FY2022/2022年度

Causal Inference Team
Shohei Shimizu
因果推論チーム　清水 昌平

RIKEN Center for Advanced Intelligence Project

統計的因果探索とは

- データを用いて因果グラフを推測するための方法論

仮定

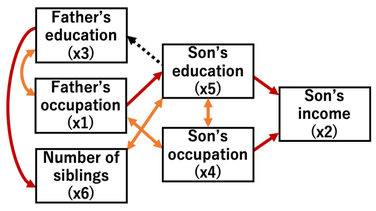
- 関数形
- 分布
- 未観測共通原因の有無
- 非巡回 or 巡回 など

データ

x1	x2	x3	x4	x5	x6
87.9	45433	17	76.3	17	1
87.9	55071	16	86	18	2
62.1	113159	16	87.9	16	0
78.5	30289	16	30.1	14	4
32.3	113159	20	63.5	20	7
60.6	55071	17	83.7	17	1
76.4	55071	16	78	14	2
63.5	37173	12	63.2	16	3
63.2	113159	14	86.5	17	1
36.5	37173	12	83.7	12	4

推測

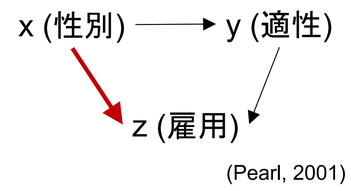
因果グラフ



Maeda and Shimizu (2020)

AIのための因果推論でも要（かなめ）

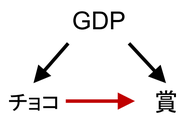
- 公平性 (Kusner et al., 2017)
- 説明性
  - 原因の確率 (Galhotra et al., 2021)
  - 予測メカニズム解析 (Blobaum et al., 2017; Sani et al., 2020)
  - 個体レベルの最適介入 (Kiritoshi et al., 2021)
- 転移学習 (Zhang et al., 2013; Zhang et al., 2020; Bareinboim et al., 2016)
- 科学的知識の取り込み (Teshima et al., 2021)
- さまざまな因果に関するクエリー(介入効果等) に答えられるかを判定するために因果グラフが必要



(Pearl, 2001)

統計的因果推論では因果グラフが要(かなめ)

- データから介入効果を推定
  - チョコ消費量を変えると
    ノーベル賞受賞者の数はどのくらい増えるのか(減るのか)
  - 機械学習のする予測
    - チョコ消費がこのくらいならノーベル賞数このくらい？
    - ノーベル賞数がこのくらいならチョコ消費このくらい？
- 介入効果を「正しく」推定するには
  因果グラフが必要 (e.g., バックドア基準)



E(賞 | do(チョコ))=E_{GDP} [E(賞 | チョコ, GDP)]

Messerli, (2012), New England Journal of Medicine

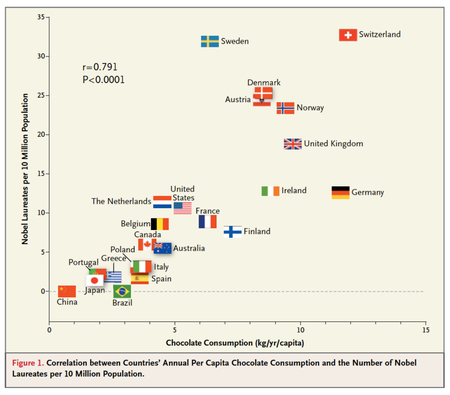


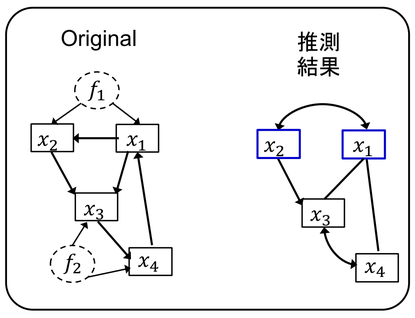
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

ノーベル賞受賞者の数

チョコレート消費量

研究内容

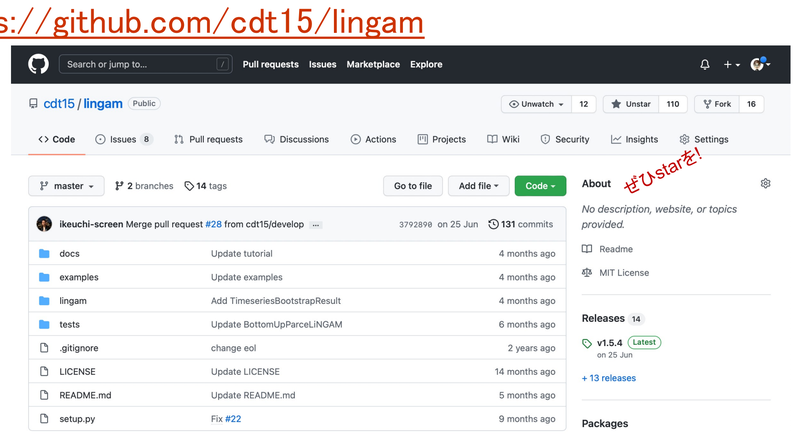
- 因果探索の研究開発
  - 未観測共通原因を許す
  - 離散と連続が混在
  - 非線形
  - 巡回
- データからわかる限界の理論的解明
- 領域知識とデータから、その時点で可能な最大限の因果グラフを分析者に提供



S. Shimizu. Statistical Causal Discovery: LiNGAM approach. SpringerBriefs in Statistics, Springer Japan, 2022.

LiNGAM Python package

- https://github.com/cdt15/lingam



T. Ikeuchi, M. Ide, Y. Zeng, T. N. Maeda, and S. Shimizu. Python package for causal discovery based on LiNGAM. Journal of Machine Learning Research, 24(14): 1-8, 2023.

- 有向道や有向辺のブートストラップ確率
- 例えば、閾値0.05を越えるものを解釈

	from	to	effect	probability
0	x3	x0	3.006190	1.00
1	x0	x1	3.004868	1.00
2	x2	x1	2.092102	1.00
3	x3	x1	20.931938	1.00
4	x0	x5	3.982892	1.00
5	x3	x5	12.024250	1.00
6	x2	x4	-0.887620	1.00
7	x3	x4	18.077244	1.00
8	x0	x4	7.993145	0.98
9	x3	x2	5.970163	0.96
10	x5	x1	0.011708	0.79
11	x2	x5	0.024284	0.72

総合効果:
20.9

x3
⋮
x0
⋮
x1

	path	effect	probability
0	[3, 0, 1]	8.644765	0.99
1	[3, 2, 1]	11.653517	0.96
2	[3, 1]	0.102936	0.10
3	[3, 2, 0, 1]	1.007787	0.08
4	[3, 0, 5, 1]	0.889629	0.05
5	[3, 4, 0, 1]	8.419805	0.02
6	[3, 2, 4, 0, 1]	-3.802326	0.01

99%
x3
x0
x1

96%
x3
x2
x1

10%
x3
x1

https://lingam.readthedocs.io/en/latest/tutorial/bootstrap.html

モデル仮定の評価

- 分析前
  - Gaussianity test
  - ヒストグラム
    - 連続変数？
  - 多重共線性
  - 領域知識
- 分析後
  - 誤差(残差)の独立性評価
    - 例えば、HSIC (Gretton et al., 2005)
  - マルコフバウンダリーによる予測の良さで評価
    (Biza et al., 2020) (実装準備中)
  - 複数のデータセットでの結果を比較
  - 領域知識による評価

連続変数と離散変数が混在する場合

Definition 1 (Linear Mixed causal model, LiM) *If a causal model for mixed data satisfies assumptions A1-A3, then this SCM is called a Linear Mixed causal model, abbreviated as LiM.*

因果グラフが識別可能

Let $\mathcal{F} = \{f^{con}, f^{dis} | f^{con}(x,e) = bx + e + c, f^{dis}(x,e) = \begin{cases} 1, & bx+e+c>0 \\ 0, & \text{otherwise} \end{cases}\}$ be a set of two functions which work on continuous and discrete variables, respectively, where x, e, b and c are a random variable of X, error term, coefficient and intercept, respectively. $\mathcal{P} = \{P^{con}, P^{dis}\}$ denotes the set of probabilistic distributions for continuous and discrete variables. Using these notations, our model can be rewritten as:

$$x_i = f_i(x_{\mathrm{pa}_i}, e_i), \quad e_i \sim P(e_i), \tag{3}$$

where $f_i \in \mathcal{F}$, and $P(e_i) \in \mathcal{P}$.

A1. Observed variables x_i $(i = 1, \ldots, p)$ form a Directed Acyclic Graph (DAG).

A2. The value assigned to each continuous variable x_i is a linear function of its parent variables denoted by $x_{\mathrm{pa}(i)}$ plus a non-Gaussian error term e_i, that is,

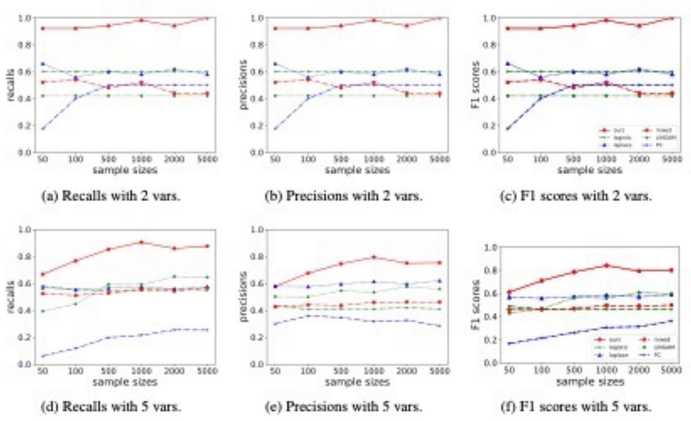
$$x_i = e_i + c_i + \sum_{j \in \mathrm{pa}(i)} b_{ij} x_j, \quad e_i \sim Non - Gaussian(\cdot), \tag{1}$$

where the error terms e_i are continuous random variables with non-Gaussian densities, and the error variables e_i are independent of each other. The coefficients b_{ij} and intercepts c_i are constants.

A3. For each discrete variable x_i, its value equals 1 if the linear function of its parent variables $x_{\mathrm{pa}(i)}$ plus a Logistic error term e_i is larger than 0, otherwise, its value equals 0. That is,

$$x_i = \begin{cases} 1, & e_i + c_i + \sum_{j \in \mathrm{pa}(i)} b_{ij} x_j > 0 \\ 0, & \text{otherwise} \end{cases}, \quad e_i \sim Logistic(0,1), \tag{2}$$

where the error terms e_i are identical to those in Eq.(1), but follow the Logistic distribution.



Y. Zeng, S. Shimizu, H. Matsui, F. Sun. Causal Discovery for Linear Mixed Data. In Proc. the First Conference on Causal Learning and Reasoning (CLeaR2022), pages 994-1009, Eureka, CA,

今後: 仮定を緩めていく