2M-01

条件付き確率密度推定を用いた非線形因果探索

高橋　大輔† 　　　佐々木　博昭‡
滋賀大学大学院†　　公立はこだて未来大学‡

1 背景

統計的因果探索は観測データから変数間の因果構造を推定する枠組みの 1 つであり，定性分析などに基づいた事前知識なしでも，統計データのみからデータ駆動的に因果構造を推定することができる [1]．また，疫学や経済学，社会科学や脳科学など様々な分野で応用されている [2]．

統計的因果探索において，単純だが重要な問題は 2 変数 X, Y の因果方向の推定であり，これまでに様々な手法が提案されている [2][3][4][5]．例えば，[5] では X と Y の回帰誤差の非対称性に着目した RECI と呼ばれる因果探索手法が提案されている．具体的には，2 変数 (X, Y) と (Y, X) をそれぞれ (説明変数, 応答変数) とみなし，$Y = f(X)$ と $X = g(Y)$ という 2 つの回帰モデルを学習する．そして，学習した回帰モデルを用いて計算された 2 つの平均 2 乗誤差を比較するこで，因果方向を決定する．RECI の利点として，ある条件下で真の因果方向が推定可能であることが理論的に保証されていることが挙げられる．その他にも，平均 2 乗誤差を比較するため他手法に比べて計算が高速である点が挙げられる．しかしながら，その一方で，因果方向を決定する際に計算される平均 2 乗誤差は，$(f(X), Y)$ もしくは $(g(Y), X)$ の条件付き確率密度関数がガウス分布であることを暗に仮定しており，非ガウス分布の場合，RECI は正しく因果方向を推定できない可能性がある．

この問題点を解消するために，本研究では, ノンパラメトリック推定法を活用した 2 変数の因果探索手法を提案する．具体的には $(f(X), Y)$ と $(g(Y), X)$ の 2 つ条件付き確率密度関数をノンパラメトリックに推定し，推定された条件付き確率密度関数を比較することで因果方向を決定する．RECI と比較すると，条件付き確率密度関数に対する仮定がマイルドであるため，広範囲なデータに対して有効であることが期待されるだけでなく，提案法は RECI を包含する一般性の高い手法とみなすことができる．最後に，人工データと実データを用いた数値実験により，提案手法が RECI よりも高精度に因果方向を推定可能であることを示す．

† Daisuke Takahashi, Graduate School of Shiga University
‡ Hiroaki Sasaki, Future University Hakodate

2 条件付き確率密度推定を用いた因果探索

2.1 問題設定

n 組のデータ $\{(x_i, y_i)\}_{i=1}^n$ が同時確率密度関数 $p(x, y)$ より生成されているとする．このとき，真の因果方向が $X \to Y$ であれば $p(y|x)p(x)$，その一方で，因果方向が $Y \to X$ であれば $p(x|y)p(y)$ と同時確率密度関数を条件付き確率密度関数と周辺確率密度関数へ分解できる．2 変数の因果探索の目的は，データ $\{(x_i, y_i)\}_{i=1}^n$ からその因果方向を推定することである．

2.2 提案法のアルゴリズム

提案法のアルゴリズムは以下のような 3 ステップ (S1-3) から構成される．RECI との主な違いは (S2) と (S3) である．

(S1) $\{(x_i, y_i)\}_{i=1}^n$ を用いて，（サポートベクター回帰（SVR）などで）2 つの回帰モデル $Y = f(X)$ と $X = g(Y)$ を学習し，その学習結果をそれぞれ $\hat{f}(x)$ と $\hat{g}(y)$ とする．

(S2) $\{(\hat{f}(x_i), y_i)\}_{i=1}^n$ をデータとみなし，条件付き確率密度関数 $p(y|\hat{f}(x))$ を推定し，その推定結果を $\hat{p}(y|\hat{f}(x))$ とする．同様に $\{(\hat{g}(y_i), x_i)\}_{i=1}^n$ からも条件付き確率密度関数の推定結果 $\hat{p}(x|\hat{g}(y))$ を

得る．

(S3) $\{(x_i, y_i)\}_{i=1}^n$ と (S2) の推定結果を用いて，

$$\hat{m} = \frac{\sum_{i=1}^n \log \hat{p}(y_i|\hat{f}(x_i))}{\sum_{i=1}^n \log \hat{p}(x_i|\hat{g}(y_i))} \tag{1}$$

を計算する．そして，$\hat{m} \geq 1$ のとき $X \to Y$，それ以外では $Y \to X$ と因果方向を決定する．

提案法と RECI との関係を議論する．(S3) における $\hat{p}(y|\hat{f}(x))$ と $\hat{p}(x|\hat{g}(y))$ へガウス分布モデルを適用するとき，式(1)は，

$$\hat{m} = \frac{\sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + (\text{定数})}{\sum_{i=1}^n (x_i - \hat{g}(y_i))^2 + (\text{定数})} \tag{2}$$

となる．式(2)を用いて (S3) と同様に因果方向の決定するとき，提案法は RECI と本質的に一致する．したがって，提案法は RECI を特別な場合として含むより一般性の高い因果探索手法である．そして，本研究では，(S2) で Least Squares Conditional Density Estimation (LSCDE)[6] を用いる．LSCDE はノンパラメトリック推定法であるため，条件付き確率密度関数の仮定がマイルドであり，提案法は RECI よりも広範囲なデータへ有効であることが期待される．

3　数値実験

最初に人工データを用いた数値実験により提案法を RECI と比較する．数値実験で使用する 2 変数データは以下のように生成した．

$$y_i = f(x_i) + \epsilon_i \tag{3}$$

上式中で x_i は生成分布，ϵ_i は確率的ノイズでここではそれぞれ同じ確率分布から生成されるとする．また，$f(x) = \sin(2\pi x) + \frac{1}{8}\exp(1-x)$ である．(3)での真の因果方向は $X \to Y$ であり，合計で $n = 1000$ 組のデータ (x_i, y_i) を生成し，それら因果方向の推定精度の平均値を評価指標とした．

数値実験結果を表 1に示す．表では，ノイズの確率分布がガウス分布，ラプラス分布，指数分布における RECI と提案法の因果方向の推定精度をそれぞれ示しており，回帰モデルの推定では RECI と提案法の両方において最小二乗法による 2 次の多項式と 3 層ニューラルネットワーク（NN），SVR（ガウスカーネル）を用いた．ノイズがガウス分布に従うとき，RECI の推定精度が提案法よりも高いことが分かる．一方，ノイズがラプラス分布と指数分布といった非ガウス分布に従うとき，提案法の推定精度が高い傾向にあり，特に NN などの非線形性が高い回帰関数において有効であることが確認される．

表1　人工データ (RECI/提案法) における平均推定精度

	多項式	SVR	NN
ガウス分布	**0.774**/0.621	**0.558**/0.279	**0.563**/0.384
ラプラス分布	**0.953**/0.479	0.268/**0.579**	0.111/**0.616**
指数分布	0.447/**0.747**	0.105/**0.753**	0.1026/**0.7368**

表2　実データによる推定精度

	多項式	SVR	NN
RECI	0.661	0.489	0.461
提案手法	**0.73**	**0.728**	**0.729**

最後に，実データセット [7] を用いた数値実験を行った．ここでは，データセットを訓練用とテスト用にランダムに分割し，複数回数値実験を行った．その推定精度の平均値が表 2に示されている．(S1) に使用された 3 つの回帰方法すべてで提案法が RECI よりも推定精度が高いことが分かる．

4　結論

本研究では 2 変数の非線形因果探索手法を提案した．既存手法と比較して，ノンパラメトリック法を用いて条件付き確率密度関数を推定しているため，提案法は広範囲なデータに対して有効であることが期待される．実際，人工データと実データを用いた数値実験によりその有効性を確認した．

参考文献

[1] 大久保将貴．因果推論の道具箱．2019.
[2] 清水昌平．講談社, 2017.
[3] P. Hoyer et al. *Curran Associates, Inc.*, 2008.
[4] D. Janzing et al. *Artificial Intelligence*, 2012.
[5] P. Blöbaum et al. *PeerJ Computer Science*, 2019.
[6] Sugiyama et al. *IEICE Transactions on Information and Systems*, 2010.
[7] J. M. Mooij et al. *Journal of Machine Learning Research*, 2016.