正規化最尤符号を用いた直接原因変数の学習

久保木優太（東京大学教養学部）
小林将理（東京大学総合文化研究科）
松島慎（東京大学総合文化研究科）

IBIS 2024
2024-11-5
ソニックシティ（埼玉）

概要

本研究では、**ターゲット変数の直接原因変数**を推定する**局所的因果探索**の問題を考える。
これは介入施策の策定にとって重要であり、全体の因果構造を推定する大域的探索に比べて、**比較的緩やかな仮定の元で効率的**に学習できる。
既存手法では仮説検定を用いるために様々な問題が生じるが、
我々は、一致性のあるNML符号を用いる2変数間の因果探索手法を拡張し、**MDL原理に基づくモデル選択の枠組み**でこの問題に取り組む。

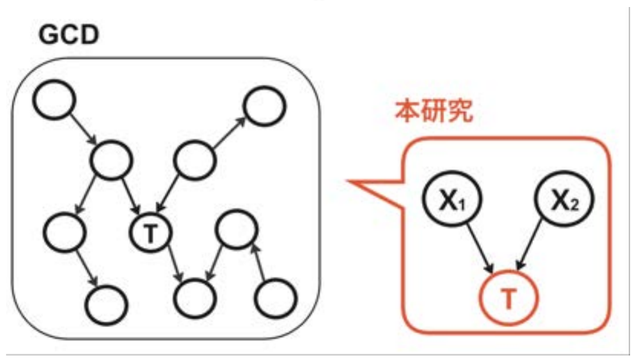
背景と問題設定

観測データのみから変数間の因果関係を推定する因果探索の問題を考える。
多変数間の因果探索には、2つのアプローチがある。

- **大域的因果探索 (GCD)** : 変数間全ての因果関係、因果グラフ全体の探索
- **局所的因果探索 (LCD)** : 因果グラフ上でターゲット変数Tの近傍（親や子）に限定した因果探索

一般にLCDは、GCDと比較して、①**計算効率が良い** ②**必要な仮定が少ない**点で優れている。

「Tの直接原因変数」＝ Pa(T) は、Tへの**介入施策**を考える上で重要な変数である。
本研究では、多変数離散データに対し、**Pa(T)のみ学習するLCDの問題**に取り組む。



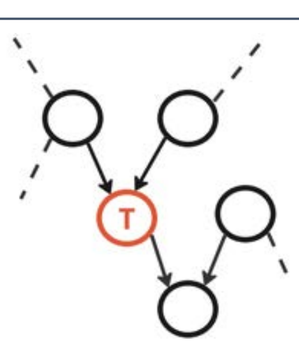
既存手法と提案手法

既存手法

LCDの研究の多くは、**条件つき独立性 (CI)** を利用して
1. MB(T)の学習
2. MB(T)内の因果探索

を行う[1,2]。
しかし、CIでPa(T), Ch(T)の**consistent**な学習はできない。

定義：
MB(T) = Pa(T)∪Ch(T) ∪ Pa(Ch(T))



Pa(T)を学習する他の手法ではすべて、**仮説検定**を必要とする[3,4]。
これらには、①**一致性がない** ②**多重検定** ③**検定の誤用** などの問題がある。

提案手法

GCDや二変数間の因果探索では、**MDL原理に基づくモデル選択**のアプローチが提案されている[5]。これは、以下のフレームワークを用いる。

1. 候補となる因果グラフGの候補集合を用意する
2. 因果グラフGが導くモデルM(G)を定義する
3. モデルM(G)のもとでのデータ列z^nの符号長$\mathcal{L}(z^n;\ M(G))$を計算する
4. $\mathcal{L}(z^n;\ M(G))$を最短にするGを真の因果グラフと推定する

本研究はこのアプローチをLCDに初めて適用し、**仮説検定を用いず/consistent**にPa(T)を学習する。
ただし、Pa(T)の候補としてPa(T)⊆Local(T)⊂X（Xは全変数）を満たすLocal(T)が与えられていることを仮定する。

より具体的には、
すべての S⊆Local(T) について、Pa(T) = Sに対応する因果グラフGとそのモデル $M_S(G)$ を定義し、$\mathcal{L}(z^n;\ M_S(G))$を計算する。
符号には、正規化最尤符号 (NML) を用いる

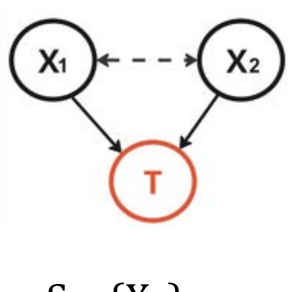
$$\mathcal{L}_{\mathrm{NML}}(z^n;\ M) = -\log \max_{P\in M} P(z^n) + \log \sum_{z'^n \in Z^n} \max_{P\in M} P(z'^n)$$

具体例

Local(T) = {X_1, X_2}の2変数である場合

各Sに対応する因果グラフ・SEMの定義とその時のz^nの符号長を示す。

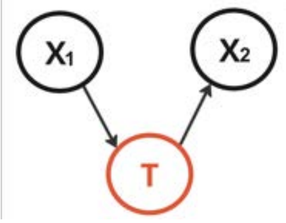
- S= {X_1, X_2}



$$\text{SEM}\begin{cases} X_1 = f_{X_1}(C, E_{X_1}) \\ X_2 = f_{X_2}(C, E_{X_2}) \\ T = f_T(X_1, X_2) + E_T \end{cases} \qquad (X_1, X_2)\sim \mathrm{Cat}(\theta_X),\ E_T\sim \mathrm{Cat}(\theta_T)$$

記述長 $\mathcal{L}(z^n;\ M_S(G)) = \mathcal{L}_{\mathrm{NML}}(z^n;\ M_S(G), f_T) + \mathcal{L}(f_T)$

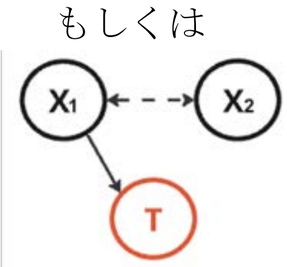
- S= {X_1}



$$\text{SEM}\begin{cases} X_1 = E_{X_1} \\ T = f_T(X_1) + E_T \\ X_2 = f_{X_2}(T) + E_{X_2} \end{cases} \qquad E_{X_1}\sim \mathrm{Cat}(\theta_{X_1}),\ E_T\sim \mathrm{Cat}(\theta_T),\ E_{X_2}\sim \mathrm{Cat}(\theta_{X_2})$$

記述長
$\mathcal{L}(z^n;\ M_S(G)) = \mathcal{L}_{\mathrm{NML}}(z^n;\ M_S(G), f_T, f_{X_2}) + \mathcal{L}(f_T) + \mathcal{L}(f_{X_2})$

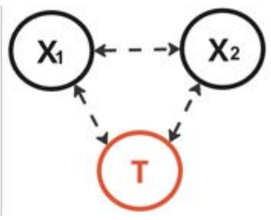
もしくは



$$\text{SEM}\begin{cases} X_1 = f_{X_1}(C, E_{X_1}) \\ X_2 = f_{X_2}(C, E_{X_2}) \\ T = f_T(X_1) + E_T \end{cases} \qquad (X_1, X_2)\sim \mathrm{Cat}(\theta_X),\ E_T\sim \mathrm{Cat}(\theta_T)$$

記述長 $\mathcal{L}(z^n;\ M_S(G)) = \mathcal{L}_{\mathrm{NML}}(z^n;\ M_S(G), f_T) + \mathcal{L}(f_T)$

- S=∅ or other cases(他のモデルで想定していないケースを全て含む)



$$\text{SEM}\begin{cases} X_1 = f_{X_1}(C, E_{X_1}) \\ X_2 = f_{X_2}(C, E_{X_2}) \\ T = f_T(C, E_T) \end{cases} \qquad (X_1, X_2, T)\sim \mathrm{Cat}(\theta)$$

記述長 $\mathcal{L}(z^n;\ M(G)) = \mathcal{L}_{\mathrm{NML}}(z^n;\ M(G))$

実験

人工データ実験

上の例のモデル通りデータ生成を行い、モデル選択の一致性を検証

n	S= {X_1, X_2}	S= {X_1} or {X_2}	S=∅ or other cases
10^3	0.84	0.97	0.74
10^4	0.99	1.00	0.96
10^5	1.00	1.00	0.95

参考文献
[1] T. Gao and Q. Ji. Local causal discovery of direct causes and effects. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc., 2015.
[2] M. Azadkia, A. Taeb, and P. Bühlmann. A fast non-parametric approach fo local causal structure learning, 2022.
[3] J. Bodik and V. Chavez-Demoulin.Structural restrictions in local causal discovery:identifying direct causes of a target variable, 2024.
[4] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals, 2015.
[5] M. Kobayashi, K. Miyaguchi, and S. Matsushima. Detection of unobserved common cause in discrete data based on the mdl principle. In 2022 IEEE International Conference on Big Data (BigData), pages 45–54, 2022.

より詳しくは↓