

MDL原理に基づく離散変数間の未観測共通原因の検知

※ IEEE International Conference on Big Data 2022に採択



2022-11-20 つくば国際会議場

小林将理 (東京大学情報理工学系研究科)
宮口航平 (IBM東京基礎研究所)
松島慎 (東京大学総合文化研究科)

概要

本研究では離散2変数間の因果関係を、直接の因果関係の2通りに加えて、変数が共通の原因を持つ場合と、変数が独立の場合を含む4つのカテゴリに分類する。

2つの離散変数の同時分布から直接因果の2方向を推定する手法はいくつか提案されているが、上記4つのカテゴリを含めた因果関係を推論する方法は限られている。

本研究ではMDL原理(最小記述長原理)に基づくNML符号長(正規化最尤符号長)を用いて、未観測共通原因の存在を許す離散変数に対する手法を初めて提案する。

計算機実験では提案手法が未観測変数を検知できることと、既存手法に比べて高い推定精度を示すことを実証した。

Reichenbach 問題

定義

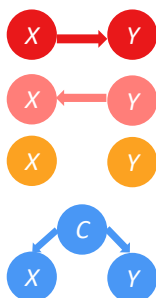
① $X \rightarrow Y$: X は Y の原因である

② $X \leftarrow Y$: Y は X の原因である

③ $X \perp\!\!\!\perp Y$: X と Y は独立である

④ $X \leftarrow C \rightarrow Y$:

X と Y の未観測共通原因 C が存在する



⇒離散2変数 X, Y の同時分布、独立な n 個のサンプルから X, Y の4通りの因果関係を推定する。

研究の動機

同時分布からの因果推論では、因果関係を表すモデルに仮定を課して表現力を制限することで因果探索が可能になる[1]。

しかし、仮定の妥当性を担保する必要がある。

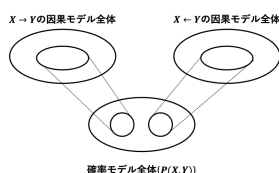
⇒ X, Y と C の関係を仮定を置かずに C を検知したい。

先行研究の限界

先行研究では

① $X \rightarrow Y$ と② $X \leftarrow Y$ の2通りの関係しか推定できないのでReichenbach問題を直接解くのは難しい。

■ 原因1: ④ $X \leftarrow C \rightarrow Y$ のSEMも仮定する必要あり



■ 原因2: ④ $X \leftarrow C \rightarrow Y$ の符号長が分からない

$$\mathcal{K}(P(X)) + \mathcal{K}(P(Y|X)) < \mathcal{K}(P(Y)) + \mathcal{K}(P(X|Y))$$

提案手法

MDL原理に基づいて、データ $z^n = (x^n, y^n) \in \{0, 1, \dots, m_X - 1\}^n \times \{0, 1, \dots, m_Y - 1\}^n$ の記述長 $\mathcal{L}(z^n; M)$ が最小となるモデル $M \in \{M_{X \rightarrow Y}, M_{X \leftarrow Y}, M_{X \perp\!\!\!\perp Y}, M_{X \leftarrow C \rightarrow Y}\}$ を推定する:

$$\hat{M}(z^n) = \underset{M \in \mathcal{M}}{\operatorname{argmin}} \mathcal{L}(z^n; M)$$

NML符号長:

$$SC(z^n, M) = -\max_{P \in \mathcal{M}} \log P(z^n) + \log C_n(M)$$

因果モデル	仮定するSEM	記述長 $\mathcal{L}(z^n; M)$
① $X \rightarrow Y$	$X = E_X$ $Y = f(X) + E_Y \pmod{m_Y}$	$SC(z^n, M_{X \rightarrow Y}^f)$ $+ \log(m_Y^{m_X-1} - 1)$
② $X \leftarrow Y$	$X = g(Y) + E_X \pmod{m_X}$ $Y = E_Y$	$SC(z^n, M_{X \leftarrow Y}^g)$ $+ \log(m_X^{m_Y-1} - 1)$
③ $X \perp\!\!\!\perp Y$	$X = E_X$ $Y = E_Y$	$SC(z^n, M_{X \perp\!\!\!\perp Y})$
④ $X \leftarrow C \rightarrow Y$	$X = f(C, E_X)$ $Y = g(C, E_Y)$	$SC(z^n, M_{X \leftarrow C \rightarrow Y})$

実験



■ 実験1: 人工データ
Reichenbach問題でテストした。

■ 実験2: 人工データ
提案手法と先行研究[2, 3, 4, 5]の $X \rightarrow Y$ の推定精度を比較した。

■ 実験3: 実世界のデータ

RESULTS ON ABALONE DATASET.					
Ground Truth	Ours(2)	Ours(4)	ACID	CISC	DR
sex → length	✓	✓	✓	✓	✓
	($\Delta = 487.54$)	($\Delta = 384.99$)			
sex → diameter	✓	✓	✓	✓	✓
	($\Delta = 467.44$)	($\Delta = 221.57$)			
sex → height	✓	✓	✓	✓	✓
	($\Delta = 659.03$)	($\Delta = 147.11$)			
RESULTS ON NLSCHOOL DATASET.					
Truth	Ours(2)	Ours(4)	ACID	CISC	DR
$X \leftarrow Y$	✓	✓	✓	✓	✓
	($\Delta = 214.89$)	($\Delta = 88.55$)			
RESULTS ON HORSE COLIC DATASET.					
Truth	Ours(2)	Ours(4)	ACID	CISC	DR
$X \rightarrow Y$	✓	✓	✓	✓	✓
	($\Delta = 19.21$)	($\Delta = 0.33$)			

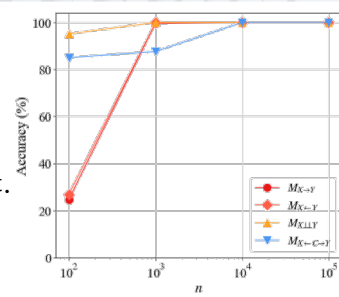


Fig1: 実験1の結果

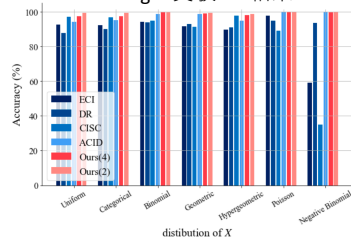


Fig2: 実験2の結果

参考文献

- [1] J. Peters, D. Janzing and B. Schölkopf: "Elements of causal inference: foundations and learning algorithms", The MIT Press (2017).
- [2] J. Peters, D. Janzing and B. Schölkopf: "Identifying cause and effect on discrete data using additive noise models", Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 597–604.
- [3] M. Kocaoglu, A. G. Dimakis, S. Vishwanath and B. Hassibi: "Entropic causal inference", Thirty-First AAAI Conference on Artificial Intelligence (2017).
- [4] K. Budhathoki and J. Vreeken: "MDL for causal inference on discrete data", 2017 IEEE International Conference on Data Mining (ICDM), pp. 751–756 (2017).
- [5] K. Budhathoki and J. Vreeken: "Accurate causal inference on discrete data", 2018 IEEE International Conference on Data Mining (ICDM), pp. 881–886 (2018).