Vol.2022-MPS-138 No.57
Vol.2022-BIO-70 No.57
2022/6/29

代謝ネットワーク解析に向けた予測誤差の状態空間を用いる時系列因果推論法の提案

大山 鷹志 [1]　遠里 由佳子 [1,a)]

概要：本研究は，生体内の代謝物などを計測して得られる非線形な時系列から，制御因子を推定することを目指し，新しい因果推論法の提案を目的とする．非線形でノイズが強い 2 つの時系列に対する因果推論の従来法に Non-Parametric Multiplicative Regression Granger 因果性テスト（NPMR）がある．NPMR は各時系列を埋め込んで作成した状態空間間で予測時系列を作成する．そこで，NPMR の状態空間で自己回帰により求めた予測時系列と元時系列との誤差から，さらに状態空間を作成し推論を行う手法を提案した．因果関係の強弱を調整できる結合ロジスティックマップで生成した短時系列（*N*=25）に対し，提案手法の推論精度が 71.0%となり NPMR を上回ることを確認した．

キーワード：因果推論, Granger 因果, 非線形, 時系列解析, Convergent Cross Mapping

1. はじめに

本研究では，大腸菌の代謝と増殖に関わる実験で得た時系列から，代謝を制御する上で重要な因子や未知の経路予測を予測するため，因果関係のネットワークを構築することを目指している．短時系列に適した因果推論法の提案し，人工的に生成した時系列を用いて，時系列の因果推論に必要な充分な長さと精度の時系列を判断する．

時系列から因果関係を推定する従来法として，入力に線形な時系列を仮定する Granger 因果性テスト[1]や，非線形な時系列を仮定する Convergent Cross Mapping (CCM) [2]がある．Granger 因果性テストは，1 つの時系列による自己回帰モデルの予測時系列と，2 つの時系列による予測時系列によって因果推論を行う．Sriyudthsak らは 乳酸菌のグルコース代謝の解糖系の主要な反応のシミュレーションで得た時系列に対し Granger 因果テストに基づき因果推論を行った[3]．一方 CCM は生態学の分野で得られる時系列に多くの実績がある．Ma らは CCM の拡張にあたる Cross Mapping Smoothness (CMS) を提案し，大腸菌や酵母の転写制御ネットワークのシミュレーションで得た時系列に適用した[4]．

本研究では，Granger 因果性テストの自己回帰の計算を，ノイズの多い非線形データに対して推論できるように，Non-Parametric Multiplicative Regression に置き換える提案した手法(以下, NPMR)に着目した[5]．そして NPMR に，自己回帰モデルにより求めた予測時系列と元時系列との誤差を用いた予測を追加した時系列因果推論の手法を提案する．

以下 2 章では先行研究である NPMR と CCM について説明する．3 章では提案手法を説明する．4 章では，結合ロジスティックマップで生成した時系列を用いて，提案手法と NPMR、CCM の精度比較を行う．5 章ではまとめと今後の課題について述べる．

1 立命館大・情報理工
Ritsumeikan University
a) yukako@fc.ritsumei.ac.jp

2. 先行研究

実験で用いる結合ロジスティックマップで生成した時系列を例に，従来法で共通した時系列の状態空間への埋め込みと予測の基本的な手続きを説明し，NPMR と CCM の因果推論の手続きの詳細を述べる．

2.1 時系列の状態空間への埋め込み

式(1)の結合ロジスティックマップで生成した非線形な 2 本の時系列を例に，時系列の状態空間への埋め込み法を説明する．

$$\begin{aligned} X_t &= X_{t-1}\left(\alpha_x - \alpha_x X_{t-1} - \beta_{xy} Y_{t-1}\right) \\ Y_t &= Y_{t-1}\left(\alpha_y - \alpha_y Y_{t-1} - \beta_{yx} X_{t-1}\right) \end{aligned} \tag{1}$$

ここで$\alpha_x, \alpha_y, \beta_{xy}, \beta_{yx} \in \mathbb{R}$は結合定数で，$\alpha_x, \alpha_y$は自身への影響率，$\beta_{xy}, \beta_{yx}$は他方への影響率にあたる．そこで，原因$X$と結果$Y$ ($X \rightarrow Y$) となるよう，β_{xy}とβ_{yx}を設定し，時系列XとYを生成した場合を考える(図 1)．正規化した時系列X, Yから，埋め込み次元Eと遅延時間τをパラメータとして状態空間$\boldsymbol{M_X}$と$\boldsymbol{M_Y}$を式(2)で作成する(図 2)．

$$\begin{aligned} \boldsymbol{X_t} &= \left[X_t, X_{t-1}, \cdots, X_{t-(E-1)\tau}\right]^{\mathrm{T}} \\ \boldsymbol{Y_t} &= \left[Y_t, Y_{t-1}, \cdots, Y_{t-(E-1)\tau}\right]^{\mathrm{T}} \end{aligned} \tag{2}$$

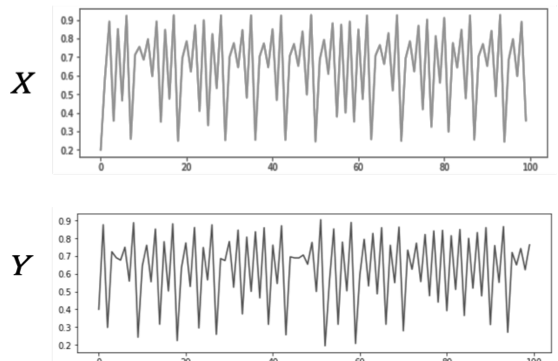

図 1　結合ロジスティックマップで得られる時系列 X, Y
($\alpha_x = 3.6, \alpha_y = 3.7, \beta_{xy} = 0, \beta_{yx} = 0.3, X_0 = 0.2, Y_0 = 0.4$)

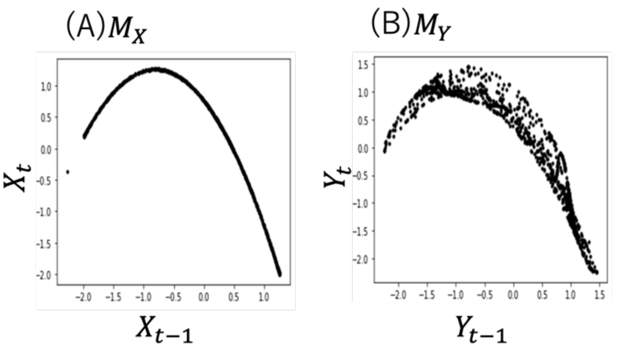

図 2　状態空間(A)$\boldsymbol{M_X}$と(B)$\boldsymbol{M_Y}$ ($E=2,\ \tau=1$)

埋め込み次元Eと時間遅れτの決定が因果推論の精度に大きく影響することが知られ，その決定方法も提案されているが，本論文では，結合ロジスティックマップで生成された時系列のみを使うため，最も基本的な条件$E=2,\ \tau=1$を用いる．

2.2 NPMR Granger 因果性テスト

NPMR における$\boldsymbol{M_X}\rightarrow\boldsymbol{M_Y}$の類推は，$Y_t$や$X_t$に対する Leave-one-out で以下のステップ 1-2 を繰り返したのち，ステップ 3 を行う．

1. 状態空間$\boldsymbol{M_Y}$で予測点$\hat{Y}_t$を，自己回帰にあたるガウスカーネルの重み$w_{i,j}$を元に求める

$$\hat{Y}_t=\frac{\sum_{i=1\ i\neq t}^{L-(E-1)\tau}W_iY_i}{\sum_{i=1\ i\neq t}^{L-(E-1)\tau}W_i},\tag{3}$$

$$W_i=\prod_{j=1}^{E}w_{i,j},\qquad w_{i,j}=\exp\left(-0.5\left[\frac{Y_{i,j}-Y_{t,j}}{\sigma_{Y_j}}\right]^2\right).\tag{4}$$

2. 予測点$\hat{Y}_t$から，状態空間$\boldsymbol{M_X}$でX_tと$X_{i\backslash t}$のそれぞれの次元$j\in E$におけるガウスカーネルに基づく重み$v_{i,j}$を用いて，$\hat{Y}_t^X$を求める

$$\hat{Y}_t^X=\hat{Y}_t+\frac{\sum_{i=1\ i\neq t}^{L-(E-1)\tau}v_{i,j}X_i}{\sum_{i=1\ i\neq t}^{L-(E-1)\tau}v_{i,j}},\tag{5}$$

$$v_{i,j}=\exp\left(-0.5\left[\frac{X_{i,j}-X_{t,j}}{\sigma_{X_j}}\right]^2\right).\tag{6}$$

3. 予測時系列$\hat{Y}$とY，$\hat{Y}^X$とYの誤差の分散から式(7)の定義を用いて評価する

$$G=\log\frac{\sigma_{Y-\hat{Y}}}{\sigma_{Y-\hat{Y}^X}}\tag{7}$$

$G\geq 0$となった場合，$X\rightarrow Y$が成立したとみなす．

2.3 Convergent Cross Mapping

CCM における$\boldsymbol{M_X}\rightarrow\boldsymbol{M_Y}$の類推は，$X_t$に対する Leave-one-out で以下のステップ 1-2 を繰り返し，ステップ 3 を行う．

1. 状態空間$\boldsymbol{M_X}$で$\boldsymbol{X_t}$から$E+1$個の近傍点$\boldsymbol{X_{t_1}},\boldsymbol{X_{t_2}},\ldots,\boldsymbol{X_{t_{E+1}}}$をユークリッド距離で決定する．

$$d\left(\boldsymbol{X_t},\boldsymbol{X_{t_n}}\right)=\sqrt{(X_{t-1}-X_{t_n-1})^2+\left(X_t-X_{t_n}\right)^2}.\tag{8}$$

2. $\boldsymbol{X_{t_1}},\boldsymbol{X_{t_2}},\ldots,\boldsymbol{X_{t_{E+1}}}$と時間的に対応する状態空間$\boldsymbol{M_Y}$上の$Y_{t_1},Y_{t_2},\ldots,Y_{t_{E+1}}$を用いて，$Y_t$の予測点$\hat{Y}_t$を求める．

$$\hat{Y}_t=\sum_{i+1}^{E+1}w_{t_i}Y_{t_i},\tag{9}$$

$$w_{t_i}=\frac{u_{t_i}}{\sum_{j=1}^{E+1}u_{t_j}},u_{t_i}=\exp\left(-\frac{d\left(\boldsymbol{X_t},\boldsymbol{X_{t_n}}\right)}{d\left(\boldsymbol{X_t},\boldsymbol{X_{t_1}}\right)}\right).\tag{10}$$

3. すべての時刻tで予測時系列$\hat{Y}$を求め，時系列Yとの相関係数ρを求める．

$\boldsymbol{M_Y}\rightarrow\boldsymbol{M_X}$の類推が正の相関で，$\boldsymbol{M_X}\rightarrow\boldsymbol{M_Y}$の類推より高い相関である場合に，$X\rightarrow Y$が成立するとみなすことを基本とする．擬似相関を見分けるたるため，状態空間の埋め込みに用いる時系列長Lで複数のサンプルを抽出し，状態空間の解像度が上がる場合に，$\hat{X}$とXの平均相関係数$\bar{\rho}$が向上し，やがて飽和するとき，$X\rightarrow Y$のみの因果関係があるとみなす[6]. CCM は少なくとも 30[2]，理想的には約 1,000 の時系列長を解析に要する[7].

2.4 短時系列やノイズを含む時系列に対する先行研究

CCM を短時系列に適した形に拡張した先行研究として，Multispectral CCM [7] や CMS [5] がある. Multispectral CCM は同一のシステムに由来する複数の短時系列を 1 つの時系列として扱うことで CCM に必要な時系列長を確保している．本研究では，複数の短時系列を前提としないため，比較対象としない．一方 CMS は CCM の類推に動径基底関数 (RBF: Radial Basis Function) ネットワークを導入している．しかし，類推における時点の対応関係を意味する「滑らかさ」という指標を定義することを主題としているため比較対象としない．

3. 提案手法

提案手法における$\boldsymbol{M_X}\rightarrow\boldsymbol{M_Y}$の類推は，$Y_t$に対する Leave-one-out で，以下のステップ 1-3 を繰り返したのち，ステップ 4 を行う．ステップ 1 は NPMR のステップ 1 と同じであり，状態空間$\boldsymbol{M_Y}$から自己回帰によって予測時系列を作成する．その後，ステップ 2-3 で元時系列との予測誤差の状態空間$\boldsymbol{M_\varepsilon}$を作成し，状態空間$\boldsymbol{M_\varepsilon}$を元に$Y$を予測する．

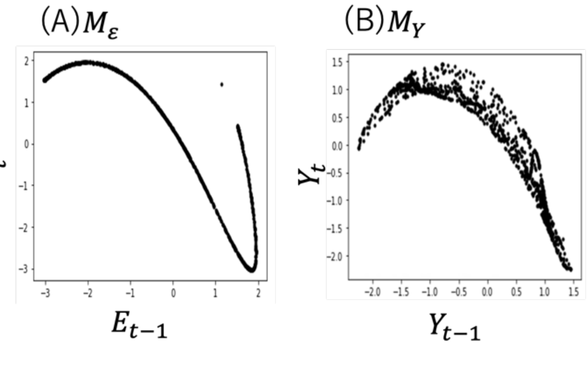

図 3　状態空間(A)$\boldsymbol{M_\varepsilon}$と(B)$\boldsymbol{M_Y}$（$E=2,\ \tau=1$）

1. 状態空間$\boldsymbol{M_Y}$で自己回帰にあたる予測点$\hat{Y}_t$を，ガウスカーネルに基づく重み$w_{i,j}$を元に求める

$$\hat{Y}_t = \frac{\sum_{i=1,i\neq t}^{L-(E-1)\tau} W_i Y_i}{\sum_{i=1,i\neq t}^{L-(E-1)\tau} W_i}, \tag{11}$$

$$W_i = \prod_{j=1}^{E+1} w_{i,j}, \qquad w_{i,j} = \exp\left(-0.5\left[\frac{Y_{i,j} - Y_{t,j}}{\sigma_{Y_j}}\right]^2\right). \tag{12}$$

2. 予測データ$\hat{Y}$と元データYの誤差$E_t = Y_t - \hat{Y}_t$を求め，状態空間$\boldsymbol{M}_\varepsilon$を作成する．
3. 予測点$\hat{Y}_t$から, 状態空間$\boldsymbol{M}_\varepsilon$で$E_t$と$E_{i\setminus t}$の各次元$j \in E$におけるガウスカーネルに基づく重み$v_{i,j}$を用いて，$\hat{Y}_t^X$を求める

$$\hat{Y}_t^X = \hat{Y}_t + \frac{\sum_{i=1\ i\neq t}^{L-(E-1)\tau} V_i X_i}{\sum_{i=1\ i\neq t}^{L-(E-1)\tau} V_i}, \tag{13}$$

$$V_i = \prod_{j=1}^{E} v_{i,j}, v_{i,j} = \exp\left(-0.5\left[\frac{E_{i,j} - E_{t,j}}{\sigma_{E_j}}\right]^2\right). \tag{14}$$

4. 誤差εとY，$\hat{Y}^X$とYの誤差の分散から等分散の検定を行い，等分散でない場合に$X \rightarrow Y$が成立したとみなす．

4. 実験

因果関係がある非線形な時系列に対する推論精度が，ノイズの有無によりどのように変化するかを，提案手法とNPMR, CCM と比較し検証する．加えて，因果関係がない時系列としてサロゲートデータを作成し，誤検出の割合も確認する．

4.1 実験 1：ノイズのない時系列で時系列長別の精度比較

提案手法と従来法である NPMR と CCM の因果推論の精度を，式(1)の結合ロジスティックマップで生成した時系列を用いて比較した．ただし，他方への影響率β_{xy}とβ_{yx}が相対的な大小関係にある場合でも，因果関係が方向を推定できるかを確かめるため[5]，自身への影響率を固定し，他方の影響率を$\beta_{xy} < \beta_{yx}$となるように，β_{yx}を区間[0.3, 0.9]から，β_{xy}を区間[0.0, β_{yx}]からランダムに選択した．そして，生成した時系列に対し$X \rightarrow Y$が推論できたときに正解とした．そして，提案手法と NPMR は，$\boldsymbol{M_X} \rightarrow \boldsymbol{M_Y}$の類推が成立し，$\boldsymbol{M_Y} \rightarrow \boldsymbol{M_X}$の類推が成立しない場合のみ，$X \rightarrow Y$とみなす．代表的な時系列長 N=25, 50, 100, 1000 とし，ランダムに生成する時系列のサンプル数を各 2000 とした．

提案手法, NPMR, CCM の正解率を表 1 に示す. N=25 の短時系列では，提案手法が最も高い精度(71.0%)を達成した．NPMR は，ノイズがない時系列に対して時系列長に関わらず推論精度は低くなった．一方 CCM は，N=25 の短時系列では，正解率は 40.0%と低いが，時系列長が長くなるに従って正解率が上昇し, N=1000 で 80.5%に達した．

図 4 に時系列長を N≤25 とした場合の提案手法の正解率の推移を示す．正解率は N=20 で最も高い 71.0%となり，N=15 で 46.9%まで低下した．これより本条件で，提案手法が約 60%の正解率を維持するためには，長さ 17 以上の時系列が必要とわかる．

図 5 に提案手法と CCM の N=25 の場合の影響率と推論結果の対応を示す．提案手法 (図 3A)も CCM (図 3B)も，$\beta_{xy} = 0$とした一方向の因果性 (unidirectional causality)に対しては推論が正確である一方，双方向の因果性 (bidirectional causality)，特に双方の因果関係が同程度か，原因となる時系列が他方から大きな影響率を受けている場合に($\beta_{xy} > 0.6$)，推論を間違う傾向がある．しかし，提案手法の方が CCM より推論できる範囲が広いことが目視で確認できる．

表 1 手法ごとの正解率の比較(%)

	N=25	N=50	N=100	N=1000
提案手法	**71.0**	22.4	3.7	0.1
NPMR	2.5	0.1	0.0	0.0
CCM	40.0	**47.9**	**50.9**	**80.5**

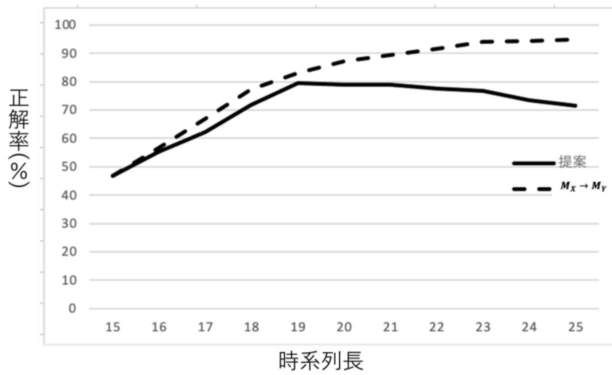

図 4 時系列長と提案手法の正解率（実線：提案手法, 破線：提案手法の$\boldsymbol{M_X} \rightarrow \boldsymbol{M_Y}$のみの類推の正解率）

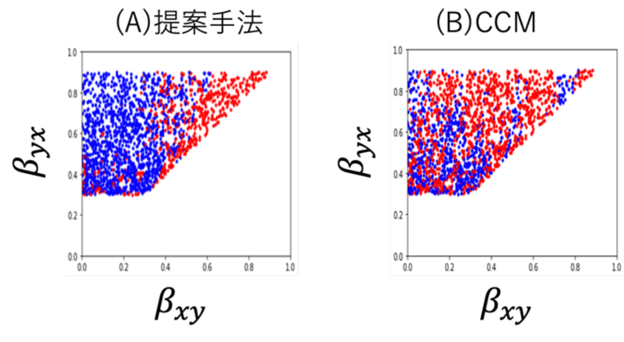

図 5 時系列長 N=25 のときの影響率の分布と推論結果 (A) 提案手法 (B) CCM（青：正解，赤：不正解）

なお，提案手法は解析に用いる時系列長が長くなると，推論精度が低下している(表 1)．その原因の一つは双方向での推論した際の因果関係の誤検出にある(図 4)．時系列長にかかわらず，$\boldsymbol{M_X} \rightarrow \boldsymbol{M_Y}$のみの類推つまり片方向の精度は 91.7%の正解率を維持しているが，時系列長が長くなると$\boldsymbol{M_Y} \rightarrow \boldsymbol{M_X}$の類推の精度が低下し，最終的な推論結果の正解率の低下につながっている．本問題に対し，時系列長の

Vol.2022-MPS-138 No.57
Vol.2022-BIO-70 No.57
2022/6/29

増加に強い改善案を検討する予定である.

4.2 実験 2：ノイズがある短時系列での精度比較

実データでは，シミュレーションデータと違い，プロセスノイズや観測ノイズが生じていることが想定される．そこで，ノイズに対する頑健性を見るため，実験 1 で用いた結合ロジスティックマップで生成した時系列に，式(15)のプロセスノイズ$\varepsilon_P \sim \mathcal{N}(0, {\sigma_P}^2)$や，式(16)の観測ノイズ$\varepsilon_O \sim \mathcal{N}(0, {\sigma_O}^2)$を加えた時系列に対する正解率の変化を確認する．

$$\begin{aligned} X_t &= X_{t-1}\left((\alpha_x + \varepsilon_P^x) - (\alpha_x + \varepsilon_P^x)X_{t-1} - \beta_{xy}Y_{t-1}\right) \\ Y_t &= Y_{t-1}\left((\alpha_y + \varepsilon_P^y) - (\alpha_y + \varepsilon_P^y)Y_{t-1} - \beta_{yx}X_{t-1}\right) \end{aligned} \tag{15}$$

$$\begin{aligned} X_t' &= X_t + \varepsilon_O^x \\ Y_t' &= Y_t + \varepsilon_O^y \end{aligned} \tag{16}$$

ただし，また，結合ロジスティック式は，自身への影響率が$1 < \alpha < 4$であるとき，カオスな性質を持つ非線形な時系列を生成するため，$1 < \alpha + \varepsilon_P < 4$となるように制御する．同様に，プロセスノイズの探索範囲は，$\sigma_P = \{0, 0.005, \cdots, 0.2\}$とする．一方，観測ノイズの探索範囲は時系列の生成に影響が少ないため，プロセスノイズよりも探索範囲を広くし $\sigma_O = \{0, 0.05, \cdots, 2\}$とする．

図 6 に提案手法と NPMR, CCM における，プロセスノイズの量と正解率の推移を示す．提案手法は常に最も高い精度を示した．ただし，プロセスノイズの量が増えると正解率は減少し，$\sigma_P = 0.2$で正解率は 43.6%となった．一方，NPMR はプロセスノイズの量が増えると正解率が上がるが，$\sigma_P = 0.2$でも正解率は 26.2%にとどまった. CCM は N=25 という条件でプロセスノイズにも弱く$\sigma_P > 0.1$で正解率は約 10%となった.

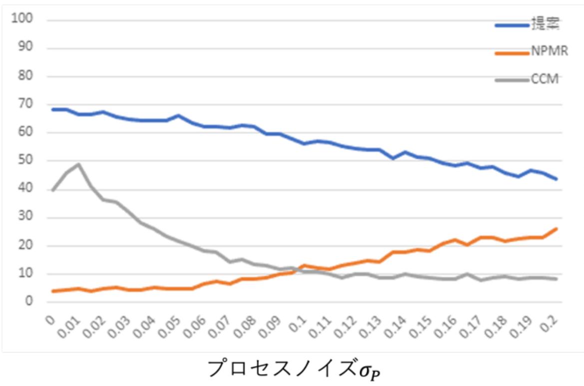

図 6 プロセスノイズと正解率の変化（N=25）

図 7 に提案手法と NPMR, CCM における，観測ノイズの量と正解率の推移を示す．提案手法は$\sigma_O < 0.4$ならば正解率が最も高く，観測ノイズが大きくなると推論精度は急激に低下する．一方 NPMR は，観測ノイズが大きくなるほど正解率が上がる．ただし，その上限は 33.2%にとどまった. CCM は N=25 という条件で観測ノイズにも弱いく，$\sigma_O >$ 0.05で正解率は 8.8%となった．正解率 60%以上を達成する上では，提案手法が最も良いと言える.

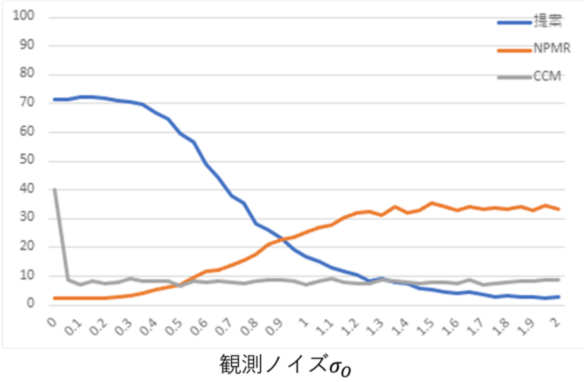

図 7 観測ノイズと正解率の変化（N=25）

以上より，提案手法はノイズが多く短い時系列の因果推論に適していることが明らかになった．ただし，観測ノイズが増えると正解率が急激に下がる可能性がある．その原因として, 2 変数の予測で 2 変数目のノイズを考慮できないことが考えられる．つまり，時系列Yのノイズの影響は，予測に必要な重みを誤差の状態空間も考慮し求める過程で弱めることができても，時系列Xのノイズの影響を弱める過程はなく，観測ノイズの量が増えると大きく精度が下がると考えている.

4.3 実験 3：サロゲートデータを用いた誤検出率の確認

得られた因果推論の正当性が因果関係のない場合に比べて有意に高いことを示すため，因果関係のない時系列に対して因果関係があると誤検出する確率を確認することが重要である．そのため，元となる時系列の平均や分散といった特徴を保存したサロゲートデータを比較に用いる手法が提案されている[8]．ここでは以下 1-3 のステップで，時系列の振幅を保持したフーリエサロゲートを用いる.

1. 時系列をフーリエ変換する

$$S_F(t) = \frac{1}{\sqrt{N}} \int x(t)\, e^{-\frac{2i\pi tn}{N}} \tag{17}$$

2. 位相S_Fを$u \sim (0, 1]$に従ってランダムに変換する

$$\hat{S}_F(t) = S_F(t) e^{2i\pi u_t} \tag{18}$$

3. $\hat{S}_F(t)$を逆フーリエ変換する

$$X_F(t) = \frac{1}{\sqrt{N}} \int \hat{S}_F(t)\, e^{-\frac{2i\pi tn}{N}} \tag{19}$$

原因Xの時系列からフーリエサロゲートデータX_Fを作成し$X_F \rightarrow Y$の推論を行い，因果関係がない時系列に対する提案手法の推論精度をみる.

図 8 に影響率の分布とサロゲートデータを用いた時の推論結果を示す．提案手法が正しく因果関係がないと推論できた確率は 89.8%であった．また，実験 1 において推論できたものかつ，因果関係がないと推論したものの結果は，90.0%であった．このことからも，実験 1 の結果のほとんど

は，因果関係があるものを推論しているといえる．

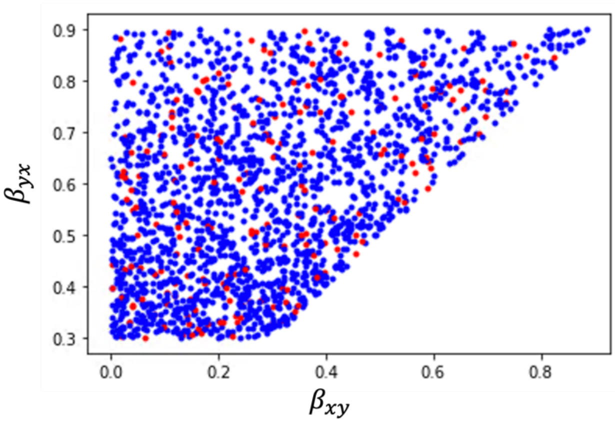

図 8　サロゲートデータにおける影響率の分布と推論結果（青：正解（因果性無し），赤：不正解（因果性有り））

5. おわりに

本発表では因果推論の手法として提案された NPMR-based Granger 因果性テストをもとに，新しい因果推論法を提案し，推論精度や誤検出の精度を比較した．結合ロジスティックマップで生成した N=25 の短時系列に対し，高い推論精度を確認した．今後の課題として，双方向の因果性が検出できる範囲の詳細な調査や，時系列長の増加に強い改善案の提案，代謝のシミュレーションデータへの応用をあげる．

謝辞　本研究は，科研費 JP19K12226 と JP20H04242 の補助による．

参考文献

[1] Granger, C. W. J., "Investigating Causal Relations by Econometric Models and Cross-spectral Methods, " *Econometrica*, 1969, vol. 37, no. 3, pp. 424–438.
[2] Sugihara, G., May, R., Ye, H., Hsieh, C. Deyle, E., Fogarty, M., and Munch, S., "Detecting Causality in Complex Ecosystems, " *Science*, 2012, vol. 338, no. 6106, pp. 496–500.
[3] Sriyudthsak, G., Shiraishi, F., and Hirai, M. Y., "Identification of a Metabolic Reaction Network from Time-Series Data of Metabolite Concentrations," PLOS One, 2013, vol. 8, no. 1, p. e51212.
[4] Ma, H., Aihara, K., and Chen, L., "Detecting Causality from Nonlinear Dynamics with Short-term Time Series," (Supplementary Information Section 2, Chapter 2), *Scientific Reports*, 2014, Vol. 4, p. 7464.
[5] Nicolaou, N., and Constandinou, T. G., "A Nonlinear Causality Estimator Based on Non-Parametric Multiplicative Regression, " *Frontiers in Neuroinformatics*, 2016, vol. 10, pp. 1–21.
[6] Clark, A. T., Ye, H., Isbell, F., Deyle, E. R., Cowles, J., Tilman, G. D., and Sugihara, G., "Spatial Convergent Cross Mapping to Detect Causal Relationships from Short Time Series," *Ecology*, 2015, vol. 96, no. 5, pp. 1174–1181
[7] 中川 新一郎，阿部 真人，岡村 寛，"Convergent Cross Mapping の紹介：生態学における時系列間の因果関係推定法"，日本生態学会誌, 2015, vol. 65, no. 3, pp. 241–253.
[8] Räth, C., and Monetti, R., "Surrogates with random Fourier Phases," In: *Topics on Chaotic Systems, Selected Papers from Chaos 2008 International Conference*, World Scientific Publishing, 2009, pp. 274–285.