周期的な非線形関係の因果関係発見

加法性正規ノイズを含む周期的非線形構造方程式モデル

長瀬麻理雄、加納裕

大阪大学大学院基礎工学研究科数理科学専攻

1　はじめに

因果関係の特定は、多くの科学分野で大きな影響を与える。例えば、細胞内シグナル伝達の因果関係を知ることは新薬の開発に、消費者ネットワークを知ることはマーケティング戦略の改善に、政策の背後にある因果関係のメカニズムを知ることはより良い政治的意思決定に、などなどである。このような因果構造を学習するために、過去数十年の間にいくつかの方法が提案され、開発されてきた。本提案を紹介する前に、加法性雑音モデルを用いた次のような簡単な設定のもとで、こうした手法とその問題点を整理しておくことにする。XとYを関心のある観測変数とし、次の3種類の関係を考えることができる:

$$\text{(a) } Y = f(X) + e_Y\text{、}\quad \text{(b) } X = g(Y) + e_X\text{、}\quad \text{(c) } Y = f(X) + e_Y\text{、}\ X = g(Y) + e_X$$

ここで、fとgは2つの変数をつなぐ関数、e_Y と e_X は加法性誤差項である。モデル(a)はモデル(b)と区別できないことはよく知られているが、fとgがともに線形関数で、誤差項が正規分布していると仮定すれば、モデル(c)も識別できない[1].線形性と正規性を仮定した場合、Directed Acyclic Graph (DAG)でも多くの因果構造で識別可能性が成立しにくいことが多くの著者によって指摘されている [2]．周期的な因果構造は、より識別が困難である。

2　モデル

Let (y, x) be random variables involved in the model. A cyclic structural equation model (SEM) has the form of

$$\begin{aligned} y &= f(x) + e_Y \\ x &= g(y) + e_X \end{aligned} \quad \rightarrow \quad \begin{aligned} e_Y &= y - f(x) \\ e_X &= x - g(y) \end{aligned} \qquad \text{ここで、}|f'(x)g'(y)| < 1\text{である。} \tag{1}$$

条件$|f'(x)g'(y)| < 1$が必要なのは、モデルが確率変数の集合を想定しているためです。(y, x)は、対応する再帰システムにおいて、平衡状態において得られる $y_{i+1} = f(x_i) + e_Y$ $x_{i+1} = g(y_i) + e_X$ で、ここで (y_i, x_i) は i) 誤差項 (e_Y, e_X) が時間的に変化しない、 ii) $\lim_{i\to\infty} y_i < \infty$, and $\lim_{i\to\infty} x_i < \infty$, and iii) $\lim_{i\to\infty} (y_i, x_i)$ が (e_Y, e_X) が与えられたときに一意に特定されると仮定して、 *i)* のシーケンス点です。式(1)に関連するヤコビアン行列は、以下の通りである。

$$\begin{pmatrix} \frac{\partial y}{\partial e_Y} & \frac{\partial y}{\partial e_X} \\ \frac{\partial x}{\partial e_Y} & \frac{\partial x}{\partial e_X} \end{pmatrix} = \begin{pmatrix} \frac{\partial e_Y}{\partial y} & \frac{\partial e_Y}{\partial x} \\ \frac{\partial e_X}{\partial y} & \frac{\partial e_X}{\partial x} \end{pmatrix}^{-1} = \frac{1}{1 - f'(x)g'(y)} \begin{pmatrix} 1 & f'(x) \\ g'(y) & 1 \end{pmatrix}\text{、}$$

となり、ヤコビアンは$1 - f'(x)g'(y) := J(x, y) \neq 0$と与えられる。$|f'(x)g'(y)| < 1$なので、反転性 $(y, x) - (e_Y, e_X)$は保証される。(1)の逆関数を $(e_Y\ e_X)^T = \mathbf{e} = \mathbf{e}(y, x)$と書きます。誤差項$\mathbf{e}$が確率密度関数

ションを$p(e_Y, e_X; \boldsymbol{\Sigma}) = $とする。 $\frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp -\frac{1}{2}\mathbf{e}^T \boldsymbol{\Sigma}^{-1} \mathbf{e}$, また、(y, x)のpdfは、$q(y, x; f, g, \boldsymbol{\Sigma}) = $として得られる。

W($\mathbf{0}, \mathbf{\Sigma}$)を持つ多変量ノルマル分布W($\mathbf{0}, \Sigma$)に従うとする。

$$\frac{|J(y,x)|}{2\pi\sqrt{|\Sigma|}}\exp\left(-\frac{1}{2}\mathbf{e}(y,x)^T\mathbf{\Sigma}^{-1}\mathbf{e}(y,x)\right).$$

3　モデルの識別可能性

$(f_1(x), g_1(y), \mathbf{\Sigma}_1)$ と $(f_2(x), g_2(y), \mathbf{\Sigma}_2)$ の2組の関数と環状SEMの分散・共分散行列を（1）のようにするとする：

$$\begin{cases} y = f_1(x) + e_Y \\ x = g_1(y) + e_X \end{cases}, \quad \mathbf{e} = (e_Y, e_X)^T \sim N(\mathbf{0}, \mathbf{\Sigma}_1)、\quad \begin{cases} y = f_2(x) + e^*_Y \\ x = g_2(y) + e^*_X \end{cases}, \quad \mathbf{e}^* = (e^*_Y, e^*_X)^T \sim N(\mathbf{0}, \mathbf{\Sigma}_2).$$

写像（y, x） '→ （e_Y, e_X）と（y, x） '→ （$e^* Y, e^* X$）は1対1で反転可能である、すなわち、1対1であると仮定する、

| モデルの識別可能性｜$f^1 j(x) g^j 1(y)$｜<0、｜$f^2 j(x) g^j 2(y)$｜<0、

$$q(y, x; f_1, g_1, \mathbf{\Sigma}_1) = q(y, x; f_2, g_2, \mathbf{\Sigma}_2)\ \forall (y, x) \Longrightarrow 。\qquad f_1 = f_2,\ g_1 = g_2,\ \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2、$$

関数fとgが有限多項式関数であると仮定して、講演の中で証明されます。

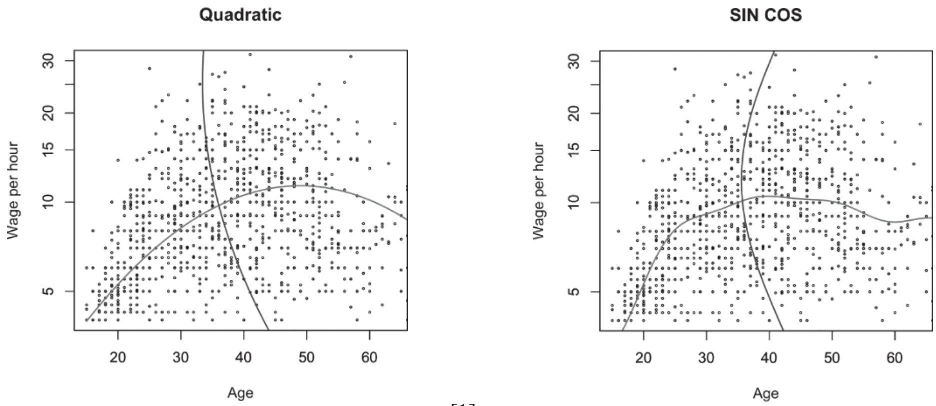
4 シミュレーション・スタディ

合成データセットを用いた以下の設定で、因果構造を完全に復元できる事実を示す。

- 多項式関数：
$$\begin{cases} y = a_0 + a_1 x + a_2 x^2 + \cdots + a_p x^p + e_Y \\ x = b_0 + b_1 y + b_2 y^2 + \cdots + b_q y^q + e_X \end{cases},\quad \boldsymbol{e} \sim W(\mathbf{0}, \mathbf{\Sigma})$$

- sin関数とcos関数：
$$\begin{cases} y = a_0 + \sum_{i=1}^{p} a_{2i-1}\sin(ix) + a_{2i}\cos(ix) + e_Y \\ x = b_0 + \sum_{j=1}^{q} b_{2j-1}\sin(jy) + b_{2j}\cos(jy) + e_X \end{cases}\quad \boldsymbol{e} \sim W(\mathbf{0}, \mathbf{\Sigma})$$

そして、それらのモデルは、些細な因果構造を持つ実世界のデータセットに適用される。例えば、米国国勢調査局が実施した1994年と1995年の人口動態調査から抽出した国勢調査データを重み付けしたデータセット「Census Income (KDD) dataset」に対してモデルを適用したところ[3]、図1に示すように、年齢は賃金に影響し、40歳と50歳前後で最も賃金が高く、賃金には年齢の影響がほとんどないという、極めて妥当な因果構造を示す結果が得られています。



[1] [2]

二次モデル、$n = 1,000$　　Sin-Cosモデル、$n = 1,000$

図1：赤線と青線はそれぞれ推定された関数fとgを表す

5 まとめと考察

最も驚いたのは、一見因果構造を持たないように見えるデータセットから、因果構造が見事に復元・発見されたことであろう。sin-cos関数を用いたモデルは、複雑さパラメータpや*qの値が*大きくなるにつれて多様な関数の形状に対応でき、複雑さパラメータをAICなどの基準で指定できるため、より実用的な場面で応用できるかもしれません。また、このモデルは、未観測の交絡因子の存在を表すと思われる誤差の共分散を扱うことができるようです。このモデルを適用することで、交絡因子が存在するかどうかを評価することができるかもしれません。

参考文献

[1] 長瀬正樹・加納洋一，非再帰式構造方程式モデルの同定可能性，*Statistics & Prob- ability Letters* **122**, pp.109-117 (2017)

[2] リチャードソン，T.，有向環状グラフの発見アルゴリズム，*人工知能における不確実性に関する第12回国際会議議事録*，モーガン・カウフマン・パブリッシャーズ・インク，pp.454-461（1996）．

[3] Mooij, J., Janzing, D., Peters, J., and Scholkopf, B., Regression by dependence minimization and its application to causal inference in additive noise models, *In Proceedings of the 26th annual international conference on machine learning*, pp.745-752.ACM, (2009)