

# 教師あり学習に基づく Granger causality の推定

近原 鷹一<sup>1,a)</sup> 藤野 昭典<sup>1,b)</sup>

受付日 2018年5月16日, 再受付日 2018年7月5日/2018年8月22日,  
採録日 2018年9月3日

**概要:** Granger causality とは変数間の因果関係の定義の 1 つであり, その推定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 Granger causality の方向を推定するが, その推定精度は, 各々のデータに対して適切な回帰モデルを選択するか否かに強く依存する. しかし, 回帰モデルの選択には, データ解析に関する深い専門知識が要求されるため, 実際には容易なことではない. 本論文では, 教師あり学習に基づく Granger causality の推定手法を提案する. 提案手法では, 過去の値で条件付けられた条件付き分布間の距離を用いた特徴量表現を用いる. この特徴量表現が, Granger causality の有無・方向の異なる時系列に対して, 十分異なる特徴ベクトルを与えることを, 人工データを用いた実験により示す. また, このような特徴ベクトルの差異によって, 提案手法が既存手法より高い推定精度を達成したことを, 人工データ・実データを用いた実験により示す.

キーワード: Granger causality, 時系列解析, カーネル法

## A Supervised Learning Approach to Granger Causality Inference

YOICHI CHIKAHARA<sup>1,a)</sup> AKINORI FUJINO<sup>1,b)</sup>

Received: May 16, 2018, Revised: July 5, 2018/August 22, 2018,  
Accepted: September 3, 2018

**Abstract:** Granger causality is one of the definitions of temporal causality between variables, and inferring Granger causality is an important task in time series analysis. Traditional methods use regression models for this task. Since the inference accuracies of these methods depend largely on whether or not we select an appropriate regression model for each time series data. However, it is not easy because such selection of regression models requires a deep understanding of the data analysis. This paper proposes a supervised learning framework that utilizes a classifier instead of regression models. Our proposed method employs 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 Granger causality. In addition, we confirmed that such difference of feature vectors enables our method to achieve higher inference accuracy than the existing methods.

**Keywords:** Granger causality, time series analysis, kernel method

### 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. たとえば, 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 ( $X \rightarrow Y$ ) 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.

時間依存する変数間の因果関係の定義として, Granger causality [8] が, 幅広い分野で用いられてきた [14], [29]. 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, (Granger causality の意味で)  $X$  は  $Y$  の原因であると定義するものである.

<sup>1</sup> 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

Granger causality を同定するために、既存手法では一般に、ベクトル自己回帰 (VAR) モデルや一般化加法モデル (GAM) などの回帰モデルが用いられる。これらの手法では、 $Y$  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、 $[X, Y]^T$  の過去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、 $X$  から  $Y$  への Granger causality がある ( $X \rightarrow Y$ ) と決定する。この際に用いる回帰モデルがデータに適合しているならば、正しく Granger causality を同定することができるが、適合しない場合は、正しく同定できない。しかし、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には、データ解析に関する深い専門知識が要求されるため、データに適合する回帰モデルを選択することは一般に容易ではない。したがって、回帰モデルの予測誤差に基づいて Granger causality を同定する既存手法では、正しくその有無・方向を推定することは一般に難しい。

本研究の目的は、データ解析に関する深い専門知識を要求しないような、Granger causality の推定手法を確立することである。そのために、本論文では、教師あり学習に基づく Granger causality 推定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、 $X \rightarrow Y$ ,  $X \leftarrow Y$ , 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問題として解く。i.i.d. データ (independent and identically distributed; すなわち、独立同分布からサンプルされたデータ) を対象とした因果推論の場合、教師あり学習に基づく手法はすで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 [3], [10], [17], [18]。時系列データを対象とした教師あり学習に基づく Granger causality の推定を実現するため、本論文では Granger causality の有無・方向の異なる時系列に対して十分異なる特徴ベクトルを与える特徴量表現を提案する。この特徴量表現は、Granger causality の定義 —  $Y$  の過去の値で条件付けた  $Y$  の未来の値に関する条件付き分布と、 $[X, Y]^T$  の過去の値で条件付けた  $Y$  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) [9]) として分布間の距離を計算する。

実験を通して、提案手法が、回帰モデルを用いて Granger causality を同定する既存手法、および、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に比べ、高い推定精度を達成することを示す。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いいかについても言及する。提案手法の基本的なアイデアは文献 [6] において述べたが、Granger causality の有無・方向の違いによる特徴ベクトルの差異を示す実験は 1 種類の人工データを用いた結果しか示されておらず、また多変数時系列データを対象とした実験は実データ実験のみで実験結果の考察も乏しかった。本論文では、まず実験結果の信頼性を高めるために、別の方法で生成した人工データにおいても Granger causality の有無・方向の違いによる特徴ベクトルの差異が確認できたことを示した。また、提案した多変数拡張の有効性を人工データを用いた実験により検証した。さらに、教師あり学習に基づく因果推論手法の関連研究を詳述し、カーネル平均の推定量を得る方法についても詳述する。

本論文の構成は以下のとおりである。初めに 2 章にて、Granger causality の定義について述べる。続く 3 章では、提案手法に関して、タスクの概要 (3.1 節)、特徴量表現を定式化するうえでのアイデア・およびその具体的な定義 (3.2 節)、多変量時系列への拡張 (3.3 節)、の順に述べる。次に 4 章において、教師あり学習に基づく既存の因果推論手法について、その概要を述べる。そして 5 章では、既存手法との比較実験の結果について述べる。最後に、6 章にて、本論文の結論を述べる。

## 2. Granger causality

よく知られているように、変数間に相関関係があるからといって因果関係があるとは限らない。しかし、変数間に因果関係がある場合、相関関係が存在する [23]。

時間依存する変数間の因果関係は、原因が結果に時間的に先行する場合と、同一の時刻で定義される変数間に原因と結果の関係がある場合の 2 つが考えられる。Granger causality は、前者の場合によって生じた変数間の相関関係を検出するために提案された概念である。

Granger causality では、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、(Granger causality の意味で)  $X$  は  $Y$  の原因であるとみなす。これは、次のように定義される：

**定義 1 (Granger causality [8])** 定常過程、すなわち定常な確率変数の系列<sup>\*1</sup>  $\{(X_t, Y_t)\} (t \in \mathbb{N})$  を考える、ただし、 $X_t$  および  $Y_t$  は  $\mathcal{X}$ ,  $\mathcal{Y}$  上にそれぞれ定義されるとす

<sup>\*1</sup> ここで確率過程の定常性を仮定することで、Granger causality の有無・方向が時間に依存せずつねに一定であることを仮定している。非定常な確率変数の系列間の因果関係を考える場合、因果関係の方向が時間とともに変化する可能性を考慮する必要があることから、問題の簡略化のため、確率過程の定常性を仮定して Granger causality は定義される。

る。ここで、 $S_X, S_Y$  をそれぞれ確率変数  $\{X_1, \dots, X_t\}, \{Y_1, \dots, Y_t\}$  の観測とする。

このとき、

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

が成立するならば、(Granger causality の意味で)  $\{X_t\}$  が  $\{Y_t\}$  の原因であると定義し、

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

が成立するならば、(Granger causality の意味で)  $\{X_t\}$  は  $\{Y_t\}$  の原因でないと定義する。

条件付き分布  $P(Y_{t+1}|S_X, S_Y), P(Y_{t+1}|S_Y)$  が同一か否かを判断するために、既存手法 [1], [8], [19], [28] では、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。たとえば、既存手法 [8] 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、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、データをよく説明する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、本論文では、教師あり学習に基づく新たなアプローチを提案する\*2。

### 3. 提案手法

#### 3.1 タスク設定

提案手法では、Granger causality を同定する問題を教師あり学習の問題として解く。具体的には、因果関係が既知であるような時系列データ (訓練データ) を用いて分類器を学習し、得られた分類器を用いて、Granger causality の有無・方向が未知であるような時系列データ (テストデータ) について、その有無・方向を推定する、教師あり学習の問題を解く。

今、訓練データが、 $N$  ペアの 2 変数時系列データ  $S^1, \dots, S^N$  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  $\{(X_1^j, Y_1^j), \dots, (X_{T_j}^j, Y_{T_j}^j)\}$  ( $j \in \{1, \dots, N\}$ ) 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 Granger causality の有無・方向、すなわち  $X^j \rightarrow Y^j, X^j \leftarrow Y^j$ 、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  $X^j = (X_1^j, \dots, X_{T_j}^j), Y^j = (Y_1^j, \dots, Y_{T_j}^j)$  を表す)。

\*2 Granger causality は潜在交絡変数 (i.e., 2 変数  $X, Y$  の共通原因として働く非観測の変数) がある場合に誤推定を生じることが知られている。このため、Granger causality に関する既存手法 [1], [5], [8], [19], [28] と同様、提案手法は潜在交絡変数が存在しないことを仮定している。

$\nu(\cdot)$  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。

3.3 節で後述するように、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も拡張可能である。

#### 3.2 分類器の設計

各時系列に causal label を割り当てる分類器を構築するため、特徴量表現  $\nu(\cdot)$  を定式化する。以下では、各時系列データを変数間の Granger causality の有無・方向によって十分異なる特徴ベクトルに変換する方法に関して述べる。

##### 3.2.1 設計指針

本論文では、Granger causality の意味で、たとえば、 $X$  が  $Y$  の原因で  $Y$  は  $X$  の原因ではない場合、causal label を  $X \rightarrow Y$  と表す。すなわち定義 1 に基づいて、3 値の causal label を次のように表す\*3：

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

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

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

式 (2), (3), (4) に基づいて各時系列に causal label を割り当てるためには、条件付き分布が同一か否かを判断する必要がある。条件付き分布が同一か否かを判定するために、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。この写像は、特性的なカーネル (たとえばガウシアンカーネル) を用いた際には、単射になる、すなわち、異なる分布が同一の点に写像されることがないことが知られている [27]。したがって、カーネル平均を用いれば、式 (2), (3), (4) における条件付き分布間の等式・不等式が表す関係をそのままに、カーネル平均間の等式・不等式として表すことができる。

カーネル平均により条件付き分布  $P(X_{t+1}|S_X, S_Y), P(X_{t+1}|S_X), P(Y_{t+1}|S_X, S_Y), P(Y_{t+1}|S_Y)$  が、それぞれ点  $\mu_{X_{t+1}|S_X, S_Y}, \mu_{X_{t+1}|S_X} \in \mathcal{H}_X, \mu_{Y_{t+1}|S_X, S_Y}, \text{ および } \mu_{Y_{t+1}|S_Y}$

\*3 ここで、 $P(X_{t+1}|S_X, S_Y) \neq P(X_{t+1}|S_X)$  かつ  $P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$  である場合、すなわち (Granger causality の意味で)  $X$  が  $Y$  の原因で、かつ  $Y$  も  $X$  の原因である場合を考えていないが、このようなケースはさらにラベルを追加すれば対応できる。

$\in \mathcal{H}_Y$  に写像されるとする. ここで,  $\mathcal{H}_X, \mathcal{H}_Y$  はそれぞれ RKHS である. このとき, 式 (2), (3), (4) は以下のように書き換えることができる.

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

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

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

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うか, いい換えれば, 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 [9]—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.

提案手法では, Granger causality を推定するための分類器を, MMD に基づく特徴量表現  $\nu(\cdot)$  を用いて, 構築する. MMD を用いれば, 条件付き分布間の距離は以下のように定義できる [9].

**定義 2 (条件付き分布間の MMD)**  $k_X, k_Y$  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.

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

同様に,  $\text{MMD}_{Y_{t+1}}^2$  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.

MMD の推定には, データをうまく説明できる適切な回帰モデルの選択も, 条件付き分布の密度関数の推定も不要である. この点において, MMD は, Kolmogorov-Smirnov 検定量 [4] やカルバックライブラーダイバージェンス [16] よりも魅力的である. というのも, 前者を用いて条件付き分布間の距離を推定する場合は, 適切な回帰モデルを選択する必要があり, 後者を用いる場合は, 条件付き分布の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.

### 3.2.2 過去の観測すべてで条件付けられた分布のカーネル平均

式 (8) の MMD を推定するには, 時刻  $t$  以前の過去の観測のすべてで条件付けられた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. 付録の A.2 章においても述べるが, カーネル平均は一般に, 写像する分布に関する期待値の形で定義され,  $\mu_{X_{t+1}|S_X, S_Y},$

$\mu_{X_{t+1}|S_X}$  の場合, それぞれ条件付き分布  $P(X_{t+1}|S_X, S_Y), P(X_{t+1}|S_X)$  に関する期待値として,

$$\mu_{X_{t+1}|S_X, S_Y} \equiv E_{X_{t+1}|S_X, S_Y}[\Phi_X(X_{t+1})] \quad (9)$$

$$\mu_{X_{t+1}|S_X} \equiv E_{X_{t+1}|S_X}[\Phi_X(X_{t+1})] \quad (10)$$

と定義される. ここで,  $\Phi_X(X_{t+1}) \equiv k_X(X_{t+1}, \cdot)$  は, カーネル関数  $k_X$  によって定義される, 特徴写像と呼ばれる関数で, たとえばガウシアンカーネル  $k_X(x, x') = \exp(-\gamma\|x - x'\|^2)$  ( $\gamma > 0$  はパラメータ) の場合, 特徴写像は  $\Phi_X(x) = \exp(-\gamma x^2)[1, \sqrt{2\gamma/1!}x, \sqrt{(2\gamma)^2/2!}x^2, \dots]^\top$  と表される.

式 (9), 式 (10) から分かるように, カーネル平均  $\mu_{X_{t+1}|S_X, S_Y}, \mu_{X_{t+1}|S_X}$  を推定するためには, 過去の観測のすべてで条件付けられた条件付き分布  $P(X_{t+1}|S_X, S_Y), P(X_{t+1}|S_X)$  に関する  $\Phi_X(X_{t+1})$  の期待値を推定する必要がある. このために提案手法では, 既存手法 Kernel Kalman Filter based on a Conditional Embedding Operator (KKF-CEO) [30] を用いる. KKF-CEO では, 非線形時系列も予測の対象とできるよう, 状態空間モデルと呼ばれる既存の生成モデルをカーネル法の概念に基づいて拡張したモデルを用いて, RKHS 上の確率変数  $\Phi_X(X_{t+1})$  の値を予測し, この予測値に基づいて  $X_{t+1}$  の値を予測する.

詳細は付録 A.1 章で述べるが, たとえば 1 次のマルコフモデルを用いる場合, 1 時刻前の観測で条件付けられた分布  $P(X_{t+1}|x_t)$  に関する期待値  $E_{X_{t+1}|x_t}[X_{t+1}]$  を予測する形になるが, 文献 [2] にもあるように, 状態空間モデルを用いれば, 過去の観測  $S_X = \{x_1, \dots, x_t\}$  のすべてで条件付けられた分布  $P(X_{t+1}|S_X)$  に関する期待値  $E_{X_{t+1}|S_X}[X_{t+1}]$  を推定できる. したがって, KKF-CEO を用いて, 状態空間モデルに基づいて  $\Phi_X(X_{t+1})$  の値を予測すれば, 式 (10) で定義されるカーネル平均, すなわち  $\mu_{X_{t+1}|S_X} \equiv E_{X_{t+1}|S_X}[\Phi_X(X_{t+1})]$  を推定できる\*4.

### 3.2.3 カーネル平均, MMD の推定

3.2.2 項では, 既存手法 KKF-CEO によって過去の観測すべてで条件付けた分布のカーネル平均を推定できる理由について述べた. 本章では, これらのカーネル平均の推定量がどのような形で表されるかについて述べる.

付録の A.2 章においても述べるが, カーネル平均は一般に, 特徴写像に関する重み付き和の形で推定される [20]. これは条件付き分布のカーネル平均  $\mu_{X_{t+1}|S_X, S_Y}, \mu_{X_{t+1}|S_X}$  の場合も例外でなく, KKF-CEO を用いれば, 関数  $\Phi_X$  に関する重み付き和の形で次のように推定される:

$$\hat{\mu}_{X_{t+1}|S_X, S_Y} = \sum_{\tau=2}^{t-1} w_\tau^{XY} \Phi_X(x_\tau) \quad (11)$$

\*4 カーネル平均を推定する際に用いる状態空間モデルには, RKHS 上の確率変数が正規分布に従うという仮定がある. この仮定が成立しない場合に, 推定量にどの程度バイアスが生じるかを考察することは今後の課題である.

$$\hat{\mu}_{X_{t+1}|S_X} = \sum_{\tau=2}^{t-1} w_{\tau}^X \Phi_X(x_{\tau}) \quad (12)$$

ここで、 $\mathbf{w}^{XY} = [w_2^{XY}, \dots, w_{t-1}^{XY}]^{\top}$ 、 $\mathbf{w}^X = [w_2^X, \dots, w_{t-1}^X]^{\top}$  ( $t > 3$ ) は実数値をとる重みベクトルである。

式 (11), (12) を式 (8) に代入すれば、 $\text{MMD}_{X_{t+1}}^2$  は、次のように推定できる。

$$\begin{aligned} \widehat{\text{MMD}}_{X_{t+1}}^2 &= \sum_{\tau=2}^{t-1} \sum_{\tau'=2}^{t-1} (w_{\tau}^{XY} w_{\tau'}^{XY} + w_{\tau}^X w_{\tau'}^X \\ &\quad - 2w_{\tau}^{XY} w_{\tau'}^X) k_X(x_{\tau}, x_{\tau'}) \end{aligned} \quad (13)$$

以下では、KKF-CEO を用いた式 (11), 式 (12) のカーネル平均の導出についてその詳細を述べる。

### 式 (12) のカーネル平均の推定：

文献 [30] にあるように、KKF-CEO では未来の時刻に関する分布のカーネル平均を、過去の時刻に関する分布のカーネル平均が、conditional embedding operator と呼ばれる作用素によって写像されたものとして表す。詳しくは付録 A.2 章で述べるが、カーネル平均が確率分布を RKHS 上の 1 点に写像する関数であるのに対し、conditional embedding operator は RKHS 上の点を別の RKHS 上の点に写像するものである。

式 (12) の  $P(X_{t+1}|S_X)$  に対するカーネル平均  $\mu_{X_{t+1}|S_X}$  の場合、KKF-CEO では、 $P(X_t|S_X)$  に対するカーネル平均  $\mu_{X_t|S_X}$  が、conditional embedding operator  $C_{X_{t+1}|X_t}$  によって写像されたものとして、

$$\mu_{X_{t+1}|S_X} = C_{X_{t+1}|X_t} \mu_{X_t|S_X}$$

と表す。文献 [30] に述べられているように、観測時系列  $S_X$  を用いれば、conditional embedding operator  $C_{X_{t+1}|X_t}$  とカーネル平均  $\mu_{X_t|S_X}$  は、次のように推定できる。

$$\begin{aligned} \hat{C}_{X_{t+1}|X_t} &= \Lambda(K + (t-2)\lambda I_{t-2})^{-1} \Upsilon^{\top} \\ \hat{\mu}_{X_t|S_X} &= \sum_{\tau=2}^t b_{\tau} \Phi_X(x_{\tau}) \end{aligned}$$

ここで、 $\Lambda$ ,  $\Upsilon$  は  $\Lambda = [\Phi_X(x_2), \dots, \Phi_X(x_{t-1})]$ ,  $\Upsilon = [\Phi_X(x_1), \dots, \Phi_X(x_{t-2})]$  であり、行列  $K$  はグラム行列で  $K_{i,j} = k_X(x_i, x_j)$ ,  $\lambda$  は正のパラメータ、 $I_{t-2}$  は  $(t-2) \times (t-2)$  の単位行列である。また、 $\mathbf{b}$  は実数値をとる重みベクトルで、その推定方法については、文献 [30] にあるため、本論文では割愛する。これらの結果から、式 (12) のカーネル平均は次のように表せる。

$$\begin{aligned} \hat{\mu}_{X_{t+1}|S_X} &= \Lambda(K + (t-2)\lambda I_{t-2})^{-1} \Upsilon^{\top} \sum_{\tau=2}^t b_{\tau} \Phi_X(x_{\tau}) \\ &= \Lambda \mathbf{w}^X \\ &= \sum_{\tau=2}^{t-1} w_{\tau}^X \Phi_X(x_{\tau}) \end{aligned} \quad (12)$$

式 (11) のカーネル平均の推定：

最後に、式 (11) に示した、 $P(X_{t+1}|S_X, S_Y)$  のカーネル平均について述べる。

この条件付き分布は、 $P(X_{t+1}|S_X)$  と異なり、 $S_X, S_Y$  により条件付けられている。付録 A.2.2 項で後述するが、一般に条件付き分布のカーネル平均は、分布を条件付ける値によって推定する際の重みベクトルの値が変わることが知られている [20]。このため、式 (12) の重み  $\mathbf{w}^X$  とは異なる記号  $\mathbf{w}^{XY}$  を用いて表す。

提案手法では、まずこの重みベクトル  $\mathbf{w}^{XY}$  を計算するため、同時分布  $P(X_{t+1}, Y_{t+1}|S_X, S_Y)$  のカーネル平均の推定量を計算する。付録 A.2.3 節で述べるように、同時分布のカーネル平均は、積カーネル  $k_{XY}$  が定義する特徴写像  $\Phi_X \otimes \Phi_Y$  に関する重み付き和の形で、

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

として推定される。ここで、重みベクトル  $\mathbf{w}^{XY}$  の値は積カーネル  $k_{XY}$  とともに KKF-CEO を用いれば同様に計算できる。

同時分布  $P(X_{t+1}, Y_{t+1}|S_X, S_Y)$  と、条件付き分布  $P(X_{t+1}|S_X, S_Y)$  は、いずれも分布を条件付ける値が  $S_X, S_Y$  で同じである。このため、重みベクトル  $\mathbf{w}^{XY}$  を用いれば、条件付き分布  $P(X_{t+1}|S_X, S_Y)$  のカーネル平均は、以下のように推定できる。

$$\hat{\mu}_{X_{t+1}|S_X, S_Y} = \sum_{\tau=2}^{t-1} w_{\tau}^{XY} \Phi_X(x_{\tau}) \quad (11)$$

### 3.2.4 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 (13) で推定できる MMD のペア  $d_t = [\widehat{\text{MMD}}_{X_{t+1}}^2, \widehat{\text{MMD}}_{Y_{t+1}}^2]^{\top}$  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) から分かるように、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 によって十分異なる MMD のペアが推定されると期待され、実際、5.2.2 項に示すように、人工データを用いた実験で、causal label の違いによる MMD のペアの差異が確認できた。

式 (5), (6), (7) は時刻  $t$  によらず成立することから、このような MMD のペアの差異は時刻  $t$  によらず存在する。このような差異を活用するため、提案手法では各時刻  $t$  に関する MMD のペア  $d_t$  を用いる。そのために、長さ  $T$  の時系列データ  $S = \{(x_1, y_1), \dots, (x_T, y_T)\}$  より、長さ  $W$  ( $W < T$ ) の部分時系列  $\{(x_{t-(W-1)}, y_{t-(W-1)}), \dots, (x_t,$

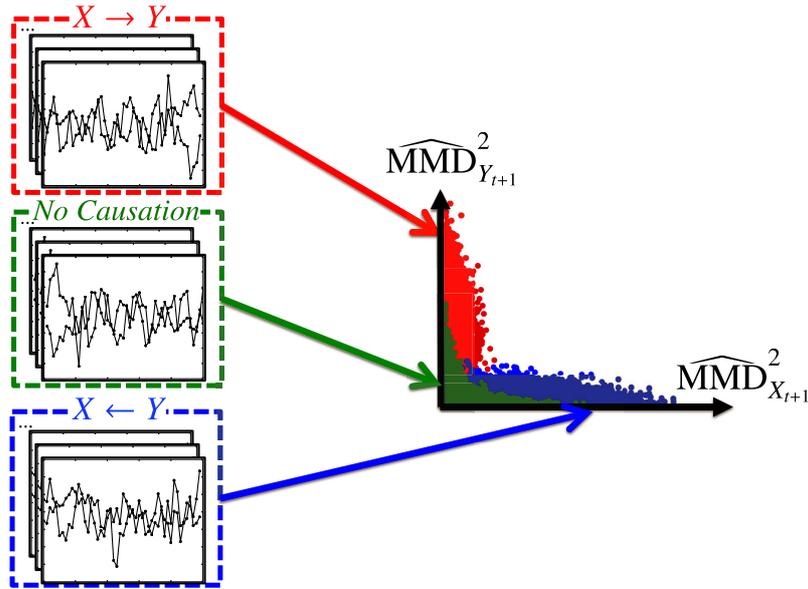


図 1 causal label の異なる時系列からは、十分異なる MMD のペアが推定される（各点は、個々の時系列から推定された MMD のペアを表している）

Fig. 1 Different MMD pairs are estimated from time series with different causal labels (Each dot represents the MMD pair estimated from each time series).

$y_t\}$  ( $t = W, \dots, T$ ) を用意し、これを用いた\*5。結果として、MMD のペアの系列  $\{d_W, \dots, d_T\}$  を得た。

時刻  $t$  を通して MMD がゼロになっているか否かを判定するため、これらの MMD のペアの系列全体を用いて 1 つの特徴ベクトルを得る。MMD のペアの系列全体をそのまま結合して 1 つの特徴ベクトルとすることも可能であるが、そのように特徴ベクトルを定めるとベクトルの次元が  $2(T - W + 1)$  となり、長さ  $T$  の異なる時系列に対して次元数が異なる特徴ベクトルを与えてしまう形になる。

そこで、提案手法では次元数が時系列の長さ  $T$  に依存しないような特徴ベクトルを得るため、MMD のペアの系列が従う分布に対するカーネル平均を考え、その推定量に基づいて特徴ベクトルを計算する\*6。そのために、 $k_X, k_Y$  とは異なる、新たなカーネル  $k_D$  を導入する。今、 $\mathcal{D}$  上の確率変数を  $D_t$  とし、MMD のペア  $\{d_W, \dots, d_T\}$  がある分布  $Q(D_t)$  に従うサンプルであるとする。分布  $Q(D_t)$  のカーネル平均の推定量を、 $\mathcal{D}$  上のカーネル  $k_D$  が定義する特徴写像  $\Phi_D(d_t) \equiv k_D(d_t, \cdot)$  を用いて表すことで特徴量表現を次のように定義する\*7。

$$\nu(S) \equiv \frac{1}{T - W + 1} \sum_{t=W}^T \Phi_D(d_t) \quad (14)$$

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

式 (14) において、特徴写像  $\Phi_D(\cdot)$  を計算するために、既存手法 Random Fourier Features (RFF) [22] 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、この特徴の数  $m$  を  $m = 100$  とし、各時系列  $S$  に対する特徴ベクトル  $\nu(S)$  を  $m$  次元ベクトルとして近似した。ここで、より大きな  $m$  を用いて実験をしても、大きく推定精度が向上することはなかった。

### 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 [7] 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。たとえば、(Granger causality の意味で) 変数  $X, Y$  間に因果関係がなく、第 3 の変数  $Z$  が  $X, Y$  の共通の原因で

\*7 式 (14) において、 $1/(T - W + 1)$  は、推定量を得るための重みの値に相当する。一般に分布から直接サンプリングしたサンプルを用いてカーネル平均を推定する際、重みの値が同一になることが知られている [20]。

\*5 このように短い部分時系列を用いることで、KKF-CEO を用いて重みベクトルを計算する際の計算量 (時系列の長さ  $T$  に対し、時間計算量は  $O(T^3)$  [30]) を小さくすることができる。

\*6 標本平均に基づいて  $\nu(S) = (d_W + \dots + d_T)/(T - W + 1)$  としても、次元数が時系列の長さ  $T$  に依存しない特徴ベクトルは得られる。しかし、このように平均のみ考慮した特徴量表現では、MMD のペアの系列が 2 つ与えられ、そのそれぞれに対し特徴ベクトルを計算する際、各系列が従う分布が、平均は等しいがたとえば分散が異なるような場合、同一の値の特徴ベクトルになる。しかし、カーネル平均を用いれば、すでに述べたように、異なる分布を同一の点には写像しないため、このような問題は生じない。

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、と誤推定しう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。また、変数  $X, Y$  間に Granger causality はないが、たとえば  $X$  が  $Y$  に対して  $Z$  を介して影響を与える場合 (i.e.,  $X \rightarrow Z, Z \rightarrow Y$ ) も、 $Z$  の影響により、 $P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$  が成立することがあるため、同様の理由で、 $X$  から  $Y$  に Granger causality があると誤推定しうることが知られている。

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  $\{Z_1, \dots, Z_t\}$  の観測  $S_Z$  で条件付けられた 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  $Y$  の原因でないと定義する。

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。たとえば、式 (2) と同様に、causal label  $X \rightarrow Y$  を、

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

とみなせ、これは次のように表せる。

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

ここで、 $\mu_{X_{t+1}|S_X, S_Y, S_Z}, \mu_{X_{t+1}|S_X, S_Z}, \mu_{Y_{t+1}|S_X, S_Y, S_Z}, \mu_{Y_{t+1}|S_Y, S_Z}$  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 変数の間に共通原因であるような変数が存在する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  $\widehat{\text{MMD}}_{X_{t+1}|Z}^2$ 、および  $\mu_{Y_{t+1}|S_X, S_Y, S_Z}, \mu_{Y_{t+1}|S_Y, S_Z}$  間の MMD である  $\widehat{\text{MMD}}_{Y_{t+1}|Z}^2$  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 (14) を、 $d_t$  を以下のようすることで拡張する\*8

$$\nu(S) \equiv \frac{1}{T - W + 1} \sum_{t=W}^T \Phi_D(d_t) \quad (15)$$

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

\*8  $d_t = \left[ \widehat{\text{MMD}}_{X_{t+1}|Z}^2, \widehat{\text{MMD}}_{Y_{t+1}|Z}^2 \right]^\top$  としても 2 変数の間に共通原因が存在するケースに対応できると考えられるが、一般に条件付き分布の密度関数は条件付ける変数が多いほど推定は難しく、これはカーネル平均についても同じであることから、式 (15) のように、条件付ける変数の数が少ない分布間の MMD に加える形で、 $d_t$  を定めている。

### 3.3.2 $n$ 変数時系列の場合 ( $n > 3$ )

式 (15) に示した 3 変数時系列に対する特徴量表現に対して、さらに MMD を  $d_t$  に加えることで、 $n$  変数時系列 ( $n > 3$ ) に対する特徴量表現に拡張することは可能である。しかし、 $n$  変数の場合、各変数のペアに対して共通原因となりうる変数の組合せの数は指数爆発するため、 $n$  変数からなる訓練データを十分に用意することはきわめて難しい。こうした理由から、提案手法では、 $n$  変数の場合においても、式 (15) の特徴量表現を用いた。

提案手法では、 $n$  変数のうち 2 変数を選んでできる  $nC_2$  通りの変数のペアに対し、各変数ペア間の Granger causality の有無・方向を、式 (15) の特徴量表現を用いて推定する。

以下、 $n$  変数のうちのある変数のペアを  $X, Y$  とし、 $X, Y$  間の因果関係を推定する方法について述べる。まず、変数の 3 つ組  $(X, Y, Z_v)$  を  $v \in \{1, \dots, n-2\}$  のそれぞれについて考え、この 3 変数に関する時系列データから、式 (15) を用いて特徴ベクトルを計算する。次に、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 ( $X \rightarrow Y, X \leftarrow Y$ , and *No Causation*) の割り当て確率を計算する。最後に、割り当て確率の最も高い 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. 関連研究

i.i.d. データから因果関係を推定する問題を教師あり学習の問題として解くタスクが初めて行われたのは、近年 Guyon らによって開催された ChaLearn [10] と呼ばれるコンペティションである。このコンペティションでは、それぞれの参加者が、平均、分散、エントロピーなど様々な統計量をもとに特徴量表現を設計して分類器を学習し、テストデータにおいて最も高い精度で causal label を割り当てる分類器を構築できた参加者が優勝であるとした。

このコンペティション以降、因果関係の定義に紐付けて特徴量表現を設計した手法がいくらか提案された。そのうち、提案手法と特に関連が深いのが Randomized Causation Coefficient (RCC) [17] である。これは、提案手法と同様、因果関係の方向によって十分異なる特徴ベクトルを得るためにカーネル平均を用いた既存手法である。しかし RCC は、提案手法とは異なる分布の特徴を得るために考案されている。具体的には、RCC はカーネル平均を用いて、周辺分布と条件付き分布の情報を特徴ベクトルとして得る。これは、*independence of cause and mechanism* (ICM) [12] と呼ばれる、因果推論のコミュニティにおいて提案された仮説から、因果関係の方向によって 2 つの分布の情報が異なると知られているためである。これに対し提案手法では、カーネル平均を用いて、過去の値で条件付けた条件付き分布間の距離を計算し、これに基づいて特徴ベクトルを得る。これはすでに述べたように、Granger causality の有無・方向によって、これらの条件付き分布間の距離が大き

く異なるためである。

RCC のほかに、多変数の i.i.d. データに対して特徴量表現を設計した手法 D2C アルゴリズム [3] や、学習した分類器を用いて静的画像中のオブジェクトと背景を識別する手法 Neural Causation Coefficient (NCC) [18] もある。

## 5. 実験

### 5.1 実験設定

提案手法 (以降, *Supervised Inference of Granger Causality* (SIGC) と呼ぶ) の性能を, i.i.d. データより教師あり学習によって因果関係を推定する既存手法 RCC [17]<sup>\*9</sup>, VAR モデル, GAM, カーネル回帰を用いて Granger causality を同定する既存手法  $\mathbf{GC}_{VAR}$  [8]<sup>\*10</sup>,  $\mathbf{GC}_{GAM}$  [1]<sup>\*7</sup>,  $\mathbf{GC}_{KER}$  [19]<sup>\*11</sup>, および回帰モデルではなく、密度関数の推定に基づいて因果関係を推定する transfer entropy TE [25]<sup>\*12</sup> と、比較した。

提案手法においては、分類器としてランダムフォレストを用いた<sup>\*13</sup>。ここでランダムフォレストを選んだのは、比較手法の 1 つである RCC が、SVM を用いた場合よりランダムフォレストを用いた場合のほうが実験的に高い推定精度を達成しているためである [17]。特徴ベクトルを用意するために、カーネル関数  $k_X, k_Y$ , および  $k_D$  としてガウシアンカーネルを用い、そのカーネルパラメータは median heuristic と呼ばれるよく知られたヒューリスティクスによって選択した [24]。提案手法のパラメータ  $W$ , および既存手法のパラメータは、後述する人工データ実験において、各手法が最良の性能が得られる値に設定した。その結果、提案手法の  $W$  は、 $W = 12$  となった。

### 5.2 2変数時系列を用いた実験

#### 5.2.1 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 [3], [17], [18] 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<sup>\*14, \*15</sup>。

- 線形時系列：以下の VAR モデルよりサンプルした。

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \frac{1}{P} \sum_{\tau=1}^P A_\tau \begin{bmatrix} X_{t-\tau} \\ Y_{t-\tau} \end{bmatrix} + \begin{bmatrix} E_{X_t} \\ E_{Y_t} \end{bmatrix} \quad (16)$$

ここで、 $\tau$  は  $\tau = 1, \dots, P$  ( $P \in \{1, 2, 3\}$ ) であり、 $E_{X_t}, E_{Y_t}$  は標準正規分布  $\mathcal{N}(0, 1)$  からサンプリングしたノイズである。ラベルが  $X \rightarrow Y$  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、 $a_\tau, d_\tau$  は一様分布  $\mathcal{U}(-1, 1)$  よりサンプリングし、 $c_\tau$  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。たとえば、ラベルが  $X \rightarrow Y$  の時系列を得る際は、 $Y_t$  が  $\{[g(X_{t-\tau}), Y_{t-\tau}]^\top\}_{\tau=1}^P$  に依存し、 $X_t$  が  $\{X_{t-\tau}\}_{\tau=1}^P$  のみに依存するような形で、サンプリングした。

#### 5.2.2 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、評価実験を行った。

- 線形なテストデータ：300 ペアの線形時系列を式 (16) に基づいて生成した。ここでラベル  $X \rightarrow Y$ ,  $X \leftarrow Y$ , *No Causation* を有する時系列の数をそれぞれ 100 とし、いくつ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  $X \rightarrow Y$ ,  $X \leftarrow Y$ , 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  $X \rightarrow Y$  の非線形時系列は次式で生成した。

$$X_t = 0.2X_{t-1} + 0.9E_{X_t} \quad (17)$$

$$Y_t = -0.5 + \exp(-(X_{t-1} + X_{t-2})^2) + 0.7 \cos(Y_{t-1}^2) + 0.3E_{Y_t} \quad (18)$$

ここで、ノイズ変数  $E_{X_t}, E_{Y_t}$  は標準正規分布  $\mathcal{N}(0, 1)$  からサンプリングした。同様にラベルが  $X \leftarrow Y$  の時系列を生成した。ラベルが *No Causation* の時系列は、式 (18) の指数関数項を無視することで用意した。

<sup>\*9</sup> [https://github.com/lopezpaz/causation\\_learning\\_theory](https://github.com/lopezpaz/causation_learning_theory)

<sup>\*10</sup> <http://people.tuebingen.mpg.de/jpeters/onlineCodeTimino.zip>

<sup>\*11</sup> <https://github.com/danielemarinazzo/KernelGrangerCausality>

<sup>\*12</sup> <https://github.com/Healthcast/TransEnt>

<sup>\*13</sup> ここで、ランダムフォレストの木の数は、 $\{100, 200, 500, 1000, 2000\}$  より交叉検証法に基づいて与えた。

<sup>\*14</sup> 線形時系列、非線形時系列ともに、平均 0、分散 1 になるように正規化を行った。これは、分散の大小のみに基づいて分類結果を出力し、訓練データを用いない既存手法と公平な比較を行うための処理である。実際、既存の因果推論手法 RCC [17] でも、同様の正規化処理がなされている。

<sup>\*15</sup> 実験では、2 種類のモデルよりサンプリングした訓練データによって高い test accuracy が得られているが、任意のテストデータで動作する保証はない。したがって、実用上は多様な種類のモデルよりサンプリングした訓練データを用いることが望ましい。

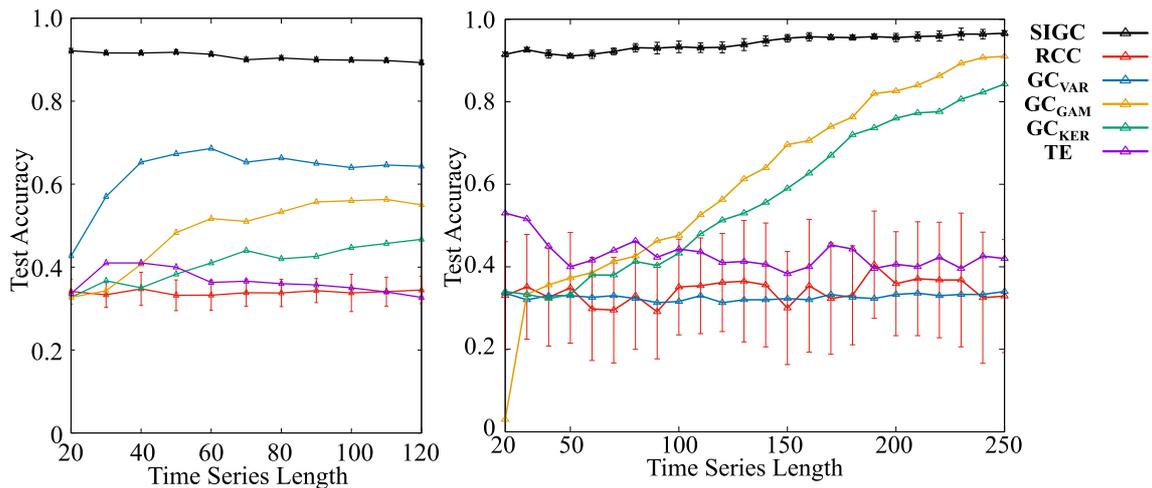


図 2 テストデータにおける推定精度 (左: 線形なテストデータ, 右: 非線形なテストデータ). 提案手法と **RCC** に関しては, 異なる訓練データを用いて行った 20 回の実験における, 推定精度の平均値, および標準偏差値 (エラーバー) を示している

Fig. 2 Test accuracies (left: linear test data; right: nonlinear test data). Means and standard deviations (error bars) are shown for our method and **RCC** based on 20 runs with different training data.

線形なテストデータ, 非線形なテストデータを用いて, 提案手法の性能を既存手法と比較した. 図 2 に, 各手法の推定精度を示す. ここで, 提案手法と **RCC** に関しては, ランダムに生成した訓練データを用いているため, 異なる訓練データを用いて 20 回実験を行った際の, 平均と標準偏差 (エラーバー) を示している.

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 ( $GC_{VAR}$ ,  $GC_{GAM}$ ,  $GC_{KER}$ ) は, 回帰モデルがデータにうまく適合するか否かによって推定精度が変わることが分かる. たとえば, VAR モデルを用いた手法  $GC_{VAR}$  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. また, 非線形なテストデータを用いた実験において,  $GC_{KER}$  は  $GC_{GAM}$  より精度が低くなっている. これは, カーネル回帰を行うためには, 時系列の長さが小さすぎるためである. 同様に, 密度関数の推定を行うためには時系列の長さが短いため, **TE** の精度も低くなっている.

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分かる. その理由は, 提案した特徴量表現にあると考えられる. これは, 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** との比較からも分かる.

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すことを確認するため, 以下のような検証実験を行った. 初めに, 非線形なテストデータを用いて, MMD のペア  $\{d_i\}$  をヒストグラムとして可視化した. すでに述べたように, 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, 個々の MMD は有限のデータサンプルから推定したもので厳密にゼロになることはないものの, 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分かった. 同様の実験を線形なテストデータを用いて行ったところ, 図 4 のようになり, 同様に特徴量表現の有効性を示唆する結果が得られた.

causal label によって十分異なる MMD のペアを得られるのならば, 訓練データを使わずとも causal label を割り当てることができるとも考えられる. 実際, 訓練データを使わず, いわば教師なし学習の形で causal label を出力する, 次のようなアプローチを考えることもできる. まず, MMD のペアの系列を用いて,  $\widehat{MMD}_{X_{t+1}}^2$  の平均がゼロか否か,  $\widehat{MMD}_{Y_{t+1}}^2$  の平均がゼロか否かを判断するため, 統計的仮説検定を行う. すると, ここで得られた 2 つの  $p$  値と何らかの閾値 (有意水準) を用いれば, 各時系列に対し, causal label ( $X \rightarrow Y$ ,  $X \leftarrow Y$ , または *No Causation*) を割り当てることができる.

しかし, こうした教師なし学習によるアプローチは, その性能が閾値に強く依存することが, 実験的に確認された. さらに, 最も性能が良くなるような閾値を選んでも, その精度は提案手法よりも低く, たとえば非線形のテストデータ (系列長  $T = 250$ ) を用いた場合, その推定精度は 0.810 となり, 提案手法の推定精度 0.966 を下回った. これらの結果は, 分類器を学習することで causal label を決定するための決定境界が得られるような, 教師あり学習のアプローチの有効性を示唆している.

### 5.2.3 実データ実験

実データを用いて提案手法の性能を評価した. ここで, 実験の信頼性を高めるため, 以下のような 2 種類のテスト

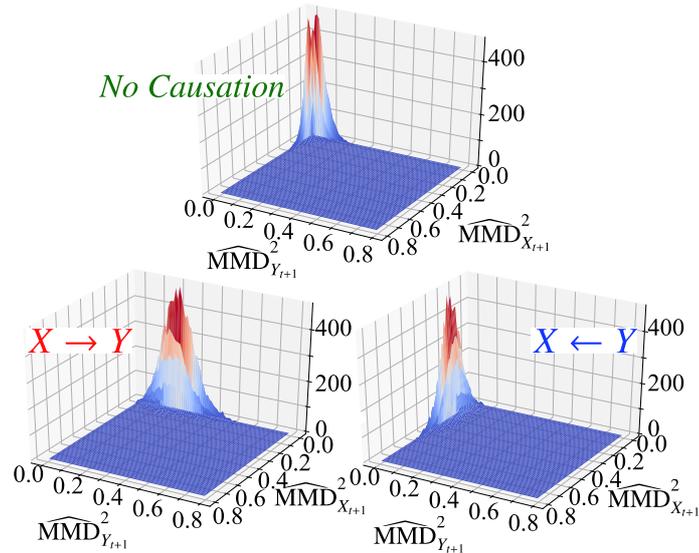


図 3 ラベル  $X \rightarrow Y$  (左下),  $X \leftarrow Y$  (右下), and *No Causation* (上) が割り当てられた, 非線形なテストデータ中の個々の時系列より得た, MMD のペアに関するヒストグラム

Fig. 3 Histogram of MMDs used to compute the feature vector for each time series in nonlinear test data with  $X \rightarrow Y$  (bottom left),  $X \leftarrow Y$  (bottom right), and *No Causation* (top).

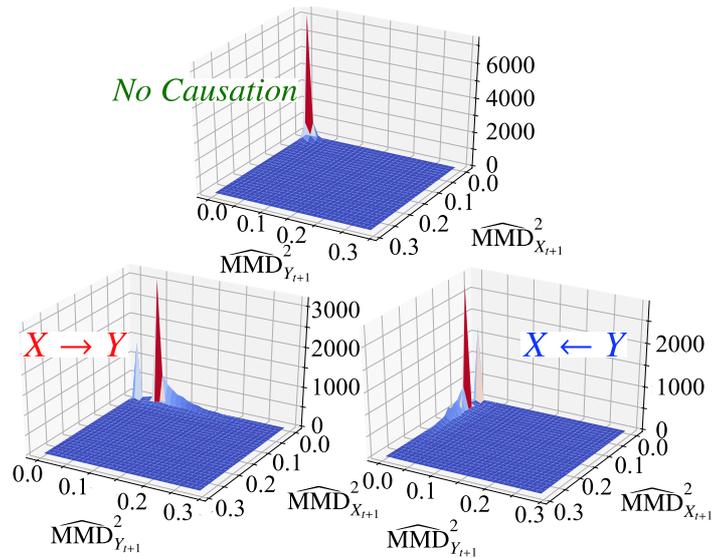


図 4 ラベル  $X \rightarrow Y$  (左下),  $X \leftarrow Y$  (右下), and *No Causation* (上) が割り当てられた, 線形なテストデータ中の個々の時系列より得た, MMD のペアに関するヒストグラム

Fig. 4 Histogram of MMDs used to compute the feature vector for each time series in linear test data with  $X \rightarrow Y$  (bottom left),  $X \leftarrow Y$  (bottom right), and *No Causation* (top).

データセットを用意した:

- 第 1 のテストデータセットは, 真の因果関係が既知のデータを集めたデータベース Cause-Effect Pairs [11] よりダウンロードした, 5 ペアの時系列データからなる. たとえば, *River Runoff* は平均降水量  $X$  と平均河川流量  $Y$  に関する 2 変数時系列データであり, 真の因果関係は  $X \rightarrow Y$  であるとされている [11].
- 第 2 のテストデータセットは, 上記 5 ペアの時系列の

それぞれから部分時系列を得ることで用意した. そのために, 各時系列を分割し, 長さ  $T = 200$  の部分時系列を複数用意した.

訓練データとしては, 人工データ実験とまったく同様にして用意した人工データを用いた.

表 1 にそれぞれのテストデータセットを用いた際の実験結果を示す. ここで, 表 1 上において, **RCC** の結果を省略した. これは, **RCC** が異なる訓練データを用いて行っ

表 1 第 1 のテストデータセットにおける推定結果の正誤 (上：ここで  $\checkmark$ ,  $\times$  はそれぞれ推定結果が正, 誤であることを表している), および第 2 のテストデータにおける推定精度 (下：提案手法と **RCC** に関しては, 異なる訓練データを用いて 20 回行った実験における, 推定精度の平均値と標準偏差値を示している)

Table 1 Causal relationships inferred from the first test dataset (top;  $\checkmark$  and  $\times$  denote correct and incorrect results, respectively) and test accuracies for the second test dataset (bottom; Means and standard deviations are shown for our method and **RCC** based on 20 runs).

|                                       | SIGC         | GC <sub>VAR</sub> | GC <sub>GAM</sub> | GC <sub>KER</sub> | TE           |
|---------------------------------------|--------------|-------------------|-------------------|-------------------|--------------|
| <i>River Runoff</i><br>( $T = 432$ )  | $\checkmark$ | $\checkmark$      | $\checkmark$      | $\times$          | $\checkmark$ |
| <i>Temperature</i><br>( $T = 16382$ ) | $\checkmark$ | $\times$          | $\checkmark$      | $\checkmark$      | $\times$     |
| <i>Radiation</i><br>( $T = 8401$ )    | $\checkmark$ | $\checkmark$      | $\checkmark$      | $\checkmark$      | $\checkmark$ |
| <i>Internet</i><br>( $T = 498$ )      | $\checkmark$ | $\checkmark$      | $\times$          | $\times$          | $\checkmark$ |
| <i>Sun Spots</i><br>( $T = 1632$ )    | $\checkmark$ | $\times$          | $\times$          | $\times$          | $\checkmark$ |

|                                      | SIGC                    | RCC              | GC <sub>VAR</sub> | GC <sub>GAM</sub> | GC <sub>KER</sub> | TE    |
|--------------------------------------|-------------------------|------------------|-------------------|-------------------|-------------------|-------|
| <i>River Runoff</i><br>( $T = 200$ ) | <b>0.958</b><br>(0.058) | 0.399<br>(0.193) | 0.684             | 0.406             | 0.155             | 0.485 |
| <i>Temperature</i><br>( $T = 200$ )  | <b>0.961</b><br>(0.011) | 0.432<br>(0.242) | 0.950             | 0.848             | 0.234             | 0.492 |
| <i>Radiation</i><br>( $T = 200$ )    | <b>0.987</b><br>(0.053) | 0.515<br>(0.345) | 0.156             | 0.0               | 0.782             | 0.394 |
| <i>Internet</i><br>( $T = 200$ )     | <b>1.0</b><br>(0.0)     | 0.478<br>(0.222) | 0.157             | 0.387             | 0.261             | 0.498 |
| <i>Sun Spots</i><br>( $T = 200$ )    | <b>1.0</b><br>(0.0)     | 0.435<br>(0.182) | 0.908             | 0.704             | 0.076             | 0.522 |

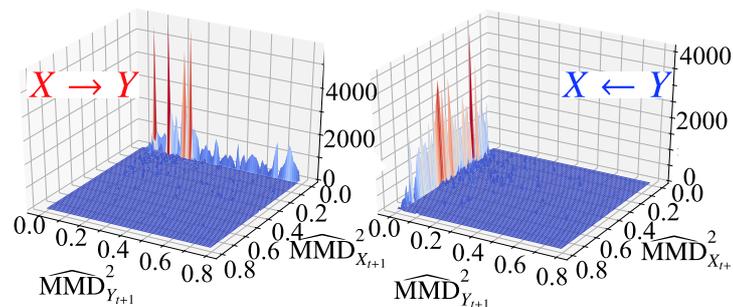


図 5 ラベル  $X \rightarrow Y$  (左),  $X \leftarrow Y$  (右) が割り当てられた, 第 2 のテストデータセット中の個々の実時系列データより得た, MMD のペアに関するヒストグラム

Fig. 5 Histogram of MMDs used to compute the feature vector for each time series in the second test dataset with  $X \rightarrow Y$  (left) and  $X \leftarrow Y$  (right).

た 20 回の実験において, 各回でまったく異なる推定結果を出力したためである. これに対し, 提案手法は 20 回の実験において, つねに同じ推定結果を出力した. 表 1 から分かるように, 提案手法は, 時系列の長さ  $T$  によらず, 他の既存手法より高い推定精度を達成した.

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すことを確認するため, 図 3, および図 4 と同様

の検証実験を行った.

具体的には, 上述した第 2 のテストデータセットに含まれる 2 変数時系列データに関して, 個々の時系列から得た MMD のペア  $\{d_t\}$  をヒストグラムとして可視化し, その結果を図 5 に示した. ここで, 図 5 左は, 真の因果関係が  $X \rightarrow Y$  である *River Runoff*, *Temperature*, *Radiation* より得た時系列データに関する結果であり, 図 5 右は, 真

表 2 3 変数人工時系列データを用いた際の推定精度 (提案手法と **RCC** に関しては, 異なる訓練データを用いて 10 回行った実験における, 推定精度の平均値と標準偏差値を示している.)

Table 2 Test accuracies for trivariate synthetic time series data. Means and standard deviations are shown for our method and **RCC** based on 10 runs with different training data.

|               | 3 変数データに対応した推定手法          |                         |                         | 2 変数データを対象とした推定手法        |                         |           | i.i.d. データを対象とした推定手法 |
|---------------|---------------------------|-------------------------|-------------------------|--------------------------|-------------------------|-----------|----------------------|
|               | <b>SIGC<sub>tri</sub></b> | <b>GC<sub>VAR</sub></b> | <b>GC<sub>KER</sub></b> | <b>SIGC<sub>bi</sub></b> | <b>GC<sub>GAM</sub></b> | <b>TE</b> | <b>RCC</b>           |
| test accuracy | 1.0<br>(0.0)              | 0.72                    | 1.0                     | 0.0<br>(0.0)             | 0.0                     | 0.0       | 0.0<br>(0.0)         |

の因果関係が  $X \leftarrow Y$  である *Internet*, *Sun Spots* より得た時系列データに関する結果である.

図 5 より, 上記の実データの場合でも, causal label に応じて十分異なる MMD が得られていることが分かる.

### 5.3 多変数時系列を用いた実験

変数の数  $n$  が  $n \geq 3$  であるような時系列データを用いて, 提案手法の性能を評価した. 訓練データとしては, 5.2 節での実験と同様にして生成した 3 変数の時系列データからなる人工データを用いた.

#### 5.3.1 人工データ実験

初めに, 3 変数の非線形な人工データを用いて, 提案手法の性能を評価した. テストデータは, 以下の式で表される three logistic map [21] に基づいて, 生成した.

$$\begin{aligned} X_t &= 0.8(1 - aX_{t-1}^2) + 0.2(1 - aY_{t-1}^2) + sE_{X_t} \\ Y_t &= 1 - aY_{t-1}^2 + sE_{Y_t} \\ Z_t &= 0.8(1 - aZ_{t-1}^2) + 0.2(1 - aX_{t-1}^2) + sE_{Z_t} \end{aligned} \quad (19)$$

ここで,  $a = 1.8$ ,  $s = 0.01$  とし, ノイズ変数  $E_{X_t}$ ,  $E_{Y_t}$ ,  $E_{Z_t}$  は, 標準正規分布  $\mathcal{N}(0, 1)$  よりサンプリングした. 初期値  $X_1, Y_1, Z_1$  を一様分布  $\mathcal{U}(0, 1)$  からサンプルして 100 通り用意し, 時系列の長さを  $T = 1,000$  とし, 100 種類の 3 変数時系列データを用意した. 式 (19) から分かるように, 変数のペア  $(X, Y)$ ,  $(Z, X)$  の, それぞれの間の真の causal label は,  $Y \rightarrow X$ ,  $X \rightarrow Z$  であり, ペア  $(Y, Z)$  間には Granger causality はない (i.e., *No Causation*).

テストデータにおける各手法の推定精度を表 2 に示す. ここで, 推定精度とは, 100 種類の時系列データのうちの, すべての変数のペア (すなわち  $(X, Y)$ ,  $(Y, Z)$ ,  $(Z, X)$ ) の間の Granger causality の有無・方向を正しく推定できた時系列データの占める割合を表す. また表 2 において, **SIGC<sub>tri</sub>** とは, 3.3 節で述べた, 3 変数時系列データに対して定義された特徴量表現を用いた提案手法を表し, **SIGC<sub>bi</sub>** は, 3.2 節で述べた, 2 変数時系列データに対して定義された特徴量表現を用いた提案手法を表す.

表 2 から, 3 変数データに対応した推定手法では高い精度で推定できているが, 2 変数データを対象とした推定

手法では推定精度に難があることが分かる. これは, 2 変数データを対象とした推定手法の場合, 変数のペア  $(Y, Z)$  間に Granger causality があると誤推定したためである. 3.3.1 項で述べたように, 変数  $Y$  が第 3 の変数  $X$  を介して  $Z$  に影響を与える場合, 定義 1 に基づく 2 変数データを対象とした Granger causality の推定手法では第 3 の変数  $X$  の影響を考慮できないため,  $Y$  が  $Z$  の原因であると誤推定してしまうと考えられる. しかし, 条件付き Granger causality に基づく 3 変数データに対応した推定手法の場合, 第 3 の変数  $X$  の影響を考慮できるため, このような誤推定は生じないと期待できる.

表 2 から分かるように, 提案手法 **SIGC<sub>tri</sub>** は, 条件付き Granger causality に基づく手法である **GC<sub>KER</sub>** と同様, すべての時系列に対して正しく Granger causality を推定し, **SIGC<sub>bi</sub>** に比べ, 十分高い精度を示した. このことから **SIGC<sub>tri</sub>** では, 3 変数時系列データを対象とした特徴量表現によって, 第 3 の変数の影響をうまく考慮できていると考えられる.

#### 5.3.2 実データ実験

実データを用いて, 提案手法の性能を評価した. テストデータとしては, 次のような時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 [26] 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に関するデータベース KEGG [15] に基づいて決定した.

本実験では, 因果関係のない遺伝子のペアの数のほうが因果関係のある遺伝子のペアの数よりはるかに大きいため, 各手法の性能を test accuracy ではなく, macro 平均 F 値および micro 平均 F 値に基づいて評価した. ここで, F 値とは, 適合率 (precision) と再現率 (recall) の調和平均で定義される値であり, macro 平均 F 値とは F 値を各クラスごとに計算しその平均をとったもの, micro 平均 F 値とはクラスの別に関係なく全事例に関して F 値を求めたものである.

表 3 マイクロアレイデータを用いた際の macro 平均 F 値, micro 平均 F 値 (提案手法と RCC に関しては, 異なる訓練データを用いて 10 回行った実験における, 平均値と標準偏差値を示している)

Table 3 Macro and micro-averaged F scores. Means and standard deviations are shown for our methods and RCC based on 10 runs.

|                        | SIGC <sub>tri</sub>   | GC <sub>VAR</sub> | GC <sub>KER</sub> | SIGC <sub>bi</sub> | GC <sub>GAM</sub> | TE    | RCC              |
|------------------------|-----------------------|-------------------|-------------------|--------------------|-------------------|-------|------------------|
| macro-averaged F-score | <b>0.483</b><br>(0.0) | 0.351             | 0.437             | 0.431<br>(0.007)   | 0.457             | 0.430 | 0.407<br>(0.096) |
| micro-averaged F-score | <b>0.637</b><br>(0.0) | 0.436             | 0.513             | 0.578<br>(0.011)   | 0.567             | 0.449 | 0.567<br>(0.161) |

表 3 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いえない結果となった. しかし, 提案した SIGC<sub>tri</sub> は Granger causality の既存手法, および SIGC<sub>bi</sub> に比べ, 高い推定精度を達成した.

## 6. まとめ

本論文では, Granger causality を同定する問題に対し, 教師あり学習に基づく新たなアプローチを提案した. Granger causality を同定する問題を教師あり学習の問題として解くために, 過去の値で条件付けられた条件付き分布間の距離に基づく特徴量表現を用いて, 分類の特徴ベクトルを得ることを提案した. この特徴量表現により, 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, 分類に基づくアプローチの有効性を示唆している.

提案手法は, 人工データ・実データを用いた比較実験において, 既存手法より高い推定精度を実現した. 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適合するか否かによって, 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. また提案手法は, i.i.d. データを対象とした分類に基づく既存の因果推論手法 RCC より高い推定精度を示した. この結果は, 提案した特徴量表現の有効性を示唆するものである.

さらに本論文では, 多変数時系列からも Granger causality を推定できるよう, 提案手法を拡張する方法についても述べ, 人工データ・実データを用いた実験により, 拡張した提案手法の有効性を確認した.

今後の展望としては, 複雑な実データへの対応があげられる. 具体的には, 非定常時系列への対応が考えられる. 非定常時系列の場合, 1) Granger causality の有無・方向が変化しない場合, 2) Granger causality の有無・方向が時間とともに変化する場合, の大きく 2 つの場合を考える必要があり, 特に後者の場合への対応が重要である. 前者に関しては, 文献 [30] にあるように, 特徴量を得る際に用いた既存手法 KKF-CEO は, 非定常な時系列に対しても分

布に対するカーネル平均を精度よく推定できるとされており, 現時点の提案手法でも十分適用可能であると考えられる. 一方後者の場合については, 提案手法では定義 1 時系列全体を通して Granger causality の有無・方向が一定であることを仮定しているため, このような場合には対応していない. このような複雑な問題設定に対応するための, 提案手法のさらなる拡張が今後の課題である.

また, 図 3, 図 4, 図 5 に示した結果では, 提案した特徴量表現によって, その 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が確認できたが, 任意の時系列データに対してこのような形で特徴ベクトルが得られる保証はない. どのような時系列であれば Granger causality の有無・方向に応じた特徴ベクトルが得られるかを理論・実験の両面から調査することも, 今後の課題として非常に重要である.

## 参考文献

- [1] Bell, D., Kay, J. and Malley, J.: A non-parametric approach to non-linear causality testing, *Economics Letters*, Vol.51, No.1, pp.7-18 (1996).
- [2] Bishop, C.M.: *Pattern Recognition and Machine Learning*, Springer (2006).
- [3] Bontempi, G. and Flauder, M.: From dependency to causality: a machine learning approach, *JMLR*, Vol.16, pp.2437-2457 (2015).
- [4] Chen, M. and An, H.Z.: A Kolmogorov-Smirnov type test for conditional heteroskedasticity in time series, *Statistics & probability letters*, Vol.33, No.3, pp.321-331 (1997).
- [5] Cheng, D., Bahadori, M.T. and Liu, Y.: FBLG: A simple and effective approach for temporal dependence discovery from time series data, *KDD*, pp.382-391 (2014).
- [6] Chikahara, Y. and Fujino, A.: Causal inference in time series via supervised learning, *IJCAI* (2018) (To appear).
- [7] Geweke, J.F.: Measures of conditional linear dependence and feedback between time series, *Journal of the American Statistical Association*, Vol.79, No.388, pp.907-915 (1984).
- [8] Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods, *Econometrica: Journal of the Econometric Society*, pp.424-438 (1969).
- [9] Gretton, A., Borgwardt, K.M., Rasch, M., Schölkopf, B. and Smola, A.J.: A kernel method for the two-sample

problem, *NIPS*, pp.513-520 (2007).

[10] Guyon, I. ChaLearn cause-effect pair challenge (2013), available from <https://www.kaggle.com/c/cause-effect-pairs/>.

[11] Jakob, Z.: Database with cause-effect pairs, available from <https://webdav.tuebingen.mpg.de/cause-effect/>.

[12] Janzing, D. and Schölkopf, B.: Causal inference using the algorithmic Markov condition, *IEEE Trans. Information Theory*, Vol.56, No.10, pp.5168-5194 (2010).

[13] Kalman, R.E.: A new approach to linear filtering and prediction problems, *Journal of basic Engineering*, Vol.82, No.1, pp.35-45 (1960).

[14] Kar, M., Nazhoğlu, Ş. and Ağır, H.: Financial development and economic growth nexus in the MENA countries: Bootstrap panel granger causality analysis, *Economic modelling*, Vol.28, No.1, pp.685-693 (2011).

[15] KEGG: Kyoto Encyclopedia of Genes and Genomes (1995), available from <https://www.genome.jp/kegg/>.

[16] Kullback, S. and Leibler, R.A.: On information and sufficiency, *The Annals of Mathematical Statistics*, Vol.22, No.1, pp.79-86 (1951).

[17] Lopez-Paz, D., Muandet, K., Schölkopf, B. and Tolstikhin, I.: Towards a learning theory of cause-effect inference, *ICML*, pp.1452-1461 (2015).

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

[19] Marinazzo, D., Pellicoro, M. and Stramaglia, S.: Kernel-Granger causality and the analysis of dynamical networks, *Physical Review E*, Vol.77, No.5, 056215 (2008).

[20] Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al.: Kernel mean embedding of distributions: A review and beyond, *Foundations and Trends® in Machine Learning*, Vol.10, No.1-2, pp.1-141 (2017).

[21] Ott, E.: *Chaos in dynamical systems*, Cambridge University Press (2002).

[22] Rahimi, A. and Recht, B.: Random features for large-scale kernel machines, *NIPS*, pp.1177-1184 (2007).

[23] Reichenbach, H.: *The Direction of Time*, University of California Press (1956).

[24] Schölkopf, B. and Smola, A.J.: *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press (2001).

[25] Schreiber, T.: Measuring information transfer, *Physical Review Letters*, Vol.85, No.2, pp.461 (2000).

[26] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, Vol.9, No.12, pp.3273-3297 (1998).

[27] Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B. and Lanckriet, G.R.: Hilbert space embeddings and metrics on probability measures, *JMLR*, Vol.11, pp1517-1561 (2010).

[28] Sun, X.: Assessing nonlinear Granger causality from multivariate time series, *ECML*, pp.440-455, Springer (2008).

[29] Yao, S., Yoo, S. and Yu, D.: Prior knowledge driven Granger causality analysis on gene regulatory network discovery, *BMC Bioinformatics*, Vol.16, No.1, pp.273 (2015).

[30] Zhu, P., Chen, B. and Principe, J.C.: Learning nonlinear generative models of time series with a Kalman filter in RKHS, *IEEE Trans. Signal Processing*, Vol.62, No.1,

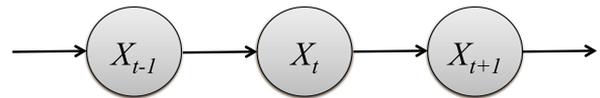


図 A.1 1次マルコフモデルのグラフィカルモデル

Fig. A.1 Graphical model for first-order Markov model. pp.141-155 (2014).

## 付 録

### A.1 過去の観測のすべてで条件付けられた分布に基づく時系列予測

本文 3.2.2 項で、観測  $S_X = \{x_1, \dots, x_t\}$  に対して予測を行う際、状態空間モデルを用いれば、過去の観測  $S_X$  のすべてで条件付けられた分布  $P(X_{t+1}|S_X)$  に関する期待値  $E_{X_{t+1}|S_X}[X_{t+1}]$  を推定できると述べた。

本章では、まず時系列予測に用いる基本的な生成モデルである 1 次マルコフモデルでは、このような期待値が推定できないことを述べ、次になぜ状態空間モデルを用いるとそのような期待値が推定できるかについて述べる。

#### A.1.1 1次マルコフモデル

時系列予測を行うための最も単純な生成モデルの 1 つとして、1 次マルコフモデルがある。1 次マルコフモデルは、ある時刻の観測は、その 1 つ前の時刻の観測のみに依存するという形で、観測変数間の関係を表した生成モデルである。1 次マルコフモデルをグラフィカルモデル (i.e., 確率変数をノード、確率変数間の依存関係をエッジとして表したグラフ) として表すと図 A.1 のようになる。

図 A.1 より分かるように、たとえば観測変数  $X_{t+1}$  は、その 1 つ前の時刻の観測変数  $X_t$  で条件付けると、それより以前のすべての観測変数 (たとえば  $X_{t-1}$ ) と独立であり、

$$X_{t+1} \perp X_{t-1} | X_t$$

が成立する。このような条件付き独立の関係があるため、過去の観測すべてで条件付けた分布  $P(X_{t+1}|x_t, \dots, x_1)$  を 1 次マルコフモデルに基づいて表すと、

$$P(X_{t+1}|x_t, \dots, x_1) = P(X_{t+1}|x_t)$$

となり、 $X_t$  でのみ条件付けた分布に一致する。したがって、1 次マルコフモデルに基づいて  $P(X_{t+1}|S_X)$  に関する期待値  $E_{X_{t+1}|S_X}[X_{t+1}]$  を推定しても、

$$E_{X_{t+1}|S_X}[X_{t+1}] = E_{X_{t+1}|x_t}[X_{t+1}]$$

となり、過去の観測すべてで条件付けられた分布に関する期待値とはならない。

#### A.1.2 状態空間モデル

幅広い時系列データの予測を考えるうえで、1 つ前の観

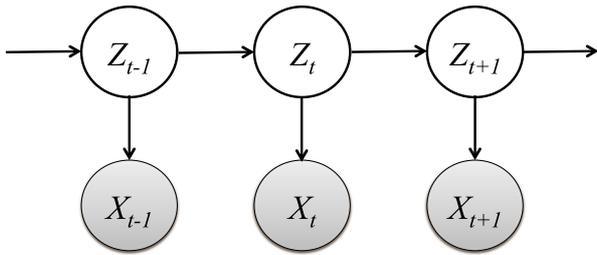


図 A.2 状態空間モデルのグラフィカルモデル (灰色の頂点は観測変数, 白色の頂点は隠れ変数を表す)

Fig. A.2 Graphical model for a state-space model (Gray nodes are observed variables while white ones are hidden variables).

測のみに依存するという1次マルコフモデルの仮定は、非常に強い制約である。この仮定を緩和するには、たとえば2つ前までの観測に依存する2次マルコフモデルを考えるなど、高次マルコフモデルを導入することが考えられるが、高次マルコフモデルはパラメータの数が多く、一般に学習しにくいという問題がある。このような問題を解決するために提案された生成モデルが、状態空間モデルである [2]。

状態空間モデルでは、各時刻の観測変数  $X_t$  に対し、対応する隠れ変数  $Z_t$  を導入し、隠れ変数どうしの関係は (1次) マルコフモデルで表されるが、観測変数の間にはマルコフ性は成立しないというモデルを考える。これをグラフィカルモデルとして表すと、図 A.2 のようになる。

図 A.2 から分かるように、隠れ変数  $Z_{t+1}$  は  $Z_t$  で条件付けると  $Z_{t-1}$  と独立であり、

$$Z_{t+1} \perp\!\!\!\perp Z_{t-1} | Z_t$$

が成立する。しかし、文献 [2] にあるように、観測変数の間に条件付き独立の関係は存在せず、 $X_{t+1}$  は過去のすべての観測に依存する<sup>\*16</sup>。したがって、状態空間モデルを用いれば、過去の観測すべてで条件付けられた分布  $P(X_{t+1}|S_X)$  に関する期待値  $E_{X_{t+1}|S_X}[X_{t+1}]$  を推定できる。

## A.2 カーネル平均

本文 3.2.3 項において、既存手法 KKF-CEO を用いた式 (12), 式 (11) のカーネル平均の推定方法について述べたが、本章では、その導出において必要となる、カーネル平均に関する基本事項について簡単に述べる。

本章では、まず A.2.1 節において、カーネル平均の基本的な定義・推定方法について述べる。次に、A.2.2 節で、条件付き分布のカーネル平均および conditional embedding operator について述べる。最後に、A.2.3 節で、同時分布のカーネル平均と積カーネルの関係について述べる。

<sup>\*16</sup> この事実は、文献 [2] にあるように、有向分離 (d 分離) と呼ばれる、グラフィカルモデルのグラフ構造に基づいて、確率変数間の (条件付き) 独立性を判断する方法を用いると容易に確認できる。

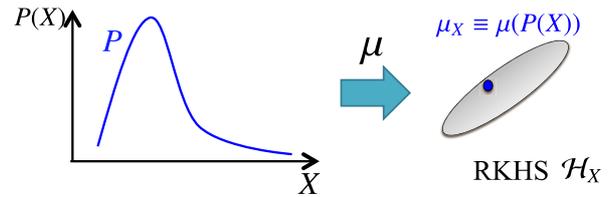


図 A.3 分布  $P(X)$  はカーネル平均  $\mu$  によって  $\text{RKHS } \mathcal{H}_X$  上の点  $\mu_X$  に写像される

Fig. A.3 By using kernel mean embedding  $\mu$ , distribution  $P(X)$  is mapped to a point in  $\text{RKHS } \mathcal{H}_X$ , i.e.,  $\mu_X$ .

### A.2.1 分布 $P(X)$ のカーネル平均

カーネル平均とは、個々の確率分布を、RKHS と呼ばれるカーネル関数が定義する特徴空間上の点として写像する関数である。 $\mathcal{X}$  上の確率変数  $X$  を考える。このとき、分布  $P(X)$  は、カーネル平均により、図 A.3 のように  $\text{RKHS } \mathcal{H}_X$  上の点  $\mu_X$  として写像される。ここで  $\mathcal{H}_X$  は、カーネル関数  $k_X$  によって定義される RKHS である。この写像は、 $k_X(x, x') = \langle \Phi_X(x), \Phi_X(x') \rangle_{\mathcal{H}_X}$  を満たす、特徴写像と呼ばれる関数  $\Phi_X$  を用いて、

$$\mu_X = \mu(P(X)) \equiv E_X[\Phi_X(X)] \tag{A.1}$$

として定義される。

カーネル平均  $\mu_X$  の推定量は、分布  $P(X)$  に従うサンプル  $\{x_1, \dots, x_n\}$  を用いれば、標本平均の形で、

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n \Phi_X(x_i) \tag{A.2}$$

として表される。

### A.2.2 条件付き分布 $P(X|y)$ のカーネル平均と conditional embedding operator

$\mathcal{X}, \mathcal{Y}$  上の確率変数  $X, Y$  を考え、 $\mathcal{X}, \mathcal{Y}$  上のカーネル関数  $k_X, k_Y$  によって定義される特徴写像を  $\Phi_X, \Phi_Y$ , RKHS を  $\mathcal{H}_X, \mathcal{H}_Y$  とする。

このとき、ある値  $Y = y$  で条件付けた分布  $P(X|y)$  は、 $\text{RKHS } \mathcal{H}_X$  上に写像され、

$$\mu_{X|y} \equiv E_{X|y}[\Phi_X(X)] \tag{A.3}$$

として定義される。

式 (A.3) から分かるように、写像先の点  $\mu_{X|y}$  は、分布を条件づける値  $y$  によって異なる。これは、 $y$  を特徴写像  $\Phi_Y$  で写像した点  $\Phi_Y(y)$  を考えれば、図 A.4 のように、 $\text{RKHS } \mathcal{H}_Y$  上の点  $\Phi_Y(y)$  を  $\mathcal{H}_X$  上の点  $\mu_{X|y}$  に変換する写像

$$\mu_{X|y} = C_{X|Y} \Phi_Y(y) \tag{A.4}$$

として表せる。ここで、 $C_{X|Y}$  は conditional embedding operator [20] と呼ばれる作用素で、その推定量は、同時分布  $P(X, Y)$  に従うサンプル  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  を用い

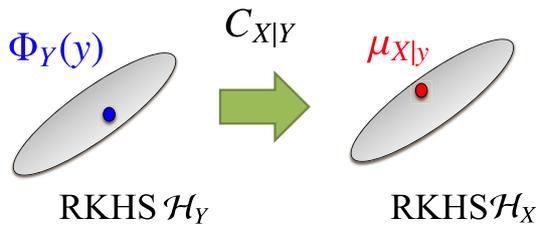


図 A.4  $\mathcal{H}_Y$  上の点  $\Phi_Y(y)$  は conditional embedding operator  $C_{X|Y}$  によって  $\mathcal{H}_X$  上の点  $\mu_{X|y}$  に写像される

Fig. A.4 By using conditional embedding operator  $C_{X|Y}$ , point in  $\mathcal{H}_Y$  (denoted as  $\Phi_Y(y)$ ) is mapped to the point  $\mu_{X|y}$  in  $\mathcal{H}_X$ .

て、以下のように表されることが知られている。

$$\hat{C}_{X|Y} = \Lambda(K + n\lambda I_n)^{-1} \Upsilon^\top \quad (\text{A.5})$$

ここで、行列  $K$  はグラム行列で  $K_{i,j} = k_Y(y_i, y_j)$ ,  $\lambda$  は正のパラメータ,  $I_n$  は  $n \times n$  の単位行列,  $\Lambda$  と  $\Upsilon$  は行列で、それぞれ  $\Lambda = [\Phi_X(x_1), \dots, \Phi_X(x_n)]$ ,  $\Upsilon = [\Phi_Y(y_1), \dots, \Phi_Y(y_n)]$  である。

式 (A.4) および式 (A.5) より、カーネル平均  $\mu_{X|y}$  は、

$$\hat{\mu}_{X|y} = \sum_{i=1}^n c_i \Phi_X(x_i) \quad (\text{A.6})$$

where  $\mathbf{c} = (K + n\lambda I_n)^{-1} \mathbf{k}_y$

として表される。ここで、 $\mathbf{k}_y = [\Phi_Y(y_1), \dots, \Phi_Y(y_n)]^\top \Phi_Y(y)$  である。式 (A.6) より、分布を条件付ける値  $y$  によって、重みベクトル  $\mathbf{c}$  が変化することが分かる。

### A.2.3 同時分布 $P(X, Y)$ のカーネル平均

同時分布  $P(X, Y)$  のカーネル平均は、積カーネルと呼ばれるカーネル関数に基づいて定義される。カーネル関数  $k_X, k_Y$  の積カーネルは、両者の積として定義され、

$$\begin{aligned} k_{XY}((x, y), (x', y')) &= k_X(x, x') \cdot k_Y(y, y') \\ &= \langle \Phi_X(x) \otimes \Phi_Y(y), \Phi_X(x') \otimes \Phi_Y(y') \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \end{aligned}$$

と表される。ここで記号  $\otimes$  はテンソル積を表す。

同時分布  $P(X, Y)$  のカーネル平均は、積カーネルによって定義される特徴写像  $\Phi_X \otimes \Phi_Y$  を用いて、

$$\mu_{X,Y} \equiv E_{X,Y}[\Phi_X(X) \otimes \Phi_Y(Y)]$$

として表され、その推定量は

$$\hat{\mu}_{X,Y} = \frac{1}{n} \sum_{i=1}^n \Phi_X(x_i) \otimes \Phi_Y(y_i) \quad (\text{A.7})$$

と表される。



近原 鷹一

2013 年慶應義塾大学理工学部生命情報学科卒業。2015 年東京大学大学院情報理工学系研究科コンピュータ科学専攻修士課程修了。同年日本電信電話株式会社 (NTT) 入社。以降、NTT コミュニケーション科学基礎研究所に

て、機械学習、知識発見に関する研究に従事。



藤野 昭典 (正会員)

1995 年京都大学工学部精密工学科卒業。1997 年同大学大学院修士課程修了。2009 年同大学院博士課程修了。博士 (情報学)。1997 年日本電信電話株式会社 (NTT) 入社。以降、機械学習、データマイニング等の研究に従事。

現在、NTT コミュニケーション科学基礎研究所主幹研究員。電子情報通信学会 PRMU 研究奨励賞 (2004 年度)、FIT 論文賞 (2005 年) 等受賞。電子情報通信学会、IEEE 各会員。