Vol.2018-MPS-118 No.24
Vol.2018-BIO-54 No.24
2018/6/15

教師あり学習に基づく時系列の因果推論

近原 鷹一[1,a)]　藤野 昭典[1,b)]

概要：時系列の因果推論は時系列解析における重要なタスクの 1 つである．従来手法では，回帰モデルを用いて因果関係の方向を推定するが，その推定精度は，回帰モデルがデータにうまくフィティングできるかどうかに強く依存するため，各々のデータに対して適切な回帰モデルを選択する必要がある．しかし，回帰モデルの選択には，データに対する深い理解が要求されるため，実際には容易なことではない．本稿では，回帰モデルではなく分類器を用いた，教師あり学習に基づく手法を提案する．そのために，過去の値で条件づけられた条件付き分布間の距離を用いた特徴量表現を導入し，この特徴量表現によって，異なる因果関係の時系列に対して，十分異なる特徴ベクトルが得られ，結果として高い精度で因果関係を推定できることを実験的に示す．

キーワード：Granger causality，時系列解析，カーネル法

A Supervised Learning Approach to Causal Inference in Time Series

Yoichi Chikahara[1,a)]　Akinori Fujino[1,b)]

Abstract: Causal inference in time series is an important task in time series analysis. Traditional methods use regression models for this task. Since their inference accuracies depend largely on whether the model can be well fitted to the data, it requires us to select an appropriate regression model. However, this is not easy because such selection of regression models requires a deep understanding of the data. This paper proposes a supervised learning framework that utilizes a classifier instead of regression models. We introduce a feature representation that utilizes the distance between the conditional distributions given past variable values. We experimentally show that the feature representation gives sufficiently different feature vectors for time series with different causal relationships, which leads our method to achieve high inference accuracy.

Keywords: Granger causality, time series analysis, kernel methods

1. はじめに

時間依存する変数間の原因と結果の関係 (因果関係) を発見することは，時系列解析における重要な問題の 1 つであり，幅広い応用が考えられる．例えば，研究開発 (R&D) に対する投資額 X が総売上 Y に影響を与えるが Y は X に影響を与えないという因果関係 ($X \to Y$) は，企業における意思決定の手助けになる．また，時系列マイクロアレイデータから，遺伝子間の因果関係 (制御関係) を発見することは，バイオインフォマティクス分野における最も重要なタスクの 1 つである．

時系列の因果関係の定義として，Granger causality [5] が，幅広い分野で用いられてきた [8], [19]．これは，変数 X の過去の値が変数 Y の未来の値を予測するのに有用であれば，X は Y の原因であると定義するものである．

Granger causality を同定するために，既存手法ではベクトル自己回帰 (VAR) モデルや一般化加法モデル (GAM) などの回帰モデルを用いる．これらの手法を用いれば，Y の未来の値に関する予測誤差が，Y の過去の値のみを用いて学習した回帰モデルで得られるものより，$[X, Y]^\top$ の過去

[1] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, Kyoto 619–0237, Japan
a) chikahara.yoichi@lab.ntt.co.jp
b) fujino.akinori@lab.ntt.co.jp

Vol.2018-MPS-118 No.24
Vol.2018-BIO-54 No.24
2018/6/15

の値を用いて学習した回帰モデルで得られるもののほうが有意に小さい場合に，X が Y の原因である ($X \to Y$) と決定することができる．この際に用いる回帰モデルがデータにうまくフィッティングできるものであれば，これらの手法によって正しく Granger causality を同定することができる．しかし，実際には，個々の時系列データに対して適切な回帰モデルを選択することは難しく，データに関する深い理解が要求されるので，こうしたモデルベースの手法を用いて正しく Granger causality を同定することは一般に容易なことではない．

本研究の目的は，データの深い理解を要求しないような，時系列の因果推論手法を確立することである．そのために，本稿では，回帰モデルではなく分類器を用いた，教師あり学習に基づく因果推論のフレームワークを提案する．具体的には，Granger causality を同定する問題を，$X \to Y$, $X \leftarrow Y$, または *No Causation* を表す 3 値のクラスラベル (causal label) を個々の時系列に割り当てる分類器を学習する，3 値分類問題として解くことを提案する．実は，i.i.d. データ (独立同分布からサンプルされたデータ) を対象とした分類に基づく因果推論手法は既にいくらか提案されており，そのどれもが実験的に高い精度を達成している [2], [7], [11], [12]．時系列データを対象とした分類に基づく因果推論を実現するため，本稿では因果関係の異なる時系列に対して十分異なる分類の特徴ベクトルを与えるような特徴量表現を定式化する．この特徴量表現は，Granger causality の定義 — Y の過去の値で条件づけた Y の未来の値に関する条件付き分布と，$[X, Y]^\top$ の過去の値で条件づけた Y の未来の値に関する条件付き分布を考えるとき，2 つの条件付き分布が異なるならば，X は Y の原因である — に基づいており，これらの条件付き分布間の距離に基づいて特徴ベクトルを返すものである．分布間の距離を計算するために，カーネル平均を用いて，個々の分布を，再生核ヒルベルト空間 (RKHS) と呼ばれる特徴空間中の点として写像し，これらの点の間の距離 (maximum mean discrepancy(MMD) [6]) として分布間の距離を計算する．

比較実験を通して，回帰モデルを用いて Granger causality を同定する既存手法，分類に基づいて i.i.d. データから因果関係を推定する既存手法より，提案手法が高い推定精度を達成したことを示す．また，提案手法の有効性を示すために，提案した特徴量表現が，Granger causality の有無・方向が異なる時系列データに対し，十分異なる特徴ベクトルを返すことを実験的に示す．さらに，多変数時系列データから Granger causality を推定するために，提案手法をどのように拡張すればよいかについても言及する．

2. Granger causality

Granger causality は，変数 X の過去の値が変数 Y の未来の値を予測するのに有用であれば，X は Y の原因であると定義するものである．これは，次のように定義される:

定義 1 (Granger causality[5]) 定常過程，すなわち定常な確率変数の系列 $\{(X_t, Y_t)\}$ $(t \in \mathbb{N})$ を考える，ただし，X_t 及び Y_t は $\mathcal{X}$，$\mathcal{Y}$ 上にそれぞれ定義されるとする．ここで，S_X，S_Y をそれぞれ確率変数 $\{X_1, \cdots, X_t\}$，$\{Y_1, \cdots, Y_t\}$ の観測とする．
Granger causality は，

$$P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$$

が成立するならば，$\{X_t\}$ が $\{Y_t\}$ の原因であると定義し，

$$P(Y_{t+1}|S_X, S_Y) = P(Y_{t+1}|S_Y) \tag{1}$$

が成立するならば，$\{X_t\}$ は $\{Y_t\}$ の原因でないと定義するものである．

2 つの条件付き分布 $P(Y_{t+1}|S_X, S_Y)$, $P(Y_{t+1}|S_Y)$ が同一であるか否かを判断するために，既存手法 [1], [5], [13], [18] では，2 つの条件付き期待値 $E[Y_{t+1}|S_X, S_Y]$，$E[Y_{t+1}|S_Y]$ が等しいか否かを，統計的仮説検定に基づいて判断する (これは式 (1) が成立するか否かを判断するより，はるかに容易な問題である)．例えば，既存手法 [5] においては，これらの条件付き期待値は (V)AR モデルを用いて表され，その予測誤差に基づいて検定統計量を計算し，Granger causality を同定する．

条件付き期待値を表す際，これらの手法では，データにうまくフィッティングできるような適切な回帰モデルが必要となる．しかし，そのような回帰モデルを選択するのは実際には容易なことではない．この問題に対し，本稿では，回帰モデルの代わりに，分類器を用いた，新たなアプローチを提案する．

3. 提案手法

3.1 分類のタスク設定

訓練データが，N ペアの 2 変数時系列データ $S^1, \cdots, S^N$ から構成されるとする．ただしそれぞれの時系列 S^j は，長さが定数 T_j で表される，確率変数 $\{(X_1^j, Y_1^j), \cdots, (X_{T_j}^j, Y_{T_j}^j)\}$ $(j \in \{1, \cdots, N\})$ の観測であるとする．ここで，個々の時系列 S^j には，causal label と呼ばれるラベル $l^j \in \{+1, -1, 0\}$ が割り当てられており，これは因果関係 $X^j \to Y^j$，$X^j \leftarrow Y^j$，もしくは *No Causation* を表すものである (ただし，X^j, Y^j は，それぞれ $X^j = (X_1^j, \cdots, X_{T_j}^j)$，$Y^j = (Y_1^j, \cdots, Y_{T_j}^j)$ を表す)．

$\nu(\cdot)$ を，時系列 S^j を単一の特徴ベクトルに変換する関数とする．提案手法では，まず $\{(\nu(S^j), l^j)\}_{j=1}^N$ を用いて，分類器を学習する．すると，2 変数時系列データ S'(テストデータ) から Granger causality を推定する問題は，学習した分類器を用いて特徴ベクトル $\nu(S')$ にラベルを割り当てる問題として換言できる．

第 3.3 章にて後述するように，このような分類のタスクは，多変数時系列データに対して拡張することも可能である．

3.2 分類器の設計

個々の時系列に causal label を割り当てる分類器を構築するために，特徴量表現 $\nu(\cdot)$ を定式化する．以下では，Granger causality に依って十分異なるような特徴ベクトルを得るための本研究のアイディアについて述べる．

3.2.1 設計指針

シンプルに Granger causality の定義 (定義 1) を用いれば，例えば，X が Y の原因で Y は X の原因ではない場合，causal label を $X \to Y$ とみなすことができる．すなわち，causal label は次のように表すことができる:

$$X \to Y \quad \text{if} \begin{cases} P(X_{t+1}|S_X, S_Y) & = P(X_{t+1}|S_X) \\ P(Y_{t+1}|S_X, S_Y) & \neq P(Y_{t+1}|S_Y) \end{cases} \tag{2}$$

$$X \leftarrow Y \quad \text{if} \begin{cases} P(X_{t+1}|S_X, S_Y) & \neq P(X_{t+1}|S_X) \\ P(Y_{t+1}|S_X, S_Y) & = P(Y_{t+1}|S_Y) \end{cases} \tag{3}$$

$$\textit{No Causation} \quad \text{if} \begin{cases} P(X_{t+1}|S_X, S_Y) & = P(X_{t+1}|S_X) \\ P(Y_{t+1}|S_X, S_Y) & = P(Y_{t+1}|S_Y) \end{cases} \tag{4}$$

式 (2)，(3)，(4) に基づいて causal label を割り当てるためには，2 つの条件付き分布が同一か否かを判断する必要がある．条件付き分布を表すのに，提案手法では，回帰モデルの代わりに，カーネル平均を用いる．カーネル平均とは，分布を RKHS と呼ばれる特徴空間中の点として写像する関数である．興味深いことに，特性的なカーネル (e.g., ガウシアンカーネル) を用いた際には，この写像は単射になる，すなわち，異なる分布は同一の点に写像されることが無いことが知られている [17].

カーネル平均が条件付き分布 $P(X_{t+1}|S_X, S_Y)$, $P(X_{t+1}|S_X)$, $P(Y_{t+1}|S_X, S_Y)$, $P(Y_{t+1}|S_Y)$ を，それぞれ点 $\mu_{X_{t+1}|S_X,S_Y}, \mu_{X_{t+1}|S_Y} \in \mathcal{H}_X$, $\mu_{Y_{t+1}|S_X,S_Y}$, 及び点 $\mu_{Y_{t+1}|S_Y} \in \mathcal{H}_Y$ に写像するとする．すると，特性的なカーネルを用いれば，式 (2)，(3)，(4) は以下のように書き換えると，

$$X \to Y \quad \text{if} \begin{cases} \mu_{X_{t+1}|S_X,S_Y} & = \mu_{X_{t+1}|S_X} \\ \mu_{Y_{t+1}|S_X,S_Y} & \neq \mu_{Y_{t+1}|S_Y} \end{cases} \tag{5}$$

$$X \leftarrow Y \quad \text{if} \begin{cases} \mu_{X_{t+1}|S_X,S_Y} & \neq \mu_{X_{t+1}|S_X} \\ \mu_{Y_{t+1}|S_X,S_Y} & = \mu_{Y_{t+1}|S_Y} \end{cases} \tag{6}$$

$$\textit{No Causation} \quad \text{if} \begin{cases} \mu_{X_{t+1}|S_X,S_Y} & = \mu_{X_{t+1}|S_X} \\ \mu_{Y_{t+1}|S_X,S_Y} & = \mu_{Y_{t+1}|S_Y} \end{cases} \tag{7}$$

式 (5)，(6)，(7) に基づいて causal label を割り当てるためには，RKHS 中の 2 点が時刻 t を通して等しいか等しくないか，言い換えれば，2 点間の距離 — これはカーネル法のコミュニティにおいて maximum mean discrepancy (MMD) と呼ばれているものである [6] — が時刻 t を通してゼロになっているか否かを判断しさえすればよい．

以上の理由から，Granger causality を推定するための分類器を構築するために，提案手法では，MMD を用いた特徴量表現 $\nu(\cdot)$ を提案する．以下では，MMD の定義と推定方法について，簡単に述べる．

定義: k_X, k_Y をそれぞれ $\mathcal{X}, \mathcal{Y}$ 上のカーネルとし，$\mathcal{H}_X$, $\mathcal{H}_Y$ をそれぞれカーネル k_X, k_Y によって定義される RKHS とする．2 つの確率分布 $P(X_{t+1}|S_X, S_Y)$, $P(X_{t+1}|S_X)$ 間の距離は，MMD を用いると，RKHS 中の 2 点 $\mu_{X_{t+1}|S_X,S_Y}$, $\mu_{X_{t+1}|S_X} \in \mathcal{H}_X$ 間の距離として次のように定義される．

$$\text{MMD}^2_{X_{t+1}} \equiv |\mu_{X_{t+1}|S_X,S_Y} - \mu_{X_{t+1}|S_X}|^2_{\mathcal{H}_X} \tag{8}$$

様にして，$\text{MMD}^2_{Y_{t+1}}$ も，2 点 $\mu_{Y_{t+1}|S_X,S_Y}$, $\mu_{Y_{t+1}|S_Y} \in \mathcal{H}_Y$ 間の距離として定義される．

推定: MMD は，回帰モデルを用いることなく，また密度関数の推定すらなしに推定することができる．この点において，MMD は，Kolmogorov-Smirnov 検定量 [3] やカルバックライブラーダイバージェンス [10] よりも魅力的である．というのも，前者を用いる場合は回帰モデルを選択する必要があり，後者を用いる場合は，密度関数の推定が必要となり，これはデータのサンプル数が不十分な場合には難しいためである．

式 (8) の MMD を推定するには，条件付き分布に対するカーネル平均 $\mu_{X_{t+1}|S_X,S_Y}$, $\mu_{X_{t+1}|S_X}$ を推定する必要がある．カーネル平均に関するレビュー論文 [14] 等にも書かれているように，一般に分布に対するカーネル平均は，特徴写像と呼ばれる関数に関する重み付き和の形で推定される．具体的には，既存手法 Kernel Kalman Filter based on a Conditional Embedding Operator (KKF-CEO) [20] を用いれば，カーネル平均 $\mu_{X_{t+1}|S_X,S_Y}$, $\mu_{X_{t+1}|S_X}$ は，関数 Φ_X に関する重み付き和の形で次のように推定される

$$\hat{\mu}_{X_{t+1}|S_X,S_Y} = \sum_{\tau=2}^{t-1} w_\tau^{XY} \Phi_X(x_\tau) \tag{9}$$

$$\hat{\mu}_{X_{t+1}|S_X} = \sum_{\tau=2}^{t-1} w_\tau^{X} \Phi_X(x_\tau) \tag{10}$$

ここで，$\Phi_X(x_\tau) \equiv k_X(x_\tau, \cdot)$ は特徴写像であり，$\mathbf{w^{XY}} = [w_2^{XY}, \cdots, w_{t-1}^{XY}]^\top$, $\mathbf{w^{X}} = [w_2^{X}, \cdots, w_{t-1}^{X}]^\top$ $(t > 3)$ は実数値をとる重みベクトルである．

式 (9)，(10) を式 (8) に代入すれば，$\text{MMD}^2_{X_{t+1}}$ は，次のように推定できる．

Vol.2018-MPS-118 No.24
Vol.2018-BIO-54 No.24
2018/6/15

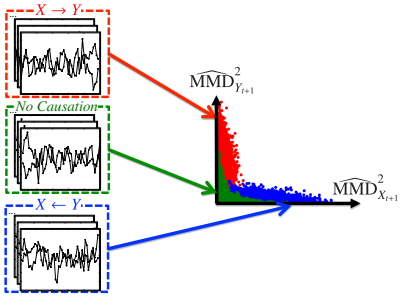

図 1　causal label の異なる時系列からは，十分異なる MMD のペアが推定される (各点は，個々の時系列から推定された MMD のペアを表している)

$$\widehat{\mathrm{MMD}}^2_{X_{t+1}} = \sum_{\tau=2}^{t-1}\sum_{\tau'=2}^{t-1}(w_\tau^{XY}w_{\tau'}^{XY} + w_\tau^{X}w_{\tau'}^{X} - 2w_\tau^{XY}w_{\tau'}^{X})k_X(x_\tau, x_{\tau'}) \tag{11}$$

3.2.2　特徴量表現

Granger causality 推定のための分類器を構築するために，提案手法では，式 (11) で推定できる MMD のペア $d_t = [\widehat{\mathrm{MMD}}^2_{X_{t+1}}, \widehat{\mathrm{MMD}}^2_{Y_{t+1}}]^\top$ を用いて特徴ベクトルを得る．

MMD のペアを用いれば，causal label の異なる時系列に対し，十分異なるような特徴ベクトルが得られると期待できる．これは，式 (5)，(6)，(7) からわかるように，causal label に依って，MMD がゼロになるか否かが異なるためである．実際には有限のデータサンプルからの推定量を用いるので MMD が厳密にゼロになることはないが，図 1 に示すように causal label に依って十分異なる MMD のペアが推定されると期待され，このことは第 4.1.2 章において実験的に確認した．

提案手法では，MMD のペアに関する経験分布をカーネル平均によって写像することで，特徴ベクトルを得ることを考える．k_X，k_Y とは異なるカーネル関数 k_D を導入して，特徴量表現を次のように定義した．

$$\nu(S) \equiv \frac{1}{T-W+1}\sum_{t=W}^{T}\Phi_D(d_t) \tag{12}$$
$$\text{where}\quad d_t = [\widehat{\mathrm{MMD}}^2_{X_{t+1}}, \widehat{\mathrm{MMD}}^2_{Y_{t+1}}]^\top$$

ここで，$\Phi_D(d_t) \equiv k_D(d_t, \cdot)$ は特徴写像である．式 (12) において，この特徴写像 $\Phi_D(\cdot)$ を計算するために，既存手法 Random Fourier Features (RFF) [15] を用いて，特徴写像を，カーネル関数に対するフーリエ変換よりサンプリングしたランダムな特徴を持つ低次元ベクトルとして近似した．実験では，特徴の数 m を $m = 100$ とし，各時系列に対する特徴ベクトルを m 次元ベクトルとして近似した．[*1]

3.3　多変数時系列への拡張

最後に，提案したアプローチを n 変数時系列 $(n \geq 3)$ に拡張するための方法論について述べる．

3.3.1　3 変数時系列の場合

3 変数時系列に対する特徴量表現は，条件付き Granger causality [4] に基づいて設計した．これは，定義 1 と異なり，多変数時系列に対して適用できるような Granger causality の定義である．

定義 1 に基づいて，3 変数時系列から Granger causality を推定すると，誤った結果を導くことが知られている．例えば，変数 X, Y 間に因果関係が無く，第 3 の変数 Z が X, Y の共通の原因である場合，X が Y の原因である，あるいは Y が X の原因である，と誤推定し得ることが知られている．これは，Z の影響により，$P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$ もしくは $P(X_{t+1}|S_X, S_Y) \neq P(X_{t+1}|S_X)$ が成立することがあるためである．

変数 Z の影響を考慮するため，条件付き Granger causality では，$\mathcal{Z}$ 上で定義される確率変数 $\{Z_1, \cdots, Z_t\}$ の観測 S_Z で条件づけられた 2 つの条件付き分布を考え，$P(Y_{t+1}|S_X, S_Y, S_Z) \neq P(Y_{t+1}|S_Y, S_Z)$ が成立するならば，Z が与えられたもとで X は Y の原因であるとし，そうでなければ Z が与えられたもとで X は Y の原因でないとするものである．

提案手法では，この条件付き Granger causality に基づいて causal label を導入することを考える．例えば，式 (2) と同様に，causal label $X \to Y$ を，

$$X \to Y \text{ if } \begin{cases} P(X_{t+1}|S_X, S_Y, S_Z) & = P(X_{t+1}|S_X, S_Z) \\ P(Y_{t+1}|S_X, S_Y, S_Z) & \neq P(Y_{t+1}|S_Y, S_Z) \end{cases}$$

とみなすことができ，これは

$$X \to Y \quad \text{if} \begin{cases} \mu_{X_{t+1}|S_X,S_Y,S_Z} & = \mu_{X_{t+1}|S_X,S_Z} \\ \mu_{Y_{t+1}|S_X,S_Y,S_Z} & \neq \mu_{Y_{t+1}|S_Y,S_Z} \end{cases}$$

と換言することができる．ただし，$\mu_{X_{t+1}|S_X,S_Y,S_Z}$，$\mu_{X_{t+1}|S_X,S_Z}$，$\mu_{Y_{t+1}|S_X,S_Y,S_Z}$，$\mu_{Y_{t+1}|S_Y,S_Z}$ は，条件付き分布 $P(X_{t+1}|S_X, S_Y, S_Z)$，$P(X_{t+1}|S_X, S_Z)$，$P(Y_{t+1}|S_X, S_Y, S_Z)$，$P(Y_{t+1}|S_Y, S_Z)$ に対するカーネル平均である．

2 変数に対し共通原因であるような変数が存在するようなケースに対応するため，$\mu_{X_{t+1}|S_X,S_Y,S_Z}$，$\mu_{X_{t+1}|S_X,S_Z}$ 間の MMD である $\widehat{\mathrm{MMD}}^2_{X_{t+1}|Z}$，及び $\mu_{Y_{t+1}|S_X,S_Y,S_Z}$，$\mu_{Y_{t+1}|S_Y,S_Z}$ 間の MMD である $\widehat{\mathrm{MMD}}^2_{Y_{t+1}|Z}$ を特徴量表現に加えることを考える．すなわち，特徴量表現 (12) を，d_t を次のようにすることで拡張する．

$$d_t = [\widehat{\mathrm{MMD}}^2_{X_{t+1}}, \widehat{\mathrm{MMD}}^2_{Y_{t+1}}, \widehat{\mathrm{MMD}}^2_{X_{t+1}|Z}, \widehat{\mathrm{MMD}}^2_{Y_{t+1}|Z}]^\top$$

3.3.2　n 変数時系列の場合 $(n > 3)$

n 変数時系列に対しては，上記の特徴量表現を用いて特徴ベクトルを計算した．提案手法では，それぞれの変数の

[*1] より大きい m を用いて実験をしても，精度の向上は特に観測されなかった．

ペア X, Y の間の因果関係を以下の 3 つのステップに基づいて推定する．第一に，変数の 3 つ組 (X, Y, Z_v) を $v \in \{1, \cdots, n-2\}$) のそれぞれに対して考え，これら 3 つの変数の観測に基づいて特徴ベクトルを計算する．第二に，学習した分類器を用いて，個々の特徴ベクトルに基づいて causal label ($X \to Y$, $X \leftarrow Y$, and *No Causation*) の割り当て確率を計算する．第三に，得られた割り当て確率の最も高いような causal label を時系列に割り当てることで，因果関係を推定する．

4. 実験

4.1 2 変数時系列を用いた実験

4.1.1 分類器の学習

2 変数時系列データを用いて，Granger causality を推定するための分類器を学習した．既存の教師あり学習に基づく手法 [2], [11], [12] と同様に，人工データ実験・実データ実験ともに，人工データを用いて分類器を学習した．これは，因果関係が既知であるような実データというのは非常に少ないためである．

訓練データとして，長さが $T = 42$ の 2 変数時系列データを 15,000 ペア用意した．具体的には，次のように線形な時系列データ，非線形な時系列データを用意した．

- 線形時系列: 以下の VAR モデルよりサンプルした．

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \frac{1}{P}\sum_{\tau=1}^{P} A_\tau \begin{bmatrix} X_{t-\tau} \\ Y_{t-\tau} \end{bmatrix} + \begin{bmatrix} E_{X_t} \\ E_{Y_t} \end{bmatrix} \quad (13)$$

ラベルが $X \to Y$ の時系列を得る際は，係数行列を

$$A_\tau = \begin{bmatrix} a_\tau & 0.0 \\ c_\tau & d_\tau \end{bmatrix}$$

とした．ここで，a_τ, d_τ は一様分布 $\mathcal{U}(-1, 1)$ よりサンプリングし，c_τ は $c_\tau \in \{-1, 1\}$ とした．同様に，ラベルが $X \leftarrow Y$, *No Causation* の時系列を得た．

- 非線形時系列: 上記と同様に VAR モデルに基づいて，標準シグモイド関数 $g(x) = 1/(1 + \exp(-x))$. を用いてサンプリングした．

4.1.2 人工データ実験

次のようにして生成した線形なテストデータ，非線形なテストデータを用いて，提案手法 (以降，*Supervised Inference of Granger Causality* (**SIGC**) と呼ぶ) の性能を評価した．

- 線形なテストデータ: 300 ペアの線形時系列を式 (13) に基づいて生成した．ここでラベル $X \to Y$, $X \to Y$, *No Causation* を有する時系列の数をそれぞれ 100 とし，いくらかのパラメータの設定は，訓練データとは異なる形で生成した (e.g., ノイズの分散は $p \in \{0.5, 1.0, 1.5, 2.0\}$ として与えた)．
- 非線形なテストデータ: 300 ペアの非線形時系列をラベル $X \to Y$, $X \to Y$, and *No Causation* を有する時系列の数が 100 となるように生成した．ここで，ラベルが $X \to Y$ の非線形時系列は次式で生成した．

$$X_t = 0.2X_{t-1} + 0.9E_{X_t} \quad (14)$$

$$Y_t = -0.5 + \exp(-(X_{t-1} + X_{t-2})^2) + 0.7\cos(Y_{t-1}^2) + 0.3E_{Y_t} \quad (15)$$

ここで，ノイズ変数 E_{X_t}, E_{Y_t} は標準正規分布からサンプリングした．同様にしてラベルが $X \leftarrow Y$ の時系列を生成した．ラベルが *No Causation* の時系列に関しては，式 (15) の指数関数項を無視することで用意した．

図 2 に，提案手法・比較手法を用いて得られた推定精度を示す．この結果から，回帰モデルを用いて Granger causality を同定する既存手法 ($\mathbf{GC}_{VAR}$, $\mathbf{GC}_{GAM}$, and $\mathbf{GC}_{KER}$) は，回帰モデルがデータにうまくフィッティングできるか否かによって推定精度が変わることがわかる．例えば，VAR モデルを用いた手法 $\mathbf{GC}_{VAR}$ は，線形なテストデータではうまく推定できているが，非線形なテストデータの場合，精度が低くなっている．これに対し，提案手法では線形なテストデータ，非線形なテストデータともに高い推定精度を達成できていることがわかる．その理由は，提案した特徴量表現に在る．このことは，同様に用意した訓練データを用いた既存の教師あり学習に基づく因果推論手法 **RCC** との比較からもわかる．

特徴量表現が causal label に応じて十分異なる特徴ベクトルを返していることを確認するため，非線形なテストデータを用いて，MMD のペア $\{d_t\}$ をヒストグラムとして可視化した．既に述べたように，これらの MMD のペアは個々の時系列に対する特徴ベクトルを計算する際に用いられるものである．結果は図 3 のようになり，確かに十分異なるような MMD のペアが得られていることがわかった．

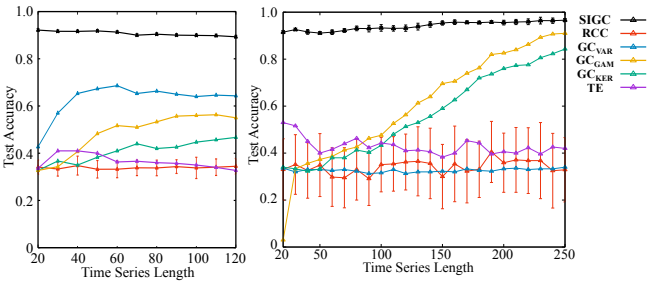

図 2 テストデータにおける推定精度 (左: 線形なテストデータ, 右: 非線形なテストデータ)

4.2 多変数時系列を用いた実験

4.2.1 実データ実験

第 3.3 章で述べた特徴量表現を用いた提案手法 $\mathbf{SIGC}_{tri}$ の性能を評価した．第 4.1 章での実験と同様，人工データを用いて分類器を学習し，テストデータは，次のような時

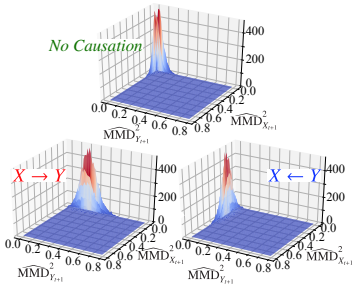

図 3 ラベル $X \to Y$ (左下), $X \leftarrow Y$ (右下), and *No Causation* (上) が割り当てられた個々の時系列に対して特徴ベクトルを得るのに用いた，MMD のペアに関するヒストグラム

表 1 マイクロアレイデータを用いた際の macro 平均 F1 値, micro 平均 F1 値

	$\mathbf{SIGC}_{tri}$	$\mathbf{SIGC}_{bi}$	RCC	$\mathbf{GC}_{VAR}$	$\mathbf{GC}_{GAM}$	$\mathbf{GC}_{KER}$	TE
macro 平均 F1	**0.483** (0.0)	0.431 (0.007)	0.407 (0.096)	0.457	0.437	0.351	0.430
micro 平均 F1	**0.637** (0.0)	0.578 (0.011)	0.567 (0.161)	0.567	0.513	0.436	0.449

系列マイクロアレイデータを用いた．

- *Saccharomyces cerevisiae*(酵母) の遺伝子発現量のデータ [16] を用いた．異なる実験条件下で測定された短い 4 本の時系列を結合し，長さ $T = 57$ の時系列を得た．ここで，遺伝子の数 (すなわち変数の数) は $n = 14$ であり，これらの間の真の因果関係は遺伝子制御ネットワークデータベース KEGG [9] に基づいて決定した．

表 1 にその結果を示す．異なる実験のもとで測定されたデータを結合したデータを用いているため，どの手法も十分に高い推定精度を達成しているとは言えない結果となった．しかし， 提案した $\mathbf{SIGC}_{tri}$ は Granger causality の既存手法より高い推定精度を達成した．また $\mathbf{SIGC}_{tri}$ は，第 3.2 章で述べた特徴量表現を用いた $\mathbf{SIGC}_{bi}$ よりも良い精度を示した．この結果は，第 3.3 章で述べたように，共通原因として働く変数の影響を考慮することの重要性を示唆している．

5. 結論

本稿では，Granger causality を同定する問題に対し，分類器を用いた新たなアプローチを提案した．回帰モデルを用いたモデルベースの既存手法では，回帰モデルがデータにうまくフィッティングできるか否かによって，推定精度が大きく変わってくるが，提案手法は，同一の特徴量表現，同一の分類器 (実験ではランダムフォレスト) を用いて，十分高い推定精度を達成した．また，Granger causality の有無・方向によって，十分異なる特徴ベクトルが得られることを実験的に示した．こうした結果は，教師あり学習に基づくアプローチの有効性を示唆している．

参考文献

[1] D. Bell, J. Kay, and J. Malley. A non-parametric approach to non-linear causality testing. *Economics Letters*, 51(1):7–18, 1996.

[2] G. Bontempi and M. Flauder. From dependency to causality: a machine learning approach. *JMLR*, 16:2437–2457, 2015.

[3] M. Chen and H. Z. An. A Kolmogorov-Smirnov type test for conditional heteroskedasticity in time series. *Statistics & probability letters*, 33(3):321–331, 1997.

[4] J. F. Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388):907–915, 1984.

[5] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[6] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520, 2007.

[7] I. Guyon. ChaLearn cause-effect pair challenge. https://www.kaggle.com/c/cause-effect-pairs/, 2013.

[8] M. Kar, Ş. Nazlıoğlu, and H. Ağır. Financial development and economic growth nexus in the MENA countries: Bootstrap panel granger causality analysis. *Economic modelling*, 28(1):685–693, 2011.

[9] KEGG: Kyoto Encyclopedia of Genes and Genomes. https://www.genome.jp/kegg/, 1995.

[10] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[11] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In *ICML*, pages 1452–1461, 2015.

[12] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. Discovering Causal Signals in Images. In *CVPR*, 2017.

[13] D. Marinazzo, M. Pellicoro, and S. Stramaglia. Kernel-Granger causality and the analysis of dynamical networks. *Physical Review E*, 77(5):056215, 2008.

[14] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[15] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2007.

[16] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

[17] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

[18] X. Sun. Assessing nonlinear Granger causality from multivariate time series. In *ECML*, pages 440–455. Springer, 2008.

[19] S. Yao, S. Yoo, and D. Yu. Prior knowledge driven Granger causality analysis on gene regulatory network discovery. *BMC bioinformatics*, 16(1):273, 2015.

[20] P. Zhu, B. Chen, and J. C. Principe. Learning nonlinear generative models of time series with a Kalman filter in RKHS. *IEEE Transactions on Signal Processing*, 62(1):141–155, 2014.