F-004

共非線形性尺度 NNR-GL を用いた変数集合発見とその実験的評価

Variable Set Discovery Using Co-nonlinearity Measure NNR-GL and Its Empirical Evaluation

大崎 美穂 * 田儀 樹 * 佐々木 捷人 *
岸本 真弥 * 片桐 滋 * 大西 圭 †
Miho Ohsaki Itsuki Tagi Hayato Sasaki
Naoya Kishimoto Shigeru Katagiri Kei Ohnishi

1. はじめに

変数間の関係を知ることは現象の解明や予測に不可欠であり，科学や工学に求められる．現象に関するドメイン知識の質・量と想定する関係によって問題解決策は異なる．原因と結果 (因果関係) を記述する因果推論は，ドメイン知識に基づき対象の変数と因果構造を絞り込んだ上で因果の強さを求める．因果探索は変数間の因果構造を発見する [1]．これらの因果関係分析は十分なドメイン知識を前提とする．もし未知の側面が多く変数も多種多様な現象に切り込むならば，因果関係分析の前段階として従属関係分析が必要であろう．

対象にすべき変数の選定と，それらの因果関係候補である従属関係を見出す尺度には，相関係数 (CC)，距離相関係数 (DCC)，ヒルベルト・シュミット独立基準 (HSIC)，最大情報係数 (MIC)，グループラッソニューラルネット回帰尺度 (NNR-GL) などがある [2, 3]．これらは従属関係を検出するが，従属関係にある変数の集合や因果関係のヒント (原因になり得る代表的な変数) までは導出できない．そこで我々は過去に，尺度が検出した従属関係から変数集合・代表変数を求める機能 IA を開発した．そして，多変数・非線形な従属関係 (共非線形性) を検出できる NNR-GL と IA の組合せ手法 (本稿では NNR-GLIA と呼ぶ) を提案し，単純な従属関係を持つデータに対する有効性を示した [4, 5]．本研究では，より複雑な従属関係を持つデータで実験を行い，NNR-GLIA の有効性をより明確にする．

2. 提案手法

提案手法 NNR-GLIA は NNR-GL による共非線形性の検出を繰り返し，IA で検出結果を集約する枠組みを持つ (図 1 参照) [4, 5]．NNR-GL は，入力層の各変数に L1 正則化を施すグループラッソ (GL) を組み込んだニューラルネットワーク回帰 (NNR) と，回帰性能と変数の重みを統合して尺度化する機能から成る．NNR-GL の NNR は複数の入力変数から 1 つの出力変数の回帰を行う，言い換えると，入出力変数間の非線形従属関係をモデル化する．NNR にはどのようなニューラルネットワークも可であるが，各入力変数の回帰への貢献度を明らかにするため，入力層のみに GL を含む必要がある (図 1 のグレーの箇所)．

式 (1) は NNR-GL の損失関数である．第 1 項は出力変数 z_j を推定する回帰の性能を高め，第 2 項は GL の効果を生む．具体的には，第 2 項は回帰への貢献度に応じて各入力変数の重み $\mathbf{w}_m^{(1)}$ を 0，あるいは大きな絶対値に二極化する．訓練・検証・試験を経て，NNR-GL は入出力変数間の共非線形性の強さの情報として，回帰性能を表す残差平方和 (RSS) と入力変数の貢献度を表す重み (RC) を導出する．最後に，RSS から求まる決定係数と RC の相乗平均を求め，共非線形性の強さの尺度値として出力する [3]．

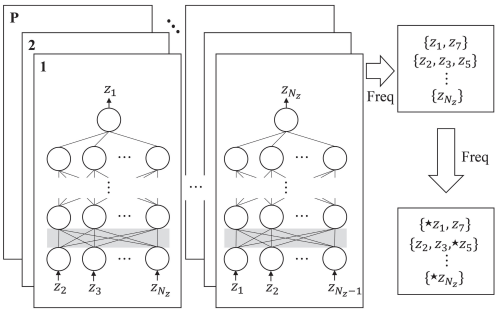

図 1: NNR-GL と IA から成る NNR-GLIA の構成図.

$$J_j(\boldsymbol{W}) = \frac{1}{N_t}\sum_{s=1}^{N_t}(z_{js} - \hat{z}_{js}(\boldsymbol{W}))^2 + \lambda \sum_{m=1,\ \neq j}^{N_z} ||\mathbf{w}_m^{(1)}||_2 \quad (1)$$

NNR-GLIA は NNR-GL の出力変数と初期乱数を変えながら，NNR-GL を繰り返し実行する．図 1 に示すように，$z_1, z_2, \cdots, z_{N_z}$ の 1 つずつを出力変数とした回帰を行い，その回帰も異なる初期乱数で複数回行う．検出結果として，非線形従属関係を持つ変数集合の候補が多数得られる．IA は閾値処理・集合集約を候補に適用して，共非線形変数集合・代表を導出する．まずは候補ごとに変数 z の出現頻度 Freq(z) を計算し，閾値 T_{Freq} を超えれば共非線形変数集合のメンバに採用する．この過程を式 (2)，(3) に従って 2 段階で行う．以上により，頻繁に共非線形性を示す変数から成る集合が得られる．さらに NNR-GL の回帰モデルの凸性に基づき，非線形従属関係の源流に位置するであろう代表変数も求める [4, 5].

$$NCS^{(j)} = \{z \mid z \in \bigcup_{i=1}^{P} CS_i^{(j)},\ T_{\mathrm{Freq1}} \leq \mathrm{Freq}(z)\} \quad (2)$$

$$FS^{(j)} = \{z \mid z \in \bigcup_{i=1}^{Q} NCS_i^{(j)},\ T_{\mathrm{Freq2}} \leq \mathrm{Freq}(z)\} \quad (3)$$

3. 評価実験

3.1. 目的と条件

本実験では複数の従属関係が混在するデータを合成し，これに NNR-GLIA と，従来の従属関係尺度と IA

*同志社大学 Doshisha University
†九州工業大学 Kyushu Institute of Technology

表 1: 5 種類の従属関係を含む人工データの条件.

原因の変数	結果の変数
$x_1 \in [0, 1]$	$x_2 = \mathrm{Line}(x_1) = x_1$
$x_3 \in [0, 10]$	$x_4 = \mathrm{Exp}(x_3) = 10^{x_3}$
$x_5 \in [-0.5, 0.5]$	$x_6 = \mathrm{Parab}(x_5) = 4x_5^2$
$x_7 \in [-1.3, 1.1]$	$x_8 = \mathrm{Cubic}(x_7) = 4x_7^3 + x_7^2 - 4x_7$
$x_9 \in [0, 1]$	$x_{10} = \mathrm{Sine}(x_9) = sin(8\pi x_9)$
$x_{11} \in [0, 1]$	なし
正解の集合と代表. 代表には $\star$ 付与.	$\{\star x_1, \star x_2\}$, $\{\star x_3, x_4\}$, $\{\star x_5, x_6\}$ $\{\star x_7, x_8\}$, $\{\star x_9, x_{10}\}$, $\{\star x_{11}\}$

の組合せ (CC-IA, DCC-IA, HSIC-IA, MIC-IA) を適用する.「従属関係にある変数の集合と代表変数を導出できるのか」という観点で性能比較して，NNR-GLIA の有効性を検討する．なお，IA による代表変数の導出には NNR-GL の回帰モデルが必要なので，比較対象の手法では代表変数は求まらない.

データ生成において従属関係は因果から生じたと仮定し，表 1 の 5 つの従属関係を用いる．分析対象の変数 $z_1, z_2, \cdots, z_{N_z}$ は，原因の変数 x_j $(j = 1, 3, 5, 7, 9)$ と，結果の変数 x_j $(j = 2, 4, 6, 8, 10)$，独立な変数 x_{11} とする．ゆえに正解の集合は 6 個である．Line は線形な従属関係のため，x_2 を原因，x_1 を結果にしても同じデータとなる．そこで Line のみ x_1 と x_2 の両方を原因，言い換えると代表変数と見なす．一様乱数で生成した原因の変数の値を従属関係式に代入し，複数の点を作る．これらに平均 0，標準偏差 SD のガウスノイズを加えて疑似的な観測データとする．SD は変数定義域の 0, 1, 5, 10 [%]，データサイズは 150, 300, 1500, 3000 (計 16 条件) とする.

NNR-GL の NNR には多層パーセプトロンを，HSIC にはガウスカーネルを採用する．検証で設定するハイパーパラメータとして，NNR-GL は層数，ニューロン数，学習率，エポック数を持つ．HSIC はカーネル幅，MIC は分割数と領域幅を持つ. 訓練で設定するパラメータは，NNR-GL は NNR の辺の重み，HSIC はカーネルの重み，MIC は変数空間の分割パターンである．そこでデータを 3 分割して訓練・検証・試験に用い，NNR-GL, HSIC，MIC の設定と検出を行う．設定不要な CC と DCC では試験のみとする．IA の尺度値の閾値は，手法ごとに最良の試験性能となる値にする.

3.2. 結果と考察

データサイズやノイズの SD が異なる 16 条件間で結果は類似していたので，抜粋して掲載する．表 2 は，3000 点 (訓練・検証・試験に各々1000 点使用) のデータに，SD が 1 [%] のノイズを加えた条件の結果である．提案手法 NNR-GLIA は正解の集合 6 個全てを発見し，かつ，集合の代表変数 (原因の変数) も正しく同定した．一方，線形性を仮定する CC-IA や非線形ながら表現能力に制限がある DCC-IA では，発見した正解は 2 個と少ない．HSIC-IA は非線形性の表現能力が高いが，今回は正解 2 個となった．MIC-IA も表現能力が高く，こちらは正解 5 個を発見した．本実験の結果

表 2: データ 3000 点，ノイズの SD 1 [%] の実験結果．正解と一致した集合は太字にした.

手法	発見した集合と代表
NNR-GLIA	**{**$\star x_1, \star x_2$**}** ,**{**$\star x_3, x_4$**}** ,**{**$\star x_5, x_6$**}** , **{**$\star x_7, x_8$**}** ,**{**$\star x_9, x_{10}$**}** ,**{**$\star x_{11}$**}**
CC-IA	**{**x_1, x_2**}** , {x_3}, {x_4}, {x_5}, {x_6}, {x_7}, {x_8}, {x_9}, {x_{10}}, **{**x_{11}**}**
DCC-IA	**{**x_1, x_2**}** , {x_3}, {x_4}, {x_5}, {x_6}, {x_7}, {x_8}, {x_9}, {x_{10}}, **{**x_{11}**}**
HSIC-IA	**{**x_1, x_2**}** , {x_3}, {x_4}, {x_5}, {x_6}, {x_7}, {x_8}, {x_9}, {x_{10}}, **{**x_{11}**}**
MIC-IA	**{**x_1, x_2**}** , {x_3}, {x_4}, **{**x_5**,** x_6**}** , **{**x_7**,** x_8**}** , **{**x_9**,** x_{10}**}** , **{**x_{11}**}**

から，他の尺度を用いた場合に比べて NNR-GLIA は共非線形変数集合の発見性能が高く，代表変数の発見も可能と言える.

4. おわりに

変数間の従属関係は因果関係の手がかり，ひいては現象の解明・予測につながるため，広い分野に役立つと考えられる．本研究では，多変数間の非線形従属関係 (共非線形性) を検出する NNR-GL と検出結果を集約する IA を用いて，共非線形変数集合・代表を発見する NNR-GLIA に着目した．NNR-GLIA，および，他の従属関係尺度に IA を組み合わせた手法を複数の非線形従属関係を含む人工データに適用し，比較評価した．その結果，NNR-GLIA は他の手法よりも正確に従属関係を持つ変数の集合を発見した．従属関係を生み出した代表変数も正確に同定できた．今後は，より複雑な従属関係を持つ人工データや，現実の問題の実測データで評価実験を行う予定である.

謝辞

本研究は JSPS 科研費 21K12018 の助成を受けた.

参考文献

[1] C. Glymour et al., “Review of Causal Discovery Methods Based on Graphical Models”, Frontiers in Genetics, vol.10, article 524 (2019).

[2] D. Lopez-Paz et al., “The Randomized Dependence Coefficient”, Int’l Conf. on Neural Information Processing Systems, vol.1, pp.1–9 (2013).

[3] M. Ohsaki et al., “NNR-GL: A Measure to Detect Co-Nonlinearity Based on Neural Network Regression Regularized by Group Lasso”, IEEE Access, vol.9, pp.132033–132052 (2021).

[4] M. Ohsaki et al., “Discovery of Sets and Representatives of Variables in Co-Nonlinear Relationships by Neural Network Regression and Group Lasso”, IEEE Int’l Conf. on Bioinformatics and Biomedicine, 10.1109BIBM.2018.8621207 (2018).

[5] M. Ohsaki et al., “Evaluation of the Neural-network-based Method to Discover Sets and Representatives of Nonlinearly Dependent Variables,” IEEE Int’l Conf. on Cybernetics, 10.1109CYBCONF51991.2021.9464151 (2021).