しています。

$$x_j = \sum_{k(j)<k(i)} B_{i,j}(x_j - \mu_j) + \varepsilon_i + \mu_i \tag{3}$$

その変わりμを推定する必要があります。そこで２変数の関係でμの事前分布を設定しベイズの枠組みでモデル選択を行うことで推定を行っています。具体的には色々な因果構造を沢山用意して、それぞれの周辺尤度を計算して比較することで最も妥当な因果構造を推定しています。これを他の関係についても行うことで最終的に全体の因果構造を推定しています。私たちの基本的な考え方は、問題全体を最適化問題に帰着させ、部分からではなく、初めから全体の因果関係を見出すことです。

最適化条件とパラメータは

- 回帰モデルが正しい（残差が最小になる）事.
- 残差の相互情報量を可能な限り小さくするように、μの分布パラメータを求める事.
- 因果関係の事前知識がある場合は、事前知識によるペナルティFを加える。
- シンプルでメンテナンスの容易なアルゴリズムを使用する。

最初の2つの条件は、LiNGAMの要件を満たすことです。残差の独立性は、未観測の共通原因がないことと同等であるというのが、LiNGAMの要件です。そして、データが線形モデルとして説明できることを意味しています。 この条件で回帰モデルの精度を高めつつ、各変数の残差が独立して局所解に陥らないように、μの分布パラメータを求める、事前知識ペナルティ付きの制約付き多目的最適化問題として解くことができます。

残差と相互情報量の両方を最小化するため多目的最適化計算を行う事になります。 残差の相互情報量ができるだけ少なくなるように、μの分布パラメータを決定します。

$$\mu = arg\ min\left[\max\left(\max(e_i), \max_{i<j}\left(\iint p(e_i,e_j)\log\left(\frac{p(e_i,e_j)}{p(e_i)p(e_j)}\right)de_i de_j\right)\right) + \theta\ penaltyF\right] \tag{4}$$

*theta penaltyF*は因果構造の事前知識が満たされていれば0であり、そうでなければ0より大きい値をとることができる関数です。 [1]

$$\mu\ \sim\ Generalized\_Gaussian(\beta,\rho,\widetilde{x}) \equiv \frac{\beta^{\frac{1}{2}}}{2\Gamma(1+\frac{1}{\rho})}\exp\left(-\beta^{\frac{1}{2}}|x-\widetilde{x}|^{\rho}\right) \tag{5}$$

μは一般化ガウス分布5を使うことでパラメータを指定することで様々な形の分布をシステマティックに与える事が可能になります。

[1] $p(e_i)$は、e_iの確率密度

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

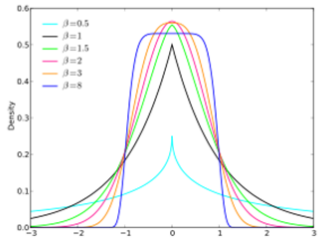

Figure 1: distribution

以上のようにして回帰モデルの精度を向上させつつ、各変数の残差が独立しているが、局所解に陥らないように$\mu\,(\beta, \rho, \tilde{x})$ の分布を決定するパラメータを決定します。そして問題は多目的最適化問題に帰着されます。

2.1.1 Loss computation

拡大チェビシェフ・スカラー化関数または加重平均スカラスカラー化関数を用いて多目的最適化を行う事ができます。

$$\textbf{loss}=max\,(w1\ independ,\ w2\ residual) + \varepsilon\,(w1\ independ + \ \ w2\ residual) \tag{6}$$

Expanded Tchebyshev-scalarization function

$$\textbf{loss}=w1\ independ + \ \ w2\ residual \tag{7}$$

Weighted average scalarization function

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

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

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

$$independ = \max_{i<j}\left(\iint p(e_i, e_j)\log\left(\frac{p(e_i, e_j)}{p(e_i)p(e_j)}\right)de_i de_j\right) \tag{8}$$

$$residual = \max_i(\mathbf{e_i}) \tag{9}$$

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

この最適化により、残差と相互情報量が最小化されますが、ほとんどの場合、残差の最大絶対値の大きさと相互情報量の大きさはかなり異なっています。多目的最適化の計算をそのまま行うと，どちらか一方しか数値的に最小化できない可能性があります。今回の実験では、損失計算に用いる残差の最大絶対値と相互情報の最大絶対値を正規化することで、どちらも最小化できることがわかりました。 正規化とは、最適化計算の1回目の評価で算出した残差の最大絶対値と相互情報量の最大絶対値を1.0にすることです。今回の実験では、 相互情報量を正規化する必要はないことがわかりました。

3 Experiment

実験に用いるデータは人工的なものであり、共通の原因となる変数を除外することで未観測の状態を作り出す。 実験では、一般化ガウス分布(5)からμを生成し、LiNGAMを用いて因果構造を推定します。この構造から、各変数の残差と相互情報量を算出します。 これを用いて，残差と相互情報量が最小になるまでμを探索することができます． 今回の実験では、この計算に、**Probabilistic Meta-Algorithm (SA)** を使用しました。

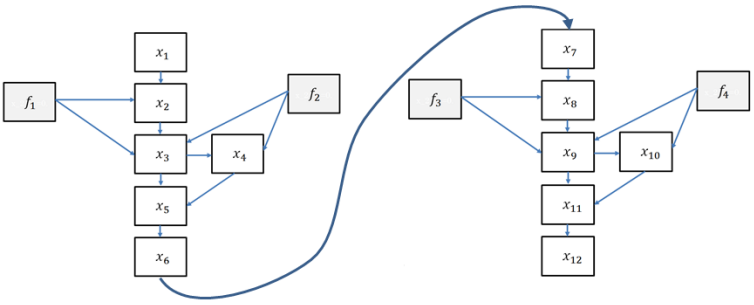

Figure 2: example causal structure

例えば、 図2に示すようなデータを19件作成しました。データ量は一律10000ラインのデータです。f_1, f_2, f_3, f_4 は交絡変数である。 これらの変数をデータに含めないことで、観察されていない共通の原因を持つデータが生成されます。

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

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

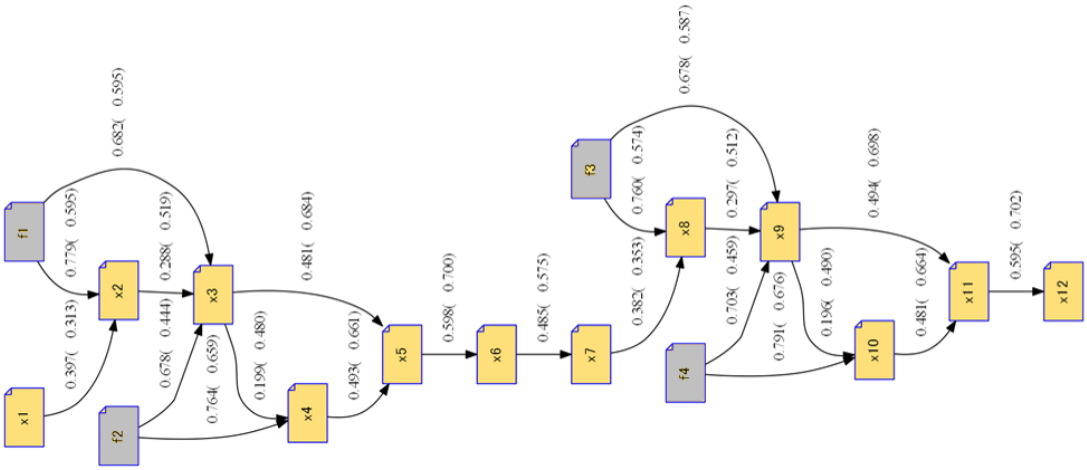

Figure 3: example

3.1 Experimental results

LiNGAMでは、$Accuracy = 70\%$で全体の30%の因果関係の方向性を間違えてしまいました(図4)。

私たちの手法では、すべての因果関係の方向性を正しく推定しました(図5)。

精度評価の値は、推定された因果構造の全エッジについて、正しい方向に推定されたエッジの数の比率で表しています。

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

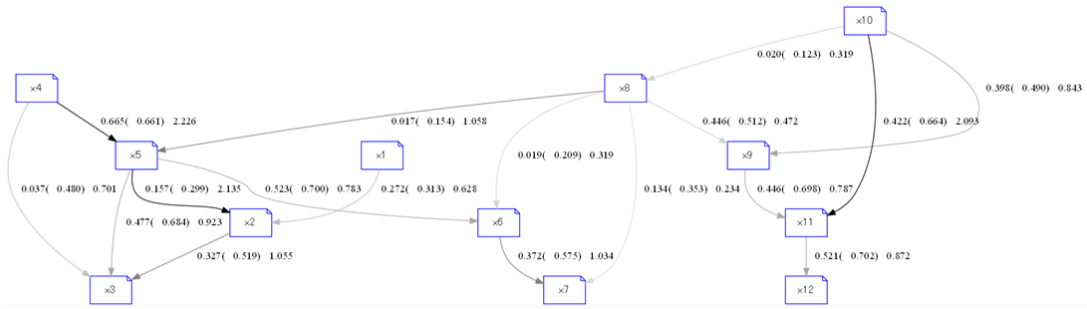

Figure 4: ICA-LiNGAM

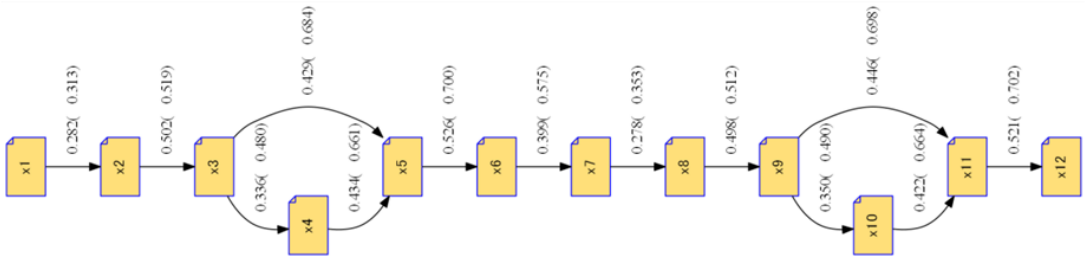

Figure 5: Our approach

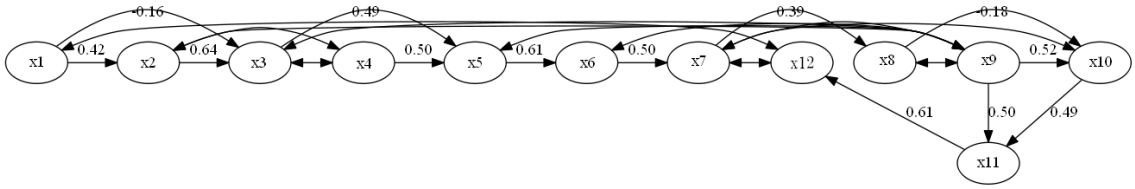

Figure 6: RCD([3])

results of the experiment for 19 cases.

Table 1: 19 cases

	ICA-LiNGAM ([5])	Our	Direct-LiNGAM([6])
Average	84%	97%	85%
Data without unobserved common cause	96%	100%	100%
Data with unobserved common cause	75%	95%	74%

未観測データがない場合でも、残差と残差の独立性を最適化により補正することで、より精度の高い線形モデルを実現しています。図7

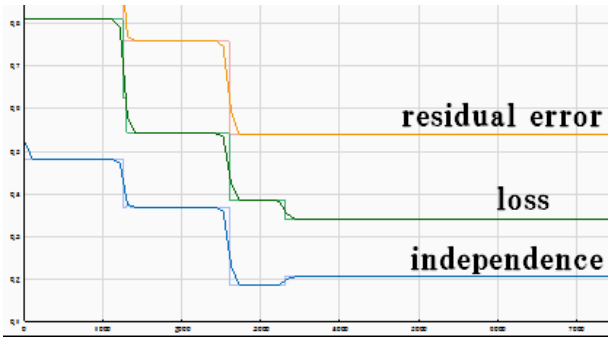

Figure 7: Optimization calculation

図7は最適化計算における損失グラフです。残差および残差間の相互情報量が下がっていく様子が見て取れます。

実験に用意した19件のデータで一つのデータで私たちの提案手法で間違った結果を推定してしまいました。図8

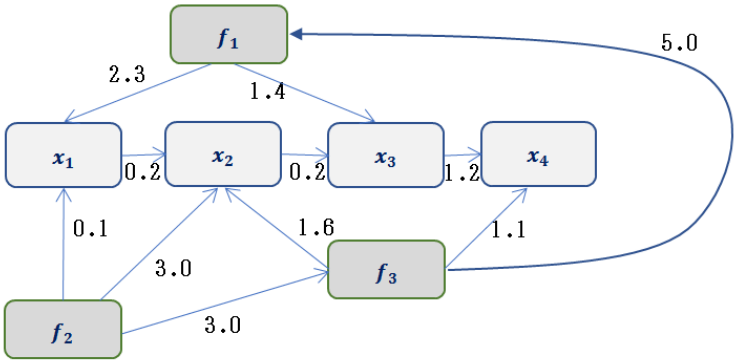

Figure 8: Example of calculating an incorrect result

$f_2 = e_{f_2}$
$f_3 = 3.0f_2 + e_{f_3}$
$f_1 = 5.0f_3 + e_{f_1}$
$x_1 = 2.3f_1 + 0.1f_2 + e_1$
$x_2 = 0.2x_1 + 3.0f_2 + 1.6f_3 + e_2$
$x_3 = 0.2x_2 + 1.4f_1 + e_3$
$x_4 = 1.2x_3 + 1.1f_3 + e_4$

このデータでは、正しい因果構造を推論することができませんでした。次の章では、先行研究で同様に未観測共通変数が在っても因果構造が推定可能なアルゴリズムRCD([3])と比較します。

3.2 Comparison with RCD([3])

私たちのアプローチは間違った結果をもたらしました。そこで、RCDを適用してみました。

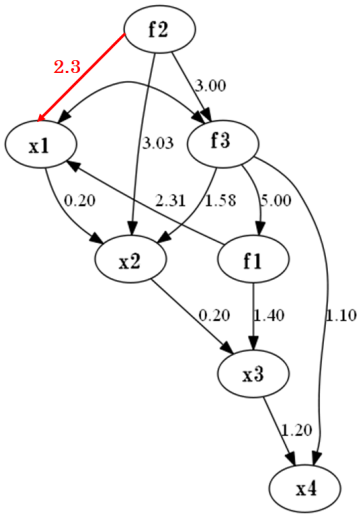

Figure 9: results for RCD

観測されていない共通原因データがない場合のRCD[2]の実験結果。

RCDは赤矢印(図9)を生成しませんでした。 x_1とf_3に双方向のパスが追加されています。これは、x_1とf_3の間に観測されていない共通原因があることを意味します。

RCDは、 x_1とf_3の間に観測されていない共通原因があると推定していますが、実際には観測されていない共通原因はありません。

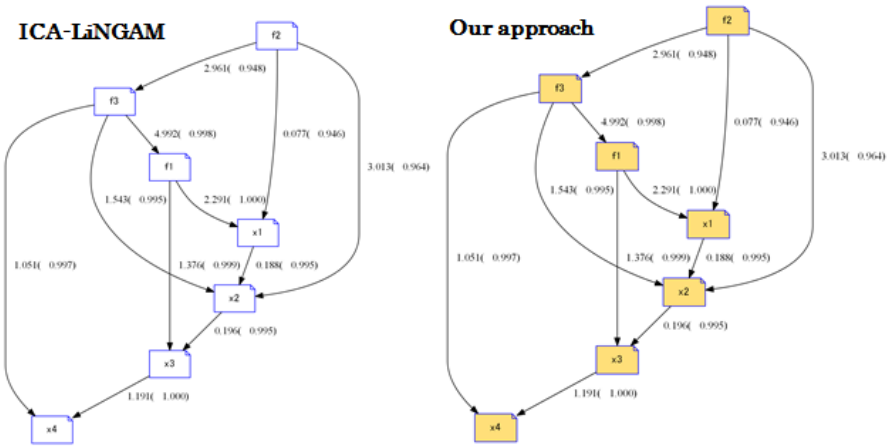

Figure 10: ICA-LiNGAM & Our approach

ICA-LiNGAMも我々の提案した手法も、 正しい結果を推定しています。 データに未観測共通変数が無いためICA-LiNGAMも正しく因果構造を推定します。 ただし，冗長な経路を削除しているため，同等の比較はできないですが，どちらも誤った経路を推定していないことがわかります。

次に未観測の共通原因データが存在する場合のRCDの実験結果を述べます。 f_1, f_2, f_3 を削除したデータです。

[2]RCDの調整パラメータはデフォルト値で検証しています

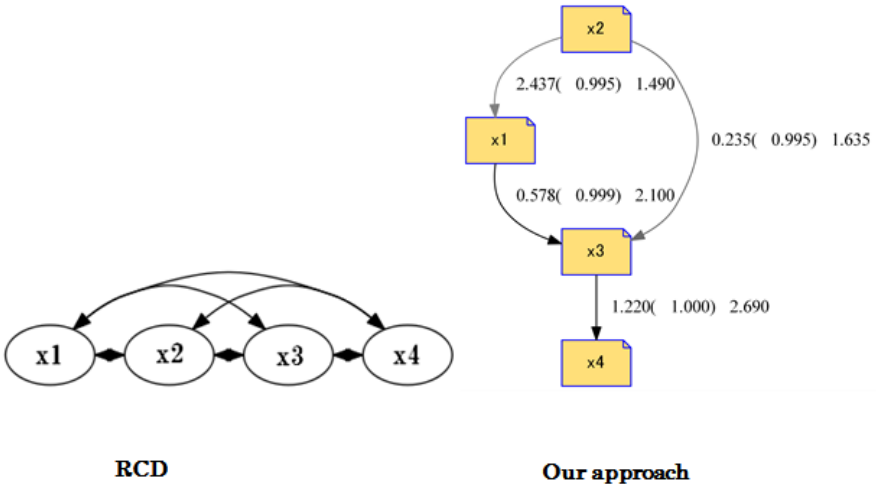

Figure 11: RCD & Our approach

我々の手法ではx_1, x_2とx_1, x_3の間で逆方向を推定しています。一方RCDでは、すべての変数の間に観測されていない共通の原因があることを推測しています。 実験では、観測されていない共通原因が他の観察された変数に比べて非常に大きい場合、誤った結果を含む可能性があることが示されました。そこで、観測されていない変数の共通原因が相対的に大きくならないようにデータを調整して実験を行ってみました。

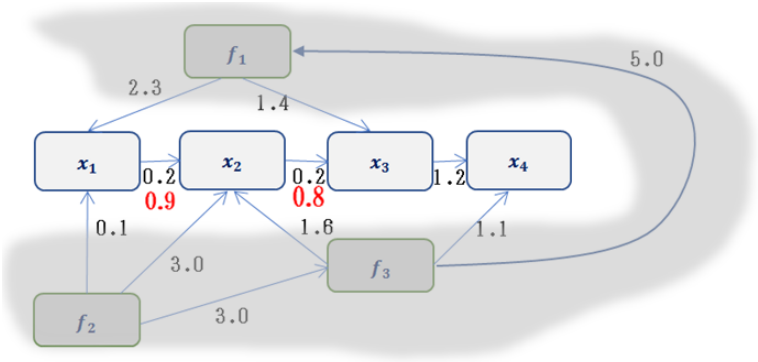

Figure 12: data

なお、因果関係グラフの構造は全く同じです。

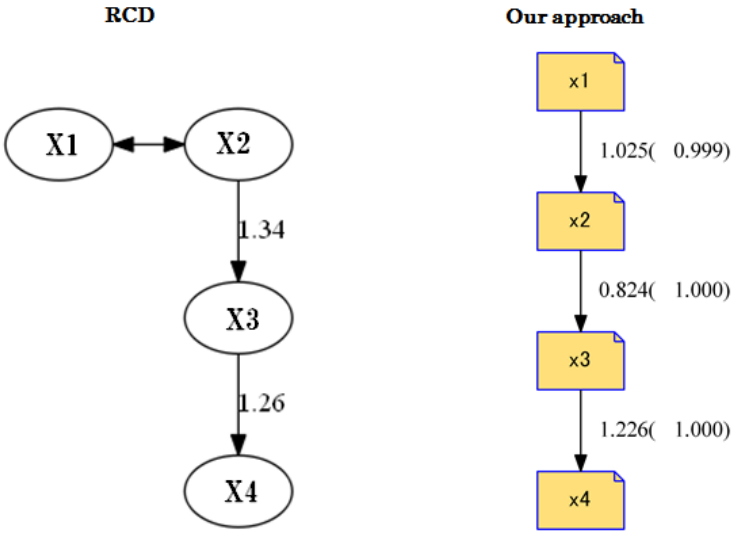

Figure 13: RCD & Our approach

RCDでは、x_1, x_2の間に未観測の共通変数があり、x_2, x_3, x_4について適切な関係を推定します。 私たちの方法では、間違ったパスはありません。また、因果関係の係数値も正しい値になっています。

4 Discussion

この手法の利点は、既存の計算技術と組み合わせることで、実用的で保守性の高いシステムを短期間で構築できることと、調整が必要なパラメータがほとんどないことです。唯一調整が必要なパラメータは、最適化の最大反復回数で、これはユーザーが設定する必要がありますが、ほとんどの場合、10000回程度の反復で収束します。 また、いくつかの問題点も見つかりました。 まず、計算時間がかかります。特に、30個の変数と10000行のデータを使用した場合、計算を終えるまでに7時間かかることがあります。これは、GPUによる高速化と並列化をうまく利用することで克服できると考えています。 最大の課題は、観測されていない潜在的共通変数の寄与度が他の変数に比べて非常に大きい場合、誤った経路を計算してしまう可能性があることです。図8 1を参照してください。

また、LiNGAMは連続したデータを前提としているため、不連続なセンサーデータには適用できないという問題もあります。 今回の実験では人工的なデータを使用しましたが、実世界のデータにも適用して適用範囲を明確にし、可能であれば改善したいと考えています。

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 Appendix Statistical Causality

A.1 ICA-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.

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

.

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

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

$$\varepsilon = \mathbf{W}\mathbf{x}$$

.

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.

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

.

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 $D = diag(\mathbf{P}\mathbf{W}_{\mathrm{ICA}})$

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

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

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.

B Appendix 計算手順

STEP1 μ_iを生成

STEP2 観測変数をμ_iによって分布状態を変更する

$$x_i - \mu_i = \sum_{k(j)<k(i)} \check{B}_{ij}(x_j - \mu_j) + \varepsilon_i \tag{10}$$

これはICA-LiNGAMによって$\check{B}$を求める事が出来る。
未観測の共通変数を含んだ式2において交絡因子の影響をtとすると

$$x_i = \sum_{k(j)<k(i)} B_{ij}x_j + t + \varepsilon_i \tag{11}$$

式11の切片$intercept$を考慮して

$$t = intercept + \mu_i \tag{12}$$

式11は次のように書ける

$$x_i = intercept + \sum_{k(j)<k(i)} B_{ij}x_j + \mu_i + \varepsilon_i \tag{13}$$

ここで分布状態を変更した因果構造$\check{B}$による残差は

$$\varepsilon_i = x_i - \sum_{k(j)<k(i)} \check{B}_{ij}x_j + t \tag{14}$$

$$\varepsilon_i = x_i - \sum_{k(j)<k(i)} \check{B}_{ij}x_j - \mu_i - intercept \tag{15}$$

STEP3 損失を計算
損失が下がったらμ_iを修正してSTEP2へ戻る
STEP1へもどる。
これを指定されて繰り返し数分繰り返す