Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Causal Inference in Time Series via Supervised Learning

Yoichi Chikahara and **Akinori Fujino**
NTT Communication Science Laboratories, Kyoto 619-0237, Japan
chikahara.yoichi@lab.ntt.co.jp, fujino.akinori@lab.ntt.co.jp

Abstract

Causal inference in time series is an important problem in many fields. Traditional methods use regression models for this problem. The inference accuracies of these methods depend greatly on whether or not the model can be well fitted to the data, and therefore we are required to select an appropriate regression model, which is difficult in practice. This paper proposes a supervised learning framework that utilizes a classifier instead of regression models. We present a feature representation that employs the distance between the conditional distributions given past variable values and show experimentally that the feature representation provides sufficiently different feature vectors for time series with different causal relationships. Furthermore, we extend our framework to multivariate time series and present experimental results where our method outperformed the model-based methods and the supervised learning method for i.i.d. data.

1 Introduction

Discovering temporal causal directions is an important task in time series analysis and has key applications in various fields. For instance, finding the causal direction that indicates that the research and development (R&D) expenditure X influences the total sales Y, but not vice versa, is helpful for decision making in companies. In addition, identifying causal (regulatory) relationships between genes from time series gene expression data is one of the most important topics in bioinformatics.

As a definition of temporal causality, Granger causality [Granger, 1969] is widely used [Kar *et al.*, 2011; Yao *et al.*, 2015]. According to its definition, the variable X is the cause of the variable Y if the past values of X are *helpful* in predicting the future value of Y.

Traditional methods for identifying Granger causality use regression models [Bell *et al.*, 1996; Cheng *et al.*, 2014; Granger, 1969; Marinazzo *et al.*, 2008; Sun, 2008] such as the vector autoregressive (VAR) model and the generalized additive models (GAM). With these methods, we can determine that X is the cause of Y if the prediction errors of Y based only on its past values are significantly reduced by additionally using the past values of X. When the regression model can be well fitted to the data, we can infer correct causal directions. However, in practice, selecting an appropriate regression model for each time series data is difficult and requires a deep understanding of the data analysis. Therefore, it is not easy to identify correct causal directions with these model-based methods.

The goal of this paper is to build an approach to causal inference in time series that does not require a deep understanding of the data analysis. To realize this goal, we propose a supervised learning framework that utilizes a classifier instead of regression models. Specifically, we propose solving the problem of Granger causality identification by ternary classification, in other words, by training a classifier that assigns ternary *causal labels* ($X \to Y$, $X \leftarrow Y$, or *No Causation*) to time series. In fact, several methods have already been proposed that perform classification to infer causal relationships from i.i.d. data, which have worked well experimentally [Bontempi and Flauder, 2015; Guyon, 2013; Lopez-Paz *et al.*, 2015; 2017]. To solve causal inference in time series via classification, we formulate a feature representation that provides sufficiently different feature vectors for time series with different causal relationships. The idea for obtaining such feature vectors is founded on the definition of Granger causality: X is the cause of Y if the following two conditional distributions of the future value of Y are different; one is given the past values of Y and the other is given the past values of X and Y. To build the classifier for Granger causality identification, we utilize the distance between these distributions when preparing feature vectors. To compute the distance, by using *kernel mean embedding*, we map each distribution to a point in the feature space called the reproducing kernel Hilbert space (RKHS) and measure the distance between the points, which is termed the *maximum mean discrepancy* (MMD) [Gretton *et al.*, 2007].

In experiments, our method sufficiently outperformed the model-based Granger causality methods and the supervised learning method for i.i.d. data by using the same feature representation and the same classifier. Furthermore, we describe how our approach can be extended to multivariate time series and show experimentally that feature vectors have a sufficient difference that depends on Granger causality, which demonstrates the effectivity of our proposed framework.

2 Granger Causality

Granger causality defines X as the cause of Y if the past values of X contain *helpful* information for predicting the future value of Y. Formally, it is defined as follows:

Definition 1 (Granger causality[Granger, 1969]) *Suppose we have a stationary sequence of random variables $\{(X_t, Y_t)\}$ ($t \in \mathbb{N}$), where X_t and Y_t are on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let S_X and S_Y be observations of $\{X_1, \cdots, X_t\}$ and $\{Y_1, \cdots, Y_t\}$, respectively.*

Granger causality defines $\{X_t\}$ as the cause of $\{Y_t\}$ if

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

and states that $\{X_t\}$ is not the cause of $\{Y_t\}$ if

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

To see if the two conditional distributions $P(Y_{t+1}|S_X, S_Y)$ and $P(Y_{t+1}|S_Y)$ are identical, traditional methods [Bell *et al.*, 1996; Granger, 1969; Marinazzo *et al.*, 2008; Sun, 2008] use statistical testing to determine whether or not the two conditional means $E[Y_{t+1}|S_X, S_Y]$ and $E[Y_{t+1}|S_Y]$ are equal, which is a much simpler problem than (1). For instance, in [Granger, 1969], the conditional means are represented by using the (V)AR model to compute the test statistic based on the prediction errors.

To represent the conditional means, these methods require us to use an appropriate regression model that can be well fitted to the data; however, such a model is difficult to select in practice. For this problem, we propose a novel approach that utilizes a classifier instead of regression models.

3 Proposed Method

3.1 Classification Setup

Suppose the training data are N pairs of bivariate time series $S^1, \cdots, S^N$, where each time series S^j with the fixed length T_j consists of the observations of random variables $\{(X_1^j, Y_1^j), \cdots, (X_{T_j}^j, Y_{T_j}^j)\}$ ($j \in \{1, \cdots, N\}$). Here, each time series S^j has a causal label $l^j \in \{+1, -1, 0\}$ that indicates $X^j \rightarrow Y^j$, $X^j \leftarrow Y^j$, or *No Causation*, where $X^j = (X_1^j, \cdots, X_{T_j}^j)$ and $Y^j = (Y_1^j, \cdots, Y_{T_j}^j)$.

Using a function $\nu(\cdot)$ that maps S^j to a single feature vector, we first train a classifier with $\{(\nu(S^j), l^j)\}_{j=1}^N$. Then, the task of inferring the causal relationship in another time series S' (test data) can be rephrased as assigning the label to $\nu(S')$ by using the trained classifier.

Such a classification task can be extended to multivariate time series as detailed in Section 3.3.

3.2 Classifier Design

To build a classifier that assigns causal labels to time series, we formulate the feature representation $\nu(\cdot)$. In what follows, we describe our ideas for obtaining feature vectors that are sufficiently different depending on Granger causality.

Basic Ideas for Granger Causality Identification

Simply by using the definition of Granger causality (Definition 1)[1], for instance, we regard the causal label as $X \rightarrow Y$ if X is the cause of Y, and if Y is *not* the cause of X. Formally, we regard the causal label as [2]

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

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

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

To assign causal labels to time series based on (2), (3), and (4), it is necessary to determine whether or not the two conditional distributions are identical. To represent information about conditional distributions, instead of using regression models, we utilize kernel mean embedding. Kernel mean embedding maps a distribution to a point in the feature space called the RKHS. Interestingly, when a *characteristic* kernel (e.g., a Gaussian kernel) is used, the mapping is *injective*: different distributions are not mapped to the same point [Sriperumbudur *et al.*, 2010].

Suppose that kernel mean embedding maps conditional distributions $P(X_{t+1}|S_X, S_Y)$, $P(X_{t+1}|S_X)$ and $P(Y_{t+1}|S_X, S_Y)$, $P(Y_{t+1}|S_Y)$ to the points $\mu_{X_{t+1}|S_X,S_Y}$, $\mu_{X_{t+1}|S_Y} \in \mathcal{H}_X$ and $\mu_{Y_{t+1}|S_X,S_Y}$, $\mu_{Y_{t+1}|S_Y} \in \mathcal{H}_Y$, respectively, where $\mathcal{H}_X$ and $\mathcal{H}_Y$ are the RKHSs. Then, when using a characteristic kernel, (2), (3), and (4) can be written as

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

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

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

To assign labels based on (5), (6), and (7), we only have to determine whether or not two points in the RKHS are the same over time t or, equivalently, to determine whether or not the distance between the two points in the RKHS, which is termed the MMD [Gretton *et al.*, 2007], is zero over time t.

For this reason, to develop the classifier for Granger causality identification, in our feature representation, we utilize the MMDs, which are defined and estimated as follows.

Definition: Let k_X and k_Y be kernels on $\mathcal{X}$ and $\mathcal{Y}$, and

[1] Note that since our approach is founded on Definition 1, which cannot address the case where there are *latent confounders* (i.e., unobserved variables that influence both X and Y), as with the existing methods [Bell *et al.*, 1996; Cheng *et al.*, 2014; Granger, 1969; Marinazzo *et al.*, 2008; Sun, 2008], this paper does not deal with such a case.

[2] Although we do not consider the case where $P(X_{t+1}|S_X, S_Y) \neq P(X_{t+1}|S_X)$ and $P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$ (i.e., X is the cause of Y, and Y is also the cause of X), we can straightforwardly address such a case by adding an extra label.

$\mathcal{H}_X$ and $\mathcal{H}_Y$ be the RKHSs defined by k_X and k_Y, respectively. The MMD for the two distributions $P(X_{t+1}|S_X, S_Y)$ and $P(X_{t+1}|S_X)$ is simply defined as the distance between $\mu_{X_{t+1}|S_X,S_Y}, \mu_{X_{t+1}|S_X} \in \mathcal{H}_X$ as follows

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

Similarly, $\mathrm{MMD}^2_{Y_{t+1}}$ is defined as the distance between $\mu_{Y_{t+1}|S_X,S_Y}, \mu_{Y_{t+1}|S_Y} \in \mathcal{H}_Y$.

Estimation: The MMD can be estimated without using regression models and without performing a density estimation. At this point, the MMD is much more attractive than the Kolmogorov-Smirnov statistic [Chen and An, 1997] and the Kullback-Leibler divergence [Kullback and Leibler, 1951] since the former requires us to select regression models and the latter requires a density estimation, which is difficult when there are insufficient samples.

To estimate the MMD (8), we estimate the kernel mean embeddings of conditional distributions $\mu_{X_{t+1}|S_X,S_Y}$ and $\mu_{X_{t+1}|S_X}$. As detailed in e.g., [Muandet *et al.*, 2017], in general, the kernel mean embedding of the distribution is estimated by taking the weighted sum of the so-called *feature mapping* function. Specifically, when using the existing method called the kernel Kalman filter based on a conditional embedding operator (KKF-CEO) [Zhu *et al.*, 2014], we can estimate $\mu_{X_{t+1}|S_X,S_Y}$ and $\mu_{X_{T+1}|S_X}$ by the weighted sum of the feature mapping Φ_X:

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

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

where $\Phi_X(x_\tau) \equiv k_X(x_\tau, \cdot)$ is a feature mapping function[3], and $\mathbf{w^{XY}} = [w_2^{XY}, \cdots, w_{t-1}^{XY}]^\top$ and $\mathbf{w^{X}} = [w_2^{X}, \cdots, w_{t-1}^{X}]^\top$ $(t > 3)$ are the real-valued weight vectors.

To compute the weight vectors $\mathbf{w^X}$ and $\mathbf{w^{XY}}$, we employed KKF-CEO. In fact, KKF-CEO provides the algorithm needed to estimate $\mathbf{w^X}$ from the observations S_X for time series prediction. Therefore, we can compute $\mathbf{w^X}$ by directly employing KKF-CEO. To estimate $\mathbf{w^{XY}}$ from S_X and S_Y, we simply used KKF-CEO with the product kernel $k_X \cdot k_Y$.

Although computing weight vectors by KKF-CEO requires the setting of several hyperparameters, they can be appropriately set for each time series by minimizing the squared errors between observations and the values predicted by KKF-CEO.

Applying (9) and (10) to (8), $\mathrm{MMD}^2_{X_{t+1}}$ is estimated as

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

[3]For instance, when using the Gaussian kernel $k_X(x, x') = \exp(-\gamma\|x - x'\|^2)$ ($\gamma > 0$ is a parameter), the feature mapping becomes $\Phi_X(x) = \exp(-\gamma x^2)\,[1, \sqrt{2\gamma/1!}\,x, \sqrt{(2\gamma)^2/2!}\,x^2, \cdots]^\top$.

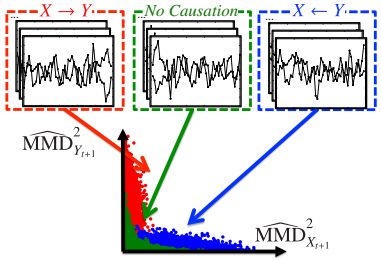

Figure 1: Different MMD pairs are estimated from time series with different causal labels. Each dot represents the MMD pair estimated from each time series.

Feature Representation

To build a classifier for Granger causality identification, we obtain the feature vectors by using the MMD pairs, where each pair $d_t = [\widehat{\mathrm{MMD}}^2_{X_{t+1}}, \widehat{\mathrm{MMD}}^2_{Y_{t+1}}]^\top$ is estimated by (11).

By using the MMD pairs, we can expect sufficiently different feature vectors to be obtained from time series with different causal labels. This is because whether or not the MMD becomes zero depends on the causal label as indicated by (5), (6), and (7). Although each MMD in d_t cannot become exactly zero since it is a finite sample estimate, we can expect sufficiently different MMD pairs to be estimated from time series with different causal labels as intuitively shown in Fig. 1, which we confirm experimentally in Section 5.2.

To prepare d_t for each t, given a time series with the length T, $S = \{(x_1, y_1), \cdots, (x_T, y_T)\}$, we use its subsequence with the length W $(W < T)$, i.e., $\{(x_{t-(W-1)}, y_{t-(W-1)}), \cdots, (x_t, y_t)\}$ $(t = W, \cdots, T)$[4]. As a result, we obtain the MMD pairs $\{d_W, \cdots, d_T\}$.

Although we can directly use these MMD pairs as a single feature vector, such a feature vector has the dimensionality $2(T - W + 1)$, which depends on the time series length T. As feature vectors whose dimensionalities are the same for time series with different lengths, we utilize the mean of the MMD pairs. However, when simply using the mean $(d_W + \cdots d_T) / (T - W + 1)$ as a feature vector, the feature vectors take the same value for the two sets of the MMD pairs whose empirical means are the same, but whose empirical distributions are different.

For this reason, to avoid mapping different distributions of the MMD pairs to the same feature vector, we again utilize kernel mean embedding. By using a different kernel function k_D from k_X and k_Y, we define our feature representation as

$$\nu(S) \equiv \frac{1}{T - W + 1} \sum_{t=W}^{T} \Phi_D(d_t) \tag{12}$$

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

which is the mean over the feature mappings $\Phi_D(d_t) \equiv k_D(d_t, \cdot)$[5].

[4]By using shorter time series, we can reduce the time complexity when computing weight vectors by KKF-CEO (i.e., $O(T^3)$ [Zhu *et al.*, 2014]).

[5]When using samples that are drawn directly from a distribution,

In (12), to compute the feature mapping $\Phi_D(\cdot)$, we employed random Fourier features (RFF) [Rahimi and Recht, 2007], which approximate a feature mapping as a low-dimensional vector of random features that are sampled from the Fourier transform of the kernel function. In experiments, we set the number of features $m = 100$ and obtained an m-dimensional feature vector for each time series, where we observed no significant improvements in the inference accuracy when using a larger m.

3.3 Extensions to Multivariate Time Series

Finally, we describe how our approach can be extended to n-variate time series ($n \geq 3$). We first present the feature representation for trivariate time series (i.e., $n = 3$) and we then address the case where $n > 3$.

Trivariate Time Series

Our feature representation for trivariate time series is founded on conditional Granger causality [Geweke, 1984], which can be applied to multivariate time series unlike Definition 1.

When using Definition 1 for trivariate time series, we can wrongly identify Granger causality. For instance, when there is no causal relationship between X and Y, if the third variable Z is their common cause, we can wrongly conclude that X is the cause of Y, or Y is the cause of X. This is because $P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$ or $P(X_{t+1}|S_X, S_Y) \neq P(X_{t+1}|S_X)$ might hold due to the influence of Z.

To address the influence of Z, conditional Granger causality compares the two conditional distributions given S_Z, i.e., the observations of the random variables $\{Z_1, \cdots, Z_t\}$, each of which is defined on $\mathcal{Z}$. Formally, it defines X as the cause of Y given Z if $P(Y_{t+1}|S_X, S_Y, S_Z) \neq P(Y_{t+1}|S_Y, S_Z)$ holds; otherwise, X is *not* the cause of Y given Z.

We introduce causal labels based on conditional Granger causality. For instance, similarly to (2), we regard the causal label $X \to Y$ as

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

which can be rewritten as

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

where $\mu_{X_{t+1}|S_X,S_Y,S_Z}$, $\mu_{X_{t+1}|S_X,S_Z}$, $\mu_{Y_{t+1}|S_X,S_Y,S_Z}$, and $\mu_{Y_{t+1}|S_Y,S_Z}$ are the kernel mean embeddings of $P(X_{t+1}|S_X, S_Y, S_Z)$, $P(X_{t+1}|S_X, S_Z)$, $P(Y_{t+1}|S_X, S_Y, S_Z)$, and $P(Y_{t+1}|S_Y, S_Z)$, respectively.

Motivated by these expressions, we address the case where Z might influence X and Y by adding $\widehat{\text{MMD}}^2_{X_{t+1}|Z}$ and $\widehat{\text{MMD}}^2_{Y_{t+1}|Z}$ to the feature representation, where they are the MMD between $\mu_{X_{t+1}|S_X,S_Y,S_Z}$ and $\mu_{X_{t+1}|S_X,S_Z}$ and the MMD between $\mu_{Y_{t+1}|S_X,S_Y,S_Z}$ and $\mu_{Y_{t+1}|S_Y,S_Z}$, respectively. For this reason, we extend the feature representation (12) simply by modifying d_t as follows

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

the kernel mean embedding of the distribution is estimated by using the same weight values (e.g., $1/(T - W + 1)$ in (12)) [Muandet *et al.*, 2017].

n-variate Time Series ($n > 3$)

Although it is possible to develop the feature representation for n-variate time series ($n > 3$) by adding the extra MMDs to d_t, it becomes very difficult to prepare enough training data to train the classifier since the number of possible combinations of the common cause variables of the variable pair X and Y grows super-exponentially in n.

For this reason, we used the feature representation for trivariate time series. From n-variate time series, we infer a causal relationship between each variable pair X and Y in three steps. First, for each $v \in \{1, \cdots, n-2\}$, we obtain the feature vector from the observations of the triplet of the variables (X, Y, Z_v). Second, based on each feature vector, we use a trained classifier to compute the probabilities of the three labels ($X \to Y$, $X \leftarrow Y$, and *No Causation*). Finally, we assign the label with the highest average probability.

Addressing the case where there are more than one common cause variables is left as our future work.

4 Related Work

It is worth noting that one of the supervised learning methods for i.i.d. data, the randomized causation coefficient (RCC) [Lopez-Paz *et al.*, 2015], also uses kernel mean embedding to obtain the features of the distribution that differ depending on the causal relationships. However, RCC and our method are designed to obtain different features of the distribution. Specifically, via kernel mean embedding, RCC obtains information about the marginal and conditional distributions, which are known to differ depending on the causal directions according to an assumption called the *independence of cause and mechanism* (ICM) [Janzing and Scholkopf, 2010]. In contrast, by using kernel mean embedding, our method measures the distance between the conditional distributions given past variable values since it becomes sufficiently different depending on Granger causality.

5 Experiments

5.1 Experimental Settings

We compared the performance of our method (hereafter referred to as the supervised inference of Granger causality (**SIGC**)) with the supervised learning method for i.i.d. data **RCC** [Lopez-Paz *et al.*, 2015][6], with the Granger causality methods **GC**$_{VAR}$ [Granger, 1969][7], **GC**$_{GAM}$ [Bell *et al.*, 1996][7], and **GC**$_{KER}$ [Marinazzo *et al.*, 2008][8], which identify Granger causality by using the VAR model, the GAM, and kernel regression, respectively, and with transfer entropy **TE** [Schreiber, 2000][9], which infers causal relationships not by using regression models, but by performing density estimation.

For our **SIGC**, we used a random forest classifier[10] to make a fair comparison with **RCC**, which has achieved better per-

[6] https://github.com/lopezpaz/causation_learning_theory

[7] http://people.tuebingen.mpg.de/jpeters/onlineCodeTimino.zip

[8] https://github.com/danielemarinazzo/KernelGrangerCausality

[9] https://github.com/Healthcast/TransEnt

[10] The number of trees is selected from {100, 200, 500, 1000, 2000} via 5-fold cross validation.

formance with the random forest classifier than with the SVM [Lopez-Paz *et al.*, 2015]. To prepare feature vectors, we used the Gaussian kernel as k_X, k_Y, and k_D and set the kernel parameter using the median heuristic, which is a well-known heuristic for selecting it [Scholkopf and Smola, 2001]. We set the parameter W in our method and the parameters in the existing methods to provide the best performance for each method in our synthetic data experiments. For our method, we selected $W = 12$.

5.2 Experiments on Bivariate Time Series Data

Classifier Training

We trained a classifier to infer causal relationships from bivariate time series data.

As with the existing supervised learning methods [Bontempi and Flauder, 2015; Lopez-Paz *et al.*, 2015; 2017], we used synthetic training data in both synthetic and real-world data experiments since there are few real-world data where the causal relationships are known.

We generated $15,000$ pairs of synthetic time series with the length $T = 42$ so that there were 5.000 instances each with causal labels $X \to Y$, $X \leftarrow Y$, and *No Causation*. Here, we used the following *linear* and *nonlinear* time series:

- Linear time series were sampled from the VAR model:

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

  where $\tau = 1, \cdots, P$ ($P \in \{1, 2, 3\}$) and E_{X_t}, E_{Y_t} were sampled from the Gaussian distribution $\mathcal{N}(0, 1)$. To obtain time series with $X \to Y$, we used the following coefficient matrix

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

  where a_τ, d_τ were drawn from the uniform distribution $\mathcal{U}(-1, 1)$, and $c_\tau \in \{-1, 1\}$. Similarly, we prepared time series with $X \leftarrow Y$, and *No Causation*.

- Nonlinear time series were also similarly generated by using the VAR model with a standard sigmoid function $g(x) = 1/(1+\exp(-x))$. For instance, we prepared time series with $X \to Y$ so that Y_t depended on $\{[g(X_{t-\tau}), Y_{t-\tau}]^\top\}_{\tau=1}^{P}$ while X_t depended only on $\{X_{t-\tau}\}_{\tau=1}^{P}$.

- Finally, we scaled each time series with mean 0 and variance 1.

Synthetic Time Series

Next, we tested the performance of our method for inferring causal relationships ($X \to Y$, $X \leftarrow Y$, and *No Causation*) from synthetic time series data. We used the following linear and nonlinear test data:

- *Linear Test Data*: We prepared 300 pairs of linear time series so that the numbers of time series with $X \to Y$, $X \to Y$, and *No Causation* were 100. In a similar way to the linear time series in the training data, each time series was sampled from the VAR model (13) although several parameter settings were different (e.g., the noise variance was given as $p \in \{0.5, 1.0, 1.5, 2.0\}$).

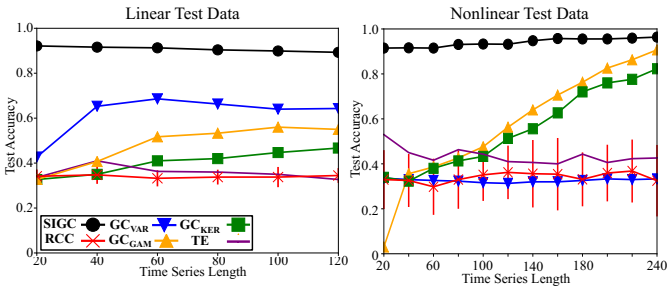

Figure 2: Test accuracies for 300 pairs of time series data against time series length (left: linear test data; right: nonlinear test data). Means and standard deviations (error bars) are shown for our method and RCC based on 20 runs with different training data.

- *Nonlinear Test Data*: We used 300 pairs of nonlinear time series, where there were 100 time series with $X \to Y$, $X \to Y$, and *No Causation* in each dataset. We generated nonlinear time series with $X \to Y$ by

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

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

  where the noise variables E_{X_t}, E_{Y_t} were sampled from $\mathcal{N}(0, 1)$. Similarly, we prepared nonlinear time series with $X \leftarrow Y$. To prepare time series with *No Causation*, we simply ignored the exponential term in (15).

Using linear and nonlinear test data, we compared the performance of our method with that of the existing methods. Fig. 2 shows the test accuracies. Note that for **SIGC** and **RCC**, we show the means and the standard deviations (error bars) in 20 experiments with different training data since these methods use randomly generated training data.

The performance of the Granger causality methods ($\mathbf{GC}_{VAR}$, $\mathbf{GC}_{GAM}$, and $\mathbf{GC}_{KER}$) depended on whether or not the regression model could be well fitted to the data. For instance, since the VAR model could be well fitted to linear test data, $\mathbf{GC}_{VAR}$ performed well on linear test data although it worked badly on nonlinear test data. In addition, with nonlinear test data, $\mathbf{GC}_{KER}$ was less accurate than $\mathbf{GC}_{GAM}$ because the time series was too short for us to perform kernel regression. Similarly, **TE** worked poorly since the time series was too short for us to perform density estimation.

In contrast, our method worked sufficiently well on linear and nonlinear test data. The main reason for the good performance lies in our feature representation. This can be seen from a comparison with the supervised learning method **RCC** since it prepares training data in the same way as our method.

To verify our feature representation, we confirmed experimentally that feature vectors are sufficiently different depending on causal labels. To do so, we used nonlinear test data to plot a histogram of the MMD pairs $\{d_t\}$ that were used to compute the feature vector for each time series. Fig. 3 shows the results. Since each MMD in d_t is a finite sample estimate, no MMD becomes exactly zero. However, we can see that the MMDs became sufficiently different according to the causal labels. We obtained similar results with linear test data although we have omitted them due to space limitations.

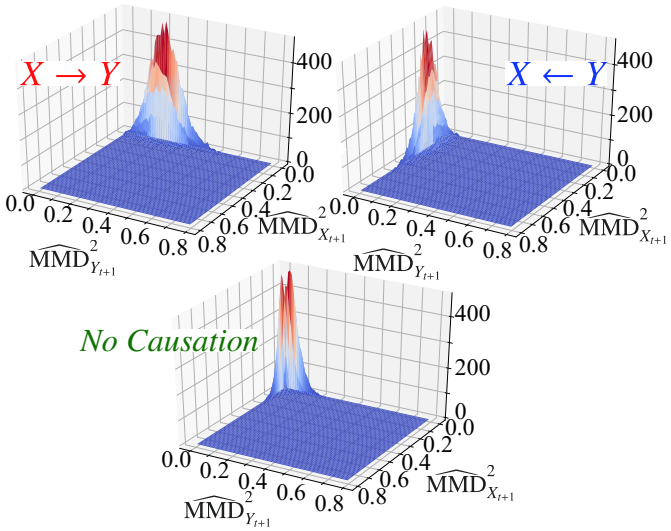

Figure 3: Histogram of MMDs used to compute the feature vector for each time series in nonlinear test data with $X \to Y$ (top left), $X \leftarrow Y$ (top right), and *No Causation* (bottom).

	SIGC	GC_{VAR}	GC_{GAM}	GC_{KER}	TE
Temperature ($T = 16382$)	✓	✗	✓	✓	✗
Radiation ($T = 8401$)	✓	✓	✓	✓	✓
Internet ($T = 498$)	✓	✓	✗	✗	✓
Sun Spots ($T = 1632$)	✓	✗	✗	✗	✓
River Runoff ($T = 432$)	✓	✓	✓	✗	✓

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

Table 1: Causal relationships inferred from the first test dataset (top; ✓and ✗ denote correct and incorrect results, respectively) and test accuracies for the second test dataset (bottom; Means and standard deviations are shown for **SIGC** and **RCC** based on 20 runs).

In fact, since the MMD pairs are sufficiently different, we can assign the causal label by taking another approach, i.e., an *unsupervised* approach that uses no training data, which outputs the causal label in two steps. First, using the MMD pairs, the two statistical tests are performed to determine if the mean of $\widehat{\mathrm{MMD}}^2_{X_{t+1}}$ is zero and if the mean of $\widehat{\mathrm{MMD}}^2_{Y_{t+1}}$ is zero. Then, by using the two p-values and some threshold value (significance level), we can assign a causal label ($X \to Y$, $X \leftarrow Y$, or *No Causation*) to each time series.

However, we confirmed experimentally that the performance of such an unsupervised approach depended greatly on the threshold value. Furthermore, it was less accurate than our method (e.g., on nonlinear test data with the length $T = 250$, its test accuracy was 0.810 (not shown in Fig. 2) while our method achieved 0.966) although we selected the threshold value that provided the best performance. These results suggest the effectiveness of our supervised learning approach, which can obtain the decision boundary needed to determine the causal label by training a classifier.

Real-world Time Series

We tested our method by using real-world time series. To improve the reliability of the experiment, we used the following two test datasets:

- The first test dataset was composed of five pairs of bivariate time series downloaded from the Cause-Effect Pairs database [Jakob,], whose true causal relationships are reported in [Jakob,] as $X \to Y$ for three pairs and as $X \leftarrow Y$ for the others. For instance, the *River Runoff* is a bivariate time series concerning average precipitation X and average river runoff Y, and the true causal relationship is regarded as $X \to Y$.
- Using the above five real-world time series, we prepared a second test dataset that consisted of subsequences in each time series. To prepare the subsequences, we simply chopped each time series into multiple subsequences of length $T = 200$.

As regards training data, we used synthetic time series that we prepared in the same way as those for synthetic data experiments.

Table 1 shows the result for each test dataset. Note that we have omitted **RCC** from the top table in Table 1 because it showed different outputs in 20 experiments where different training data were used as in the synthetic data experiments, while our **SIGC** always output the same causal directions. As shown in Table 1, our **SIGC** outperformed the other existing methods regardless of the time series length T.

5.3 Experiments on Multivariate Time Series Data

We tested $\mathbf{SIGC}_{tri}$, which utilizes a feature representation for trivariate time series. For classifier training, we used synthetic trivariate time series that were generated in a similar way to those used in the experiments described in Section 5.2. As test data, we used the following time series gene expression data:

- We used the *Saccharomyces cerevisiae* (yeast) cell cycle gene expression dataset collected by [Spellman *et al.*, 1998]. By combining four short time series that were measured in different microarray experiments, we prepared a time series with the length $T = 57$, where the number of genes was $n = 14$. To determine the true causal relationships between the genes, we used the gene network database KEGG [KEGG, 1995].

Since the number of non-causally-related gene pairs was much larger than the number of causally-related gene pairs, we evaluated the performance of each method in terms of the macro and micro-averaged F1 scores rather than test accuracy.

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

Table 2: Macro and micro-averaged F1 scores. Means and standard deviations are shown for our methods and **RCC** based on 10 runs.

Table 2 shows the results. Since the data were measured in different microarray experiments, all the methods could not sufficiently work well. However, our $\mathbf{SIGC}_{tri}$ worked better than the existing Granger causality methods. It also performed better than $\mathbf{SIGC}_{bi}$, which uses the feature representation for bivariate time series, thus indicating that it is important to consider the influence of the common cause variable as described in Section 3.3.

6 Conclusions

We have proposed a classification approach to Granger causality identification. Whereas the performance of the model-based methods depended hugely on whether the regression model could be well fitted to the data, our method performed sufficiently well by using the same feature representation and the same classifier (random forest classifier). Furthermore, we demonstrated experimentally the reason for such good performance by showing a sufficient difference between the feature vectors that depends on Granger causality. These results demonstrate the effectiveness of classification approaches to Granger causality identification.

Addressing complicated real-world scenarios (e.g., inferring the causal directions that change over time) constitutes our future work.

References

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

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

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

[Cheng *et al.*, 2014] Dehua Cheng, Mohammad Taha Bahadori, and Yan Liu. FBLG: a simple and effective approach for temporal dependence discovery from time series data. In *KDD*, pages 382–391, 2014.

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

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

[Gretton *et al.*, 2007] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520, 2007.

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

[Jakob,] Zscheischler Jakob. Database with cause-effect pairs. https://webdav.tuebingen.mpg.de/cause-effect/.

[Janzing and Scholkopf, 2010] Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[Kar *et al.*, 2011] Muhsin Kar, Şaban Nazlıoğlu, and Hüseyin Ağır. Financial development and economic growth nexus in the MENA countries: Bootstrap panel granger causality analysis. *Economic modelling*, 28(1):685–693, 2011.

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

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

[Lopez-Paz *et al.*, 2015] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In *ICML*, pages 1452–1461, 2015.

[Lopez-Paz *et al.*, 2017] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering Causal Signals in Images. In *CVPR*, 2017.

[Marinazzo *et al.*, 2008] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel-Granger causality and the analysis of dynamical networks. *Physical Review E*, 77(5):056215, 2008.

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

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

[Scholkopf and Smola, 2001] Bernhard Scholkopf and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[Schreiber, 2000] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

[Spellman *et al.*, 1998] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce 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.

[Sriperumbudur *et al.*, 2010] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

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

[Yao *et al.*, 2015] Shun Yao, Shinjae Yoo, and Dantong Yu. Prior knowledge driven Granger causality analysis on gene regulatory network discovery. *BMC bioinformatics*, 16(1):273, 2015.

[Zhu *et al.*, 2014] Pingping Zhu, Badong Chen, and Jose 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.