Deep Learning Approach to Granger Causality

by

Yuting Li

Student Number: 20139547

Sept 2022

Supervised by Prof. Tomaso Aste

A Dissertation in part fulfilment of the

Degree of MSc Scientific and Data Intensive Computing

Department of Physics and Astronomy

University College London

CONTENTS

LIST OF FIGURES

List of Tables

ACKNOWLEDGEMENTS

First, I would like to express my deepest appreciation to Prof.Tomaso for his invaluable patience and feedback despite lots of difficulties and challenges during the pandemic. Secondly, I also could not have undertaken this journey without constant support from my parents, especially for their financing of my master study. Additionally, my sincere thanks also go to my classmates and my colleagues for their moral support. Lastly, I would like to recognize the assistance and help that I received from the Physics department and UCL.

DECLARATION

I, Yuting Li, confirm that the work presented in this paper is my own. Where information has been derived from other sources, I confirm that this has been indicated in the dissertation.

The code for the Python package "DLcausality", which is developed and used in this paper, can be found at https://github.com/oliviayt1224/Deep-Learning-Causality. For more useful instructions for it, please refer to Section 4 Python Implementation.

ABSTRACT

The questions of how to define causality has kept philosophers debating for more than two thousand years, but still no consensus has been achieved. Due to the difficulties in qualitatively interpreting causality, scientists were soon converted to describing causality quantitatively. Granger causality is one of the most representative quantitative measures, whose effectiveness has been validated by researchers on various fields. However, the traditional Granger causality is only applicable to linear dependence, which drives scientists to develop other alternatives for detecting nonlinear causal relationships. To break the limitation of linearity, a non-parametric measure called transfer entropy (TE) is adopted in this paper to investigate both linear and nonlinear Granger causality. For studying linear Granger causality, vector autoregressive models are used, whilst neural networks are introduced for studying the nonlinear causality. One of the meaningful contributions of this paper is exploring causality for not only two-variable cases but also three-variables cases. The significance of the results are assessed using the z-scores at 99% confidence interval. There are four synthetic data distributions discussed in this paper and all models involved are programmed into an open-source Python package called "DLcausality". After analyzing the results for each example, the proposed model in this paper has shown incredible effectiveness in detecting both linear and nonlinear Granger causality.

Keywords: Granger Causality, Transfer Entropy, Neural Network

CHAPTER

1. INTRODUCTION

There is a famous collection of books called "I Wonder Why", which should be familiar to most readers who may have read it during their childhoods. The reason it is so popular with children is because of everyone's instinctive curiosity about the world around them. Every time we ask a question "why?", it reflects our curiosity about the causal correlations between two terms. However, what does the term causality mean? Some people think it is equated with correlation, which is a common misconception. As an example, the sales of ice cream and swimsuit will both rise in summer, revealing that they are positively correlated. However, apparently there is not a causality between them. The confusion between correlation and causality results from a lack of a clear and widely-accepted definition of causality. Even though scholars have been debating this topic in both philosophy and statistics throughout history, no agreement has been achieved in philosophy. Fortunately in statistics, a widely-recognized theory called "Granger Causality" was proposed and made it possible to study the causality.

Granger causality has been widely used in various fields like economics, environmental science and neuroscience, all recognising its effectiveness. However, it still has many limitations due to its restrictive assumptions. One of the greatest limitations arises from its assumption of the linear dependence between variables, which is inconsistent with real-world situations. Scientists have realized that applying the linear Granger causality model to explore the causal relations for nonlinearly dependent variables would lead to incorrect conclusions.[18] To address this limitation, an alternative non-parametric measure called "Transfer Entropy" was introduced to investigate nonlinear Granger causality, which was found to be more applicable to complex real-world cases with higher accuracy and precision.[31] However, no matter which measure is used, a regression model is always required due to the need to test the predictive capability of historical data. Instead of utilizing the conventional linear regression method such as the Vector Autoregression (VAR) model that the Granger causality test usually uses, neural networks are a more appropriate choice to capture and model nonlinear dependence. Up to now, limited research has been done in studying the field of how neural networks can be used in detecting the nonlinear Granger causality by estimating the values of transfer entropy, which brings up the main innovation of this paper.

The majority of the current research is confined to simply the causality for two-variable cases. This paper not only measures the transfer entropy for traditional two-variable cases, but more importantly, also examined the effect on transfer entropy of incorporating a third

variable into regression, which is one of the factors that makes this paper differ from the existing literature. To be more specific, there are two statistics considered in ternary cases: the conditional transfer entropy and the high-order transfer entropy. For example, when studying the Granger causality from X and Y, the conditional transfer entropy studies the information transfer from X and Y given the past information of a third variable Z. A significant conditional transfer entropy indicates that incorporating Z improves the regression for Y. As for the high-order transfer entropy, it is defined to be the difference between the conditional and unconditional transfer entropies. If this measure is significant, it proves that adding the past values of Z into the models can greatly decrease the regression residuals. Additionally, the proposed models in this paper investigates the causality in both directions (bidirectional causality) for both coupled and ternary cases. In order to validate these ideas, an open-resource Python package is developed, which contains functions for generating four synthetic datasets, building up regression models and calculating statistics for Granger causality. It provides users various choices to self-define their own models and also great potential for further development.

Based on this background, this paper aims to detect and investigate the nonlinear Granger causality from the perspective of transfer entropy, using the techniques of neural networks. This paper has been organised into six sections: Section 1 introduces the background literature and the methodology involved in this paper; Section 2 describes the data generation processes for four distributions; Section 3 includes the usage of the Python package called "DLcausality", which is developed for investigating both linear and nonlinear Granger causality more conveniently; Section 4 summarizes all the results and analyzes the effectiveness of the model; Section 5 provides some of the potential aspects for future research and finally Section 6 is a brief conclusion of all works involved in this paper.

2. BACKGROUND AND METHODOLOGY

Although causality is a term mentioned frequently in our daily life, no one is able to bring up a precise and consistent definition of it. There is no doubt that the term causality narrows the scopes of cointegration and interaction between two events. However, the concept of causality is still very abstract, which is difficult to identify and capture its regular patterns and features. Due to its complexity, causality has been one of the hottest topic in both philosophy and statistics throughout history.

People's understanding regarding causality has experienced a long evolution. Among all the existing diverse opinions, the theory brought up by David Hume is the most representative and influential one. According to Hume's opinions, if for every event that belongs to type A, it is followed by an event of type B, then this observed regular pattern can be regarded as a causal effect, drawing a conclusion that A causes B.[12] Although Hume has made remarkable contributions to the definition of causality, this topic has kept philosophers discussing for more than two thousand years and has never been resolved. It is indisputable that these philosophical thoughts are wise and reasonable, providing us with a conceptual framework to understand causality. However, they are too theoretical to be directly applied to solve practical problems. Therefore, a lot of scientists turn their attention to measuring causality in an objective and quantitative way.

2.1 Granger Causality

The problem of defining "causality" mathematically is non-trivial and challenging. The reason beyond is that it is very difficult to design a measure which can capture the commonalities of causality for the majority of real-world cases as the causal links in different situations are variable. One of the most representative causality theory in a time-series context is called Granger causality, which has been widely recognized by academia and also contributed to Granger's winning of the Nobel Prize in 2003.[2] Granger proposed the idea of studying causal relations by decomposing the cross spectrum between two variables, which was a great progress for operationally investigating causality by means of mathematical procedures.[2]

2.1.1 Formalization of Granger Causality

Granger causality is firstly targeted at two-variable cases. Granger designed a linear vector autoregressive model for testing bivariate causality, which is usually referred to as Granger causality test.[6] The regression equations involved in the test are shown below:

$$Y_t = a + bY_{t-L} + u_{Y,t} \tag{1}$$

$$Y_t = c + dY_{t-L} + eX_{t-L} + u'_{Y,t} \tag{2}$$

where X_t and Y_t are two time series, L is the time-lag, and the regression errors $u_{Y,t}$ and $u'_{Y,t}$ are assumed to be mutually independent, and individually independent and identically distributed (i.i.d.) with zero mean and constant variance. Granger causality can be determined by

comparing the variance of two residual terms.

$$var(u'_{Y,t}) < var(u_{Y,t}) \tag{3}$$

If this inequality holds, it is reasonable to conclude that the past information of X is valuable for forecasting the future values of Y, which indicates that X is a Granger cause for Y.

Geweke put forward a standard measure called "F-statistic" ($F_{X\to Y}$), which can be used as the evaluation criterion of Granger causality.[8]

$$F_{X\to Y} = \ln\left(\frac{var(u_{Y,t})}{var(u'_{Y,t)}}\right) \tag{4}$$

According to the definition, apparently F-statistic is always non-negative, which is only equal to zero when X is not a Granger cause for Y. In other words, if it is positive, it supports for the existence of Granger causal effects.

2.1.2 Applications and Limitation of Granger Causality

Granger causality has been widely adopted in various research fields especially in economics. In the first beginning, Granger used his theory to study the relation between aggregate advertising and aggregate consumption spending.[5] The results from his work proved that there was a unidirectional causal relationship from consumption to advertising, which was consistent with an infinite number of alternative models run by other researchers. His findings proved the effectiveness of Granger causality and have been widely adopted to study economics. For example, it was applied to detect the relationship between innovation and economic growth in the European Economic Area (EEA) countries. The results verified the existence of a bidirectional causality between these two variables.[32]

Despite its popular recognition in academia, it has many limitations, among which the restriction of linearity will be the most glaring one. Traditionally, Granger causality relies on the assumption of linear vector autoregressive (VAR) models, making it only applicable for linear dependence. Unfortunately, most of the correlations are nonlinear in reality.[35] Owning to this fact, applying Granger causality sometimes will produce unreasonable conclusions. For example, there is a paper, in which the results suggested that the US gross national product was a Granger cause for sunspots.[7]

2.2 Transfer Entropy

To break the limitation of linearity for Granger causality, Schreiber introduced an information-theoretic notion called transfer entropy (TE), which is a measure for nonlinear causality.[11] TE quantifies the amount of directed information transfer between processes, therefore it is asymmetrical. Different from Granger causality, the value of TE is directly estimated from the dataset without assuming any model regarding the relation. Thus, it is a non-parametric measure, which does not suffer from model constraints.[29]

2.2.1 Formulation of Transfer Entropy

Transfer entropy is based on the concept of Shannon entropy, whose general form for (X, Y) is:

$$H(X,Y) = -\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i, y_j)\ln(p(x_i, y_j)) \tag{5}$$

$$h(X,Y) = -\int_{R_x}\int_{R_y} p(x,y)\ln(p(x,y))dxdy \tag{6}$$

where equation (5) represents the joint entropy for discrete random variables and equation (6) is the one for continuous random variables. Similarly, the conditional entropy of X given Y can be calculated by:

$$H(X|Y) = -\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i, y_j)\ln(p(x_i|y_j)) \tag{7}$$

$$h(X|Y) = -\int_{R_x}\int_{R_y} p(x,y)\ln(p(x|y))dxdy \tag{8}$$

The transfer entropy from X_{t-L} to Y_t for lag L can be represented by:

$$TE^{(L)}_{X\to Y} = H(Y_t|Y_{t-L}) - H(Y_t|X_{t-L}, Y_{t-L}) \tag{9}$$

This paper will not only study the causality for two-element cases but also step forward to three-element cases. For the latter, suppose there is another variable Z, we set up a term called "conditional transfer entropy" ($TE^{(L)}_{X\to Y|Z}$), which is the transfer entropy from X_{t-L} to Y_t given the information of Z_{t-L}. Its formulation is shown below:

$$TE^{(L)}_{X\to Y|Z} = H(Y_t|X_{t-L}, Y_{t-L}) - H(Y_t|X_{t-L}, Y_{t-L}, Z_{t-L}) \tag{10}$$

2.2.2 Calculation of Linear Transfer Entropy

Transfer entropy has been proven to be equivalent to Granger causality for Gaussian variables, exponential Weinman and log-normal distribution.[16, 21] More importantly, scholars have sorted out a more practical method to associate transfer entropy and Granger causality. Recalling equation (4), equation (11) is the linear method used to calculate transfer entropy in this paper.

$$TE^{(L)}_{X\to Y} = \frac{1}{2}\ln\left(\frac{var(u_{Y,t})}{var(u'_{Y,t})}\right) = \frac{1}{2}F_{X\to Y} \tag{11}$$

where the left-hand side represents the transfer entropy, and the right-hand side refers to the F-statistic, which is introduced previously as a tool of measuring Granger causality.

As for the conditional transfer entropy, equation (13) is the solution.

$$Y_t = f + gY_{t-L} + hX_{t-L} + iZ_{t-L} + u''_{Y,t} \tag{12}$$

$$TE^{(L)}_{X\to Y|Z} = \frac{1}{2}\ln\left(\frac{var(u'_{Y,t})}{var(u''_{Y,t)}}\right) \tag{13}$$

We can expect that if including the information of Z_{t-L} improves the regression of Y_t, then $TE^{(L)}_{X\to Y|Z}$ will be larger than 0.

Besides, this paper also contributes to studying the Granger causality bidirectionally, which means the proposed model will also try to verify the existence of Granger causality from Y to X. Similar to the previous procedures, it is pretty straightforward to derive the equations for calculating $TE^{(L)}_{Y\to X}$ and $TE^{(L)}_{Y\to X|Z}$.

$$X_t = a + bX_{t-L} + u_{X,t}$$
$$X_t = c + dY_{t-L} + eX_{t-L} + u'_{X,t}$$
$$X_t = f + gY_{t-L} + hX_{t-L} + iZ_{t-L} + u''_{X,t}$$

$$TE^{(L)}_{Y\to X} = \frac{1}{2}\ln\left(\frac{var(u_{X,t})}{var(u'_{X,t)}}\right)$$

$$TE^{(L)}_{Y\to X|Z} = \frac{1}{2}\ln\left(\frac{var(u'_{X,t})}{var(u''_{X,t)}}\right)$$

2.3 Neural Network

Nowadays, as technology rapidly develops, the computational abilities of computers are much more powerful than ever, and even far exceeding the abilities of humans. Although a computer can solve any problem and execute any command that can be described in a logical formula, it is far less intelligent than a human brain. For example, it can not make inferences and judgments from past experience. In other words, it has no self-learning ability. To enable machines to think intelligently as humans do, scientists simulated the realistic model of neural activities in human brains and created an "artificial neural network" to be the brain for a computer.

2.3.1 Multilayer Perceptron

Before introducing the concept of Multilayer Perceptron, it is necessary to first start with the concept of Feedforward Neural Network (FNN). FNN is the first and simplest type of neural network, in which the information will only flow forward.[26] The structure of A simple three-layer FNN is displayed in Figure 1. At first, FNN only had a single-layer, which was called Rosenblatt's perceptron. It has an inability on solving nonlinear separable problems.[1, 3] To address this issue, one of the solutions is to build up hidden layers in FNN. Cybenko has proved that a FNN with one hidden layer can simulate the pattern for any function based on the universal approximation theorem.[10] This is one of the reasons why Multilayer Perceptron (MLP) has been developed, which is a subset of FNN whose layers are fully connected. A MLP consists of at least three layers: an input layer, a hidden layer and an output layer. Figure 1 above is actually a graph for MLP. In fact, most commonly

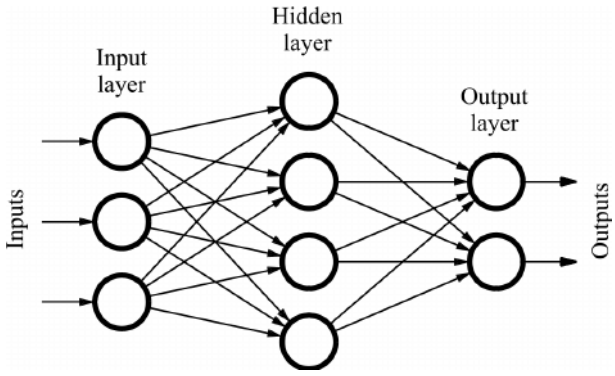

Figure 1: 3-layer Feedforward Neural Network[20]

there are more than one hidden layers in a MLP.

Figure 2 presents the functions of a single neuron in the hidden layer, from which it can be seen that the whole process consists of two parts: (1) calculate a weighted sum of inputs $(x_1, x_2...x_n)$; (2) input the sum into an activation function and forward the result to the next layer.

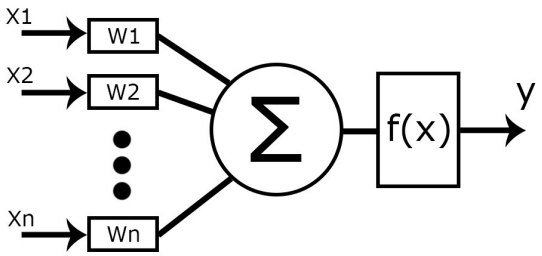

Figure 2: Functions of Neuron (from Good Audience)

2.3.2 Backpropagation with Stochastic Gradient Descent

The key factor that differentiates MLP from Rosenblatt's perceptron in the effectiveness of solving nonlinear separable problems is the Backpropagation training algorithm, which was first introduced in the 1970s as a general method for performing automatic differentiation of complex nested functions.[4] Backpropagation had never been widely applied in training neural networks until Rumelhart and his partners described it as a procedure of repeatedly adjusting the weights of the connections in a MLP.[9]

The idea of backpropagation is:

(1) Assign part of the blame for the training error in the final layer to each neuron in the previous layer, and then further split up blame if there exist other hidden layers, and so on;
(2) Find out the correlation between the change of error and weights;
(3) Use an optimization technique like Stochastic Gradient Descent to compute the optimal weights for minimizing the error.

An mathematical introduction of this algorithm is provided in the following section. Before

going through the equations, it is necessary to first define the notations:

w_{mn}^l: the weight regarding the information flow from the nth neuron in the $(l-1)$th layer to the mth neuron in the lth layer.
b_m^l: the bias of the mth neuron in the lth layer.
a_m^l: the output from the activation function in the mth neuron of the lth layer.
$\sigma(net)$: the activation function used in the neural network.
z_m^l: the weighted value being input to the activation function in the mth neuron of the lth layer.
δ_m^l: the error for the mth neuron in the lth layer.

The outputs of activation functions from two adjacent layers satisfy the following relation:

$$a_m^l = \sigma(\sum_{i=1}^{k} w_{mi}^l a_i^{l-1} + b_m^l) = \sigma(z_m^l) \tag{14}$$

where k represents the total number of neurons in the $(l-1)$th layer.

The partial derivatives $\frac{\partial L(x_j)}{\partial w}$ and $\frac{\partial L(x_j)}{\partial b}$ of the loss function $L(x_j)$ (or the cost function C) with respect to weight w or bias b are the key elements of backpropagation that needed to be calculated. For regression problems, the quadratic loss function $L(x_j)$ for a specific training sample x_j is given below:

$$L(x_j) = \frac{1}{2}||y(x_j) - \hat{y}(x_j)||^2 = \frac{1}{2}||y(x_j) - a^L(x_j)||^2 = \frac{1}{2}\sum_i (y_i(x_j) - a_i^L(x_j))^2 \tag{15}$$

where i is the number of neurons in the final layer L, $y(x_j)$ is the expected output for input x_j, L is the number of layers in the neural network and $a^L(x_j)$ is the output from the activation function in the last layer taking x_j as input.

It is universally acknowledged that the purpose of using any statistical model including neural networks, is to obtain the most possibly accurate predictions for dependant variables. The criteria of evaluating the quality of the model is how much error there is between y and $\hat{y}$, which is measured by the loss function $L(x)$. Obviously, the goal of training a neural network is to find out the optimal weights and biases, which lead to a minimum loss. Then the question comes out: how to find the optimal model? To answer this question, in this paper the optimization algorithm "Stochastic Gradient Descent" (SGD) is covered.
The gradient of a function f, denoted as $\bigtriangledown f$, is a vector containing all its partial derivatives, which is shown in equation (16).

$$\bigtriangledown f(x,y) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{16}$$

The gradient is a vector that points to the direction in which the function increases fastest, and it will become zero at a local maximum or a local minimum.[15] Based on these characteristics, an iterative optimisation algorithm called "Gradient Descent" (GD) is purposed by the famous mathematician Cauchy, to find the local minimum of a

differentiable function.[23] To perform the GD algorithm, there are usually two iterative procedures:

(1) Compute the gradient at current point (x_n, y_n).
(2) Find the opposite direction of this gradient and move the current point towards it by a specific step.

Equation (17) and (18) present the iterative processes from a mathematical perspective:

$$x_{n+1} = x_n - \eta_n \frac{\partial f}{\partial x}(x_n, y_n) \tag{17}$$

$$y_{n+1} = y_n - \eta_n \frac{\partial f}{\partial y}(x_n, y_n) \tag{18}$$

where η_n is a hyperparameter representing the learning rate (also called the step size) for the nth iteration, which might take different values for each iteration. The value of η_n affects whether the iteration is convergent or not and also the convergent speed for the iteration.

Stochastic Gradient Descent (SGD) is one of the commonly used optimization algorithm based on GD, which is stochastic in nature because it only chooses one random instance of the training data and then computes its gradient for each iteration. By contrast, there is another optimization algorithm called Batch Gradient Descent (BGD), which considers all the training samples into each iteration by calculating the average of all their gradients and then use it to adjust variables. Based on these facts, SGD usually converges faster than BGD does when dealing with a large dataset because it takes steps more frequently.

The following section provides step-by-step guide for using Backpropagation to train the neural network and SGD to obtain an optimal model (the one with minimum loss):

(1) Prepare a set of training examples.

(2) Initialize parameters like w^l, b^l, σ, η and determine the number of iterations.

(3) Shuffle the training set and randomly choose an instance.

(4) Forwardly compute z^l and a^l for each layer l.

(5) Measure the error δ^L for output layer L.

$$\delta^L = \bigtriangledown_a L \odot \sigma'(z^L) = (a^L - y) \odot \sigma'(z^L) \tag{19}$$

where $\bigtriangledown_a L$ is a vector including all the partial derivatives $\frac{\partial L}{\partial a_m^L}$, $\sigma'(z^L)$ contains all the partial derivatives $\frac{\partial a_m^L}{\partial z_m^L}$, and the operator $\odot$ denotes the Hadamard product (or element-wise product).

(6) For the rest of the layers l $(l < L)$, calculate δ^l.

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l) \tag{20}$$

where the weight matrix w^{l+1} is a vector containing the weights connecting to neurons in layer $l+1$. According to equation (20), the calculation of the error δ^l is dependent on the error in the next layer, showing that the computation of the error proceeds backward from the output layer down to the input layer. This is why the algorithm is called "Backpropagation".

(7) Update parameters using the Gradient Descent method:

$$w^l_{mn} = w^l_{mn} - \eta a^{l-1}_n \delta^l_m \tag{21}$$

$$b^l_m = b^l_m - \eta \delta^l_m \tag{22}$$

(8) Repeat step 3 to step 7 until the model reach a local minimum.

2.3.3 Calculation of Nonlinear Transfer Entropy

After training and optimizing a MLP with the training set, the neural network can be used to forecast Y for the testing set. Suppose the neural network is a generalized function f, then the predicted values of Y can be calculated by applying f to the data points of specific independent variables. The nonlinear methods of calculating transfer entropy are listed below:

$$u_{Y,t} = Y_t - f(Y_{t-L})$$

$$u'_{Y,t} = Y_t - f(X_{t-L}, Y_{t-L})$$

$$u''_{Y,t} = Y_t - f(X_{t-L}, Y_{t-L}, Z_{t-L})$$

$$TE^{(L)}_{X\to Y} = \frac{1}{2}\ln\left(\frac{var(u_{Y,t})}{var(u'_{Y,t)}}\right) \tag{23}$$

$$TE^{(L)}_{X\to Y|Z} = \frac{1}{2}\ln\left(\frac{var(u'_{Y,t})}{var(u''_{Y,t)}}\right) \tag{24}$$

where $u_{Y,t}$, $u'_{Y,t}$ and $u''_{Y,t}$ are the corresponding residuals calculated from three different neural networks. As for the transfer entropy for the reverse direction, it would be almost the same procedures as above except for replacing the input of the neural network.

2.3.4 Literature Review of Deep Learning Causality

Granger causality measures the extent to which the past information of one time series is predictive of another one. The majority of traditional methods for detecting Granger causality, such as the vector autoregressive (VAR) model, strictly assumes that the dependence between two time series is linear. However, in the real world, most of the time series are nonlinearly related, therefore using the traditional approaches will arrive at misleading and wrong conclusions. For instance, a research used linear Granger causality model to analyze the causal effect between greenhouse gases and the problem of global warming . The results drew a wrong conclusion that there was no Granger causality from CO_2 to temperature.[14] To handle this problem, using transfer entropy will be more ideal

giving the credit to its nonlinear and non-parametric features.

To substitute for the traditional Granger causality, transfer entropy has been widely used to study the causal effects in multiple subjects. In the research field of finance, transfer entropy is a powerful tool for exploring the contagion mechanisms of financial crisis. For example, empirical research found that the amount of information flow between industry sectors during a financial crisis became greater in Korean stock markets.[25] What's more, it was used to detect the causal relationship between the returns of financial products and their possible influence factors, such as social sentiment and cryptocurrency returns.[33] In the research field of neuroscience, when comparing with Granger causality, it has been proved that transfer entropy is a more reliable and robust measure of studying effective connectivity in electrophysiological communication.[22] In the research field of epidemiology, transfer entropy has been proved to be practically applicable for epidemiological problems, which was used to extract the correct direction of information flow between smoking and lung cancer, and obesity and diabetes risk.[27]

Although scientists have made some research in applying transfer entropy to investigate causality, most of the papers estimated transfer entropy by approximating the probability density functions. For instance, multidimensional histogram approaches are commonly employed to measure density.[33] However, there is not much literature applying neural networks to achieve this goal. Using neural networks to detect nonlinear Granger causality, which is shortened to "Deep Learning Causality" in this paper, is an emerging topic proposed in this century. To date, there is only little literature studied in this field, but it is still necessary to summarize the existing works and find the research gaps.

Regarding the same example above, a completely different conclusion was drawn when using neural networks, which provided evidence of significant unidirectional non-linear Granger causality from CO_2 to global temperature.[18] This conclusion was tallied with the actual situation. Scholars also employed neural networks to measure the transfer entropy for estimating the causal effect between the two uncertainty proxies - Economic Policy Uncertainty (EPU) and Chicago Board Options Exchange Volatility Index (VIX), and precious metals prices including gold, silver, palladium and platinum.[34] They successfully determined that gold was still the dominant 'safe-haven' asset to hedge against uncertainty. Interestingly, these precious metals were also observed to affect EPU and VIX, while they are immune to EPU shocks but not to those from the VIX. They successfully determined that only a unidirectional causal relationship existed from precious metals to EPU, which suggested that the prices of precious metals were immune to EPU confirming the 'safe-haven' feature of these assets. However, there was a bidirectional relationship between precious metals and VIX, which indicated that the prices of precious metals were sensitive to the volatility of the US equity market.

Although scholars have made some achievements in this field, the research concerned is limited because this is still relatively new research field. Based on the current situation, this paper aims to develop an open-source Python package for detecting the nonlinear Granger causality by measuring the transfer entropy using MLP. Besides studying the causality for two variables like the majority of the current literature did, this Python package contributes to investigating the causality for three variables.

2.4 The Significance of Transfer Entropy

In this paper, there is a measure called "the value of significance" Z_{TE} will be adopted to check whether our model can continuously and successfully identify the causality for the synthetic data distribution while generating different data points for it each time. This is achieved by the following three steps:

(1) Generate a dataset and calculate its TE;
(2) Shuffle the dataset to make the time series mutually independent, and calculate its $TE_{shuffle}$;
(3) Repeat step (1) and (2) for a large enough number of times (100 times in this paper), then calculate the mean and standard deviation of all $TE_{shuffle}$, and compute the value for Z_{TE} according to equation (25).

$$Z_{TE} = \frac{TE - \mu^{TE}_{shuffle}}{\sigma^{TE}_{shuffle}} \tag{25}$$

where $\mu^{TE}_{shuffle}$ represents the mean of $TE_{shuffle}$ and $\sigma^{TE}_{shuffle}$ is the standard deviation. It is worth pointing out that when dealing with three-variable cases, the Z_{TE} will be calculated on the conditional TE.

After dividing the difference by the standard deviation, the value is standardized to be a comparable scale with standard normal distribution. Shuffling the time series eliminates the dependency inside, therefore $\mu^{TE}_{shuffle}$ should be approximately zero, so Z_{TE} measures the significance of the TE results. If Z_{TE} is larger than 2.58, then it is reasonable to draw a conclusion that the result is reliable for a 99% significance interval.

2.5 The Significance of Higher-order TE

In addition to studying conditional transfer entropy, it is also worth comparing the unconditional one with it. In other words, this paper will also analyze the difference of $(TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y})$, which is referred to as "Higher-order TE" (HOTE) in this paper, to further investigate the effect of including Z on the calculation of transfer entropy and to prove that the method proposed in this paper is effective. Therefore. there is also a significance value Z_{HOTE} for the higher-order TE. It aims at detecting whether there is a significant difference between these two transfer entropies. This is achieved by going through these procedures:

(1) Calculate the difference $(TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y})$ for both non-shuffled and shuffled results, noting them as $HOTE$ and $HOTE_{shuffle}$ respectively;
(2) Compute the mean and standard deviation for $HOTE_{shuffle}$, noting them as $\mu^{HOTE}_{shuffle}$ and $\sigma^{HOTE}_{shuffle}$ respectively;
(3) Calculate the value for Z_{HOTE} according to equation (26).

$$Z_{HOTE} = \frac{HOTE - \mu^{HOTE}_{shuffle}}{\sigma^{HOTE}_{shuffle}} \tag{26}$$

Similarly, If Z_{HOTE} is larger than 2.58, then it is reasonable to draw a conclusion that the difference is significant for a 99% significance interval.

2.6 The Summary of Analytical Processes

After introducing the methodologies involved in this paper, it is more straightforward to visualize the processes using a flow chart as shown in Figure 3.

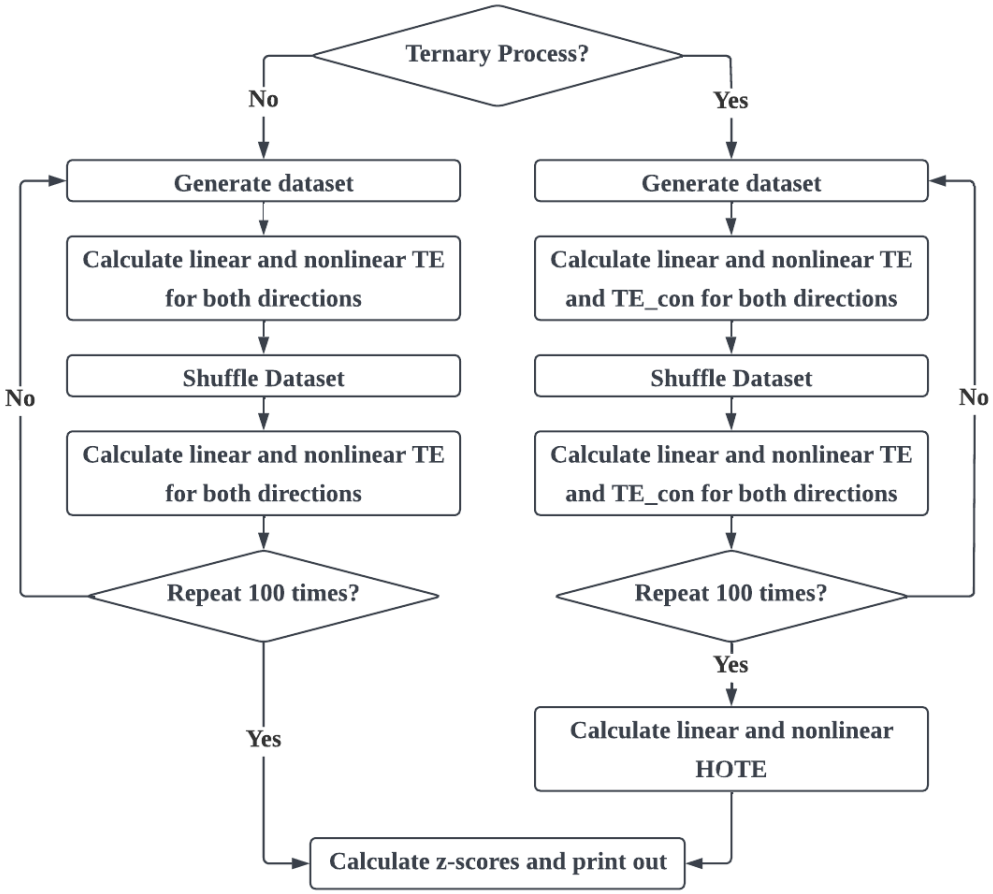

Figure 3: Flow chart for the analytical processes

3. DATA GENERATION

In this paper, there are four synthetic distributions generated to be used for validating the effectiveness of the model. Two of them are two-variable distributions and the others are three-variable distributions.

3.1 Coupled Wiener Process

A wiener process (also called as Brownian Motion) is one of the most important stochastic processes, which is named in honor of a remarkable mathematician Norbert Wiener.[17, 30]

A wiener process $W(t)$ has the following properties:

(1) $W(0) = 0$.
(2) $W(t)$ satisfies a normal distribution with mean 0 and variance t. In other words, $W(t) \sim N(0, t)$.

In the course of probability theory, the wiener process is acting as the base for many limit theorems and is also a basic tool for modeling many phenomena involving randomness. It has been applied frequently in different fields like mathematics, economics, finance and physics. Therefore, building up a coupled dataset based on wiener processes is of much practical significance. First, let's construct two independent time series $X(t)$ and $V(t)$ following the equations below:

$$X(t) = W_1(t) = \sqrt{t}Z_1 \tag{27}$$
$$V(t) = W_2(t) = \sqrt{t}Z_2 \tag{28}$$

where Z_1 and Z_2 are independent standard normal distribution $N(0, 1)$. After generating these two sequences, another time series $Y(t)$ will be generated by incorporating the past value of X and the current value of V (equation (29)).

$$Y(t) = \alpha X(t - L) + (1 - \alpha)V(t) \tag{29}$$

where α measures the degree of dependency (casual effect) between X_{t-L} and Y_t.

3.2 Coupled Logistic Map

The logistic map is firstly published in 1976 by a biologist, which is represented by a simple quadratic recurrence equation but exhibits great complexity.[13]

$$X_t = f(X_{t-1}) = \gamma X_{t-1}(1 - X_{t-1}) \tag{30}$$

where X_t is the value of X at time t, γ is a positive influencing factor (similar to the growth rate) and f is the mapping function. To achieve the goal of generating another time series

Y_t dependent to X_t, the model proposed by Hahs and Pethel is used in this paper:[19]

$$Y_t = (1-\alpha)f(Y_{t-1}) + \alpha g(X_{t-1}) \tag{31}$$

where α is the coupling strength in range of $[0,1]$, $g(X_{t-1}) = (1-\epsilon)f(X_{t-1}) + \epsilon f^2(X_{t-1})$ is the the coupling function and ϵ is the coupling strength in $g(X)$. It's worth noting that the notation f^2 does not represent a second-order derivative but instead, it means the mapping function f is used twice, suggesting that there exists nonlinear causality from X_{t-1} to Y_t. Meanwhile, as $f^2(X_{t-1}) = f(f(X_{t-1})) = f(X_t) = X_{t+1}$, therefore, equation (31) can be rewritten into the form below:

$$Y_t = (1-\alpha)f(Y_{t-1}) + \alpha(1-\epsilon)f(X_{t-1}) + \alpha\epsilon X_{t+1} \tag{32}$$

which shows linear causality from Y_t to X_{t+1}. However, due to the great complexity of the logistic map, it is hard to tell whether Y_t nonlinearly causes X_{t+1} or not by simply looking at the equation, which will be one of the aspects that the model will try to figure out later.

Although theoretically, γ can take any positive number, it is recommended to set $\gamma = 4$, which will make the map to be a chaotic logistic map.[19]

3.3 Ternary Wiener Process

Similarly, as the coupled case explained above, the first two wiener processes $Z(t)$ and $V(t)$ are generated as below:

$$Z(t) = W_1(t) = \sqrt{t}Z_1 \tag{33}$$

$$V(t) = W_2(t) = \sqrt{t}Z_2 \tag{34}$$

where Z_1 and Z_2 are independent standard normal distribution $N(0,1)$. With a slight difference from before, there is another variable $X(t)$ generated by incorporating a wiener process and the current value of Z.

$$X(t) = (1-\phi)W_3(t) + \phi Z(t) = (1-\phi)\sqrt{t}Z_3 + \phi Z(t) \tag{35}$$

where ϕ is the weight for $Z(t)$. Afterwards, the final time series $Y(t)$ is generated by the following equation:

$$Y(t) = \alpha X(t-L) + \beta Z(t-L) + (1-\alpha-\beta)V(t) \tag{36}$$

where α is the factor measuring how much $X(t-L)$ will contributes to the current $Y(t)$, the other two coefficients do the similar functions.

3.4 Ternary Logistic Maps

To generate a ternary logistic map, similarly, firstly X_t is computed by using the following equation:

$$X_t = f(X_{t-1}) = \gamma X_{t-1}(1-X_{t-1}) \tag{37}$$

Then another time series Z_t is generated based on using a uniform distribution U ranging in [0,1].

$$Z_t = U \tag{38}$$

After getting the values for both X_t and Z_t, eventually it is able to calculate for Y_t, which is designed to further incorporate Z_{t-1} into calculation.

$$Y_t = (1 - \alpha)f(Y_{t-1}) + \alpha g(X_{t-1})(Z_{t-1})^2 \tag{39}$$

where function f and g remain the same as ones in the coupled case above.

When using these three data distributions for the models, the dataset will be split into a training set and a testing set for building a reliable and unbiased model. In this paper, the previous 70% of data points will be used for training the model, and the rest 30% of data points will be used for testing the accuracy and calculating for TE and significance Z.

4. PYTHON IMPLEMENTATION

One of the meaningful contributions of this paper is to develop a user-friendly Python package called "DLcausality", which stands for Deep Learning Causality. In the following sections, a comprehensive introduction of its functions will be given to help readers know more about its structure and usage.

4.1 Package Installation

Users should first download the files for the package from Github before starting to install it. They should turn on the terminal under the directory where they plan to keep the files and then input the following command:

```
git clone https://github.com/oliviayt1224/Deep-Learning-Causality
```

Listing 1: Cloning files from Github

After downloading the files to the local side, users can install the package by using the codes below:

```
pip install .
```

Listing 2: Installing the package

Afterwards, the package is installed into the Python working-environment along with all the prerequisite packages and then users are able to use all the commands set up in the package.

4.2 Command Introduction

There are four different commands implemented in this package: "cwp", "clm", "twp" and "tlm", each of them refers to one of the synthetic data distributions mentioned before and is introduced individually in the following sections. When a command is successfully executed, the results will be printed out in the terminal with all relative z-scores.

4.2.1 Command "cwp"

This command enables users to investigate the causality for coupled wiener processes, and the codes for executing it have been shown in the block below. There are five parameters that users can specify their values, which are summarized in table 1.

Parameter	Definition	Type	Range	Default Value
T	the total time length of the series	float	$(0, +\infty)$	1
N	the number of time steps	integer	$(0, +\infty)$	300
alpha	the coefficient measures the degree of dependency between X_{t-L} and Y_t	float	$[0, 1]$	0.5
lag	the number of time-lag	integer	$(0, +\infty)$ smaller than N	5
num_exp	the number of experiments running for calculating the value of significance	integer	$(0, +\infty)$	100

Table 1: Parameters for command "cwp"

Users can execute the command by following the syntax:

```
cwp --T <time length> --N <time steps> --alpha <coefficient> --lag <time
   lag> --num_exp <number of experiments>
```

Listing 3: Executing the command "cwp" with specific input values

Since every single parameter has a corresponding default value, therefore it is not necessary for users to specify a number for each of them. The easiest way to use this command is simply just to use all the defaults values as inputs by doing:

```
cwp
```

Listing 4: Executing the command "cwp" with all default values

4.2.2 Command "clm"

This command enables users to investigate the causality for coupled logistic maps, and the codes for executing it have been shown in the block below. There are five parameters that users can specify their values, which are summarized in table 2.

Parameter	Definition	Type	Range	Default Value
T	the total time length of the series	float	$(0, +\infty)$	1
N	the number of time steps	integer	$(0, +\infty)$	1000
alpha	the coupling strength	float	$[0, 1]$	0.4
epsilon	the coupling strength	float	$[0, 1]$	0.9
num_exp	the number of experiments running for calculating the significance value	integer	$(0, +\infty)$	100

Table 2: Parameters for command "clm"

In fact, generating the coupled logistic map requires specific initial values for X and Y, but users do not need to input the initial values in this command. It is because for each experiment, the program will generate a random number ranging in $(0, 1)$ as the initial value of X or Y. The reason for doing it is that if the program keeps using the same initial numbers, then the dataset for each experiment will be exactly the same, which makes it impossible to calculate the z-scores. Users can execute the command by following the syntax:

```
clm --T <time length> --N <time steps> --alpha <coefficient> --epsilon <
    coefficient> --num_exp <number of experiments>
```

Listing 5: Executing the command "clm" with specific input values

Similarly, this command can also be executed without specifying the input values:

```
clm
```

Listing 6: Executing the command "clm" with all default values

4.2.3 Command "twp"

This command enables users to investigate the causality for ternary wiener processes, and the codes for executing it have been shown in the block below. There are seven parameters that users can specify their values, which are summarized in table 3.

Parameter	Definition	Type	Range	Default Value
T	the total time length of the series	float	$(0, +\infty)$	1
N	the number of time steps	integer	$(0, +\infty)$	300
alpha	the coefficient measures the degree of dependency between X_{t-L} and Y_t	float	$[0, 1]$	0.5
phi	the coefficient measures the degree of dependency between X_t and Z_t	float	$[0, 1]$	0.5
beta	the coefficient measures the degree of dependency between Z_{t-L} and Y_t	float	$[0, 1]$	0.5
lag	the number of time-lag	integer	$(0, +\infty)$ smaller than N	5
num_exp	the number of experiments running for calculating the value of significance	integer	$(0, +\infty)$	100

Table 3: Parameters for command "twp"

Users can execute the command by following the syntax:

```
twp --T <time length> --N <time steps> --alpha <coefficient> --phi <
    coefficient> --beta <coefficient> --lag <time lag> --num_exp <number
    of experiments>
```

Listing 7: Executing the command "twp"

Similarly, this command can also be executed without specifying the input values:

```
twp
```

Listing 8: Executing the command "twp" with all default values

4.2.4 Command "tlm"

This command enables users to investigate the causality for ternary logistic maps, and the codes for executing it has been shown in the block below. There are five parameters that users can specify their values, which are summarized in table 4.

Parameter	Definition	Type	Range	Default Value
T	the total time length of the series	float	$(0, +\infty)$	1
N	the number of time steps	integer	$(0, +\infty)$	700
alpha	the coupling strength	float	$[0, 1]$	0.4
epsilon	the coupling strength	float	$[0, 1]$	0.9
num_exp	the number of experiments running for calculating the significance value	integer	$(0, +\infty)$	100

Table 4: Parameters for command "tlm"

Users can execute the command by following the syntax below:

```
tlm --T <time length> --N <time steps> --alpha <coefficient> --epsilon <
    coefficient> --num_exp <number of experiments>
```

Listing 9: Executing the command "tlm" with specific input values

Similarly, this command can also be executed without specifying the input values:

```
tlm
```

Listing 10: Executing the command "tlm" with all default values

4.3 Unit Tests

In order to check whether all functions work as expected and also check whether readable messages about errors show up when inappropriate inputs are used in the commands, there are various unit tests set up in the package to help verify the results, which can be easily executed by running the code below:

```
pytest -v
```

Listing 11: Running all unit tests

5. RESULT AND DISCUSSION

After setting up the package, it is time to shift focus to analyzing the results. In order to test the performance of the proposed method in measuring Granger causality, the following sections will perform sensitive tests for parameters in the models for each data distribution. In other words, there will be only one parameter changed for each section while the others keep using the same values (default values as introduced in the tables above unless specified).

5.1 Results for Coupled Wiener Processes

5.1.1 Experiments of varying N

When the number of time steps N ranging from 100 to 1000, it is obvious that either linear or nonlinear measure is able to detect the significant existence of Granger causality from X to Y under the convenience interval of 99%. It is supported by the fact that $\overline{Z}_{TE}$ (mean of z-scores) is larger than 2.58 for every N in Figure 4. There is an overall upward trending showing in curves when N takes a larger number. It indicates that both models will have better capabilities for capturing the causal effect within the datasets. Besides, $\overline{Z}_{TE}$ calculated by using linear measures is larger than the nonlinear one, which is reasonable because the variables in the coupled wiener processes are linearly dependent with each other. This also proves the effectiveness of the model.

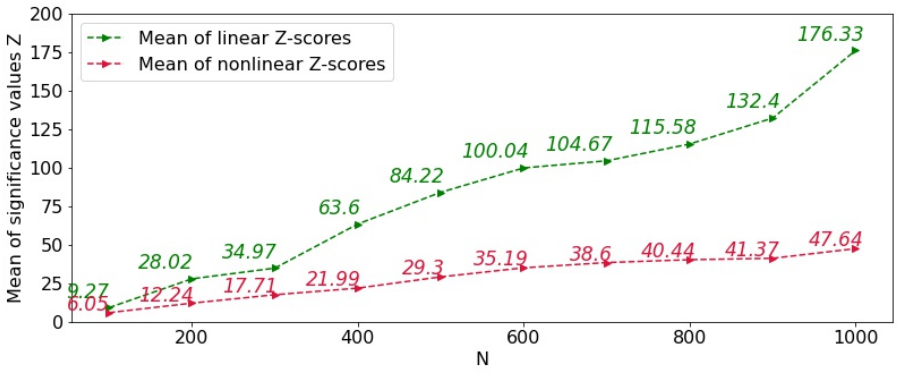

Figure 4: Mean of Z-scores for $TE^{(L)}_{X\to Y}$ for coupled wiener processes (varying N while T=1, $\alpha = 0.5$ and lag=5)

However, when reproducing the analytical procedures reversely, all $\overline{Z}_{TE}$ fluctuate up and down around 0, and none of them is larger than 2.58 or smaller than -2.58 (Figure 5). It suggests that there is no significant evidence proving the existence of Granger causality from Y to X. Combined with what we observed before, the results show that there is only a unidirectional causal relationship from X to Y. This exactly matches the feature of the data distribution because the lagged value of X is used to generate Y. Again, it proves the accuracy of the model.

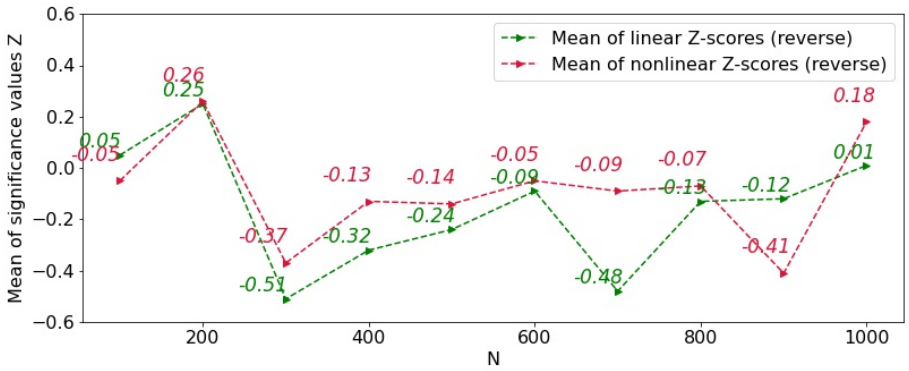

Figure 5: Mean of Z-scores for $TE^{(L)}_{Y\to X}$ for coupled wiener processes (varying N while T=1, $\alpha = 0.5$ and lag=5)

5.1.2 Experiments of varying alpha

When varying the coefficient α between X_{t-L} and Y_t from 0.1 to 0.9, similarly two curves present general uptrends. It is theoretically interpretable because alpha measures the dependence degree between X and Y. Therefore, as alpha gets larger, X and Y have a stronger dependency (causality), which should result in larger values of transfer entropy and then Z_{TE}. There are two special cases where the results are not significant when alpha takes 0.1 or 0.2, it is because the correlation between X and Y is not strong enough when alpha is small, which causes the model incapable to detect the significant causality.

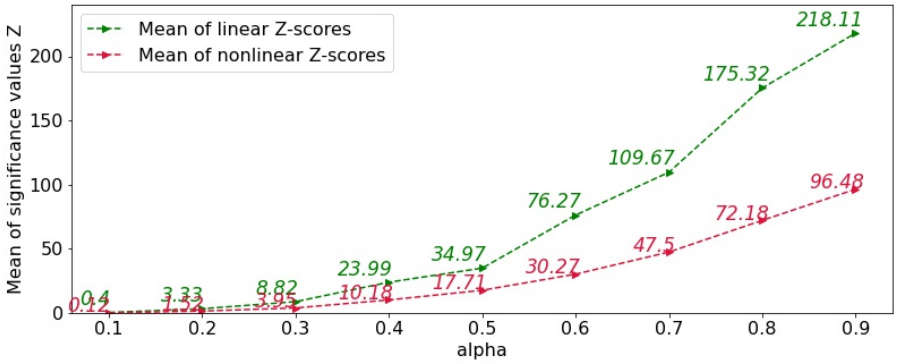

Figure 6: Mean of Z-scores for $TE^{(L)}_{X\to Y}$ for coupled wiener processes (varying alpha while T=1, N=300 and lag=5)

If assigning a larger N into the model, even though alpha is small, it is still possible that the model can output significant results. As seen in Figure 7, while keeping alpha equal to 0.2, both measures get significant results when N is larger than 500.

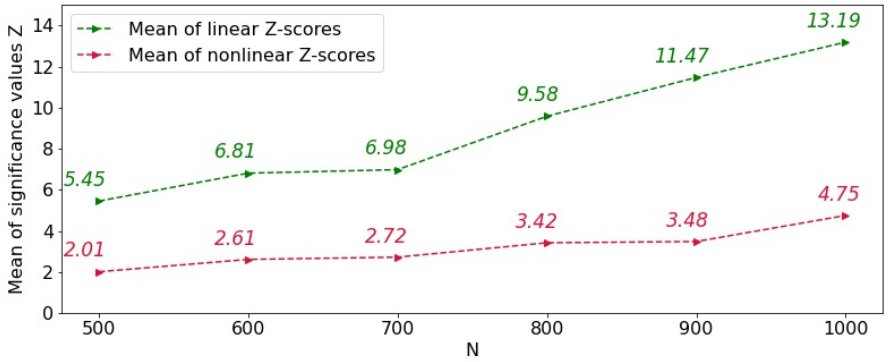

Figure 7: Mean of Z-scores for $TE^{(L)}_{X\to Y}$ for coupled wiener processes (varying N while T=1, $\alpha = 0.2$ and lag=5)

Similarly, there is still no significant causality from Y to X observed in the dataset no matter which value that α takes (Figure 8).

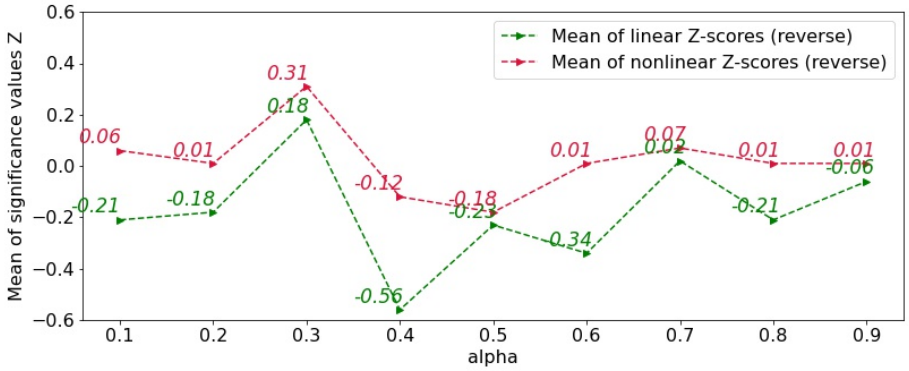

Figure 8: Mean of Z-scores for $TE^{(L)}_{Y\to X}$ for coupled wiener processes (varying alpha while T=1, N=300 and lag=5)

5.2 Results for Coupled Logistic Maps

5.2.1 Experiments of varying N

When it comes to coupled logistic maps, $\overline{Z}_{TE}$ is also upward trending while increasing the number of N. However, unlike coupled ternary processes, the linear results are below the nonlinear one in this case, which is because that coupled logistic map is a nonlinear distribution. The results support that the model is applicable for capturing the causality for coupled logistic maps. The nonlinear model detects significant causality much earlier than the linear one does. As it has been shown in Figure 9, the nonlinear curve gets higher than 2.58 since $N = 300$ while the linear one could not get significant results until $N = 400$. The reason behind it is that when the correlation is relatively small, more data points are needed for the models in order to enable them to detect the causal patterns.

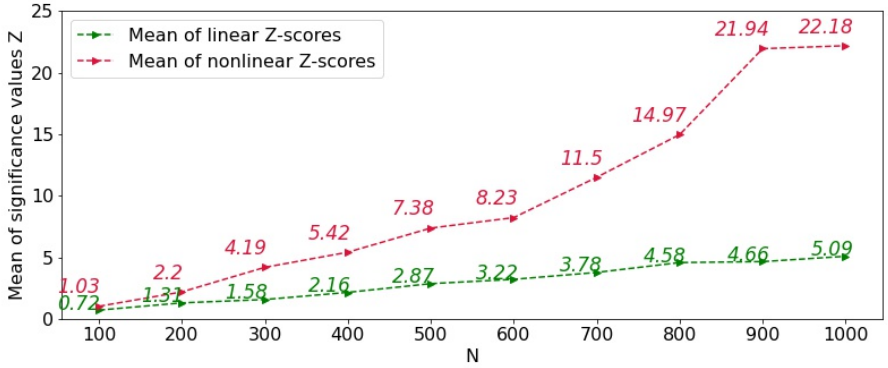

Figure 9: Mean of Z-scores for $TE^{(L)}_{X\to Y}$ for coupled logistic maps (varying N while T=1, $\alpha = 0.4$ and $\epsilon = 0.9$)

However, when reproducing the analytical procedures reversely, all $\overline{Z}_{TE}$ are larger than 2.58 (Figure 10). It suggests that there exists a significant Granger causal effect from Y to X. Combined with what we observed above, the results show that there are bidirectional causal relations between X and Y. This exactly matches the feature of this data distribution as we have shown in section 3.2 that X_{t-1} is used to generate Y_t while Y_t is used to produce X_{t+1}. Unlike the nonlinear dependence between X_{t-1} and Y_t, the correlation between Y_t and X_{t+1} is actually linear. That is why the curve of the linear measure is above the other.

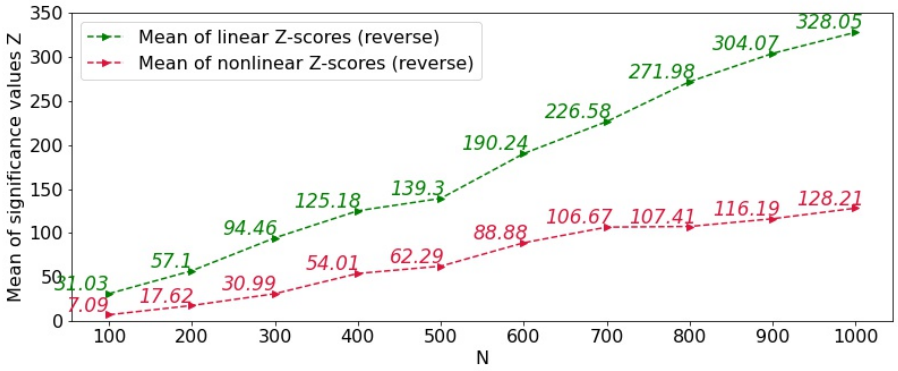

Figure 10: Mean of Z-scores for $TE^{(L)}_{Y\to X}$ for coupled logistic maps (varying N while T=1, $\alpha = 0.4$ and $\epsilon = 0.9$)

Besides, when comparing the results in Figure 9 and 10, over all the values of $\overline{Z}_{TE}$ in Figure 10 are much greater than the ones in Figure 9. It suggests that although the causal effects between X and Y are bidirectional, the causal link from Y to X is stronger. To explain it, first let's recall the equation for coupled logistic maps.

$$Y_t = (1-\alpha)f(Y_{t-1}) + \alpha(1-\epsilon)f(X_{t-1}) + \alpha\epsilon X_{t+1}$$
$$Y_t = (1-\alpha)f(Y_{t-1}) + 4\alpha(1-\epsilon)X_{t-1}(1-X_{t-1}) + \alpha\epsilon X_{t+1}$$

Now, if we replace α and ϵ with their values in this example and calculate for the two coefficients in front of X_{t-1} and X_{t+1}, then the equation will be in this form:

$$Y_t = 0.6f(Y_{t-1}) + 0.16X_{t-1}(1-X_{t-1}) + 0.36X_{t+1}$$

It is understandable that the absolute value of the coefficient reflects the strength of causality between variables. The coefficient in front of the X_{t-1} term is only 0.16, which is

smaller than the coefficient (0.36) between Y_t and X_{t+1}. In this case, as Figure 9 is the results of the causality from X_{t-1} to Y_t and Figure 10 is the results of the causality from Y_t to X_{t+1}, therefore it is reasonable that the values of $\overline{Z}_{TE}$ in Figure 10 are much larger. All these results above prove the accuracy of the model.

5.2.2 Experiments of varying alpha

Regarding the experiments of varying the coupling strength α, these two curves shows entirely different trends. As alpha gets larger, the curve of nonlinear results experiences first a slight decrease but then a sharp increase while the curve of linear results does the opposite. The nonlinear measure can produce significant results at all conditions of alpha, but the linear measure starts to be ineffective once alpha gets larger than 0.5. Considering this is a nonlinear distribution, it is expected that the nonlinear measure performs better. This suggests that the linear measure does not always work and users should try both methods if they are not sure about the data distribution.

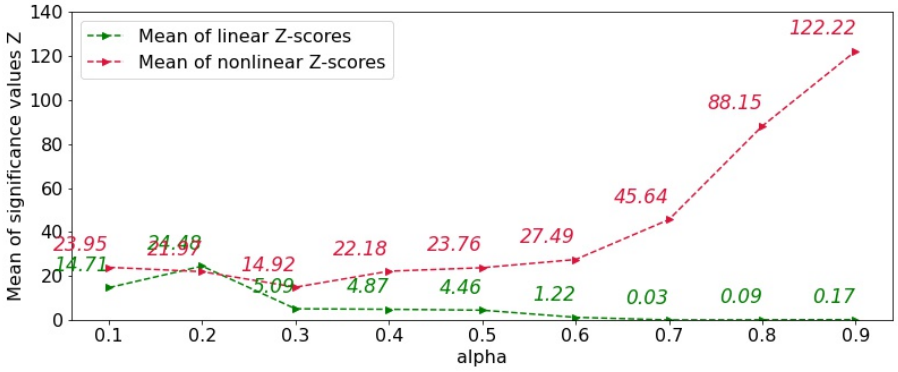

Figure 11: Mean of Z-scores for $TE^{(L)}_{X\to Y}$ for coupled logistic maps (varying alpha while T=1, N=1000 and $\epsilon = 0.9$)

When analyzing the causality in the reverse direction, all $\overline{Z}_{TE}$ are larger than 2.58 (Figure 12) indicating that the model successfully detects significant causality from Y to X. For most of the cases, $\overline{Z}_{TE}$ obtained from the linear method is larger than the one from the nonlinear method owing to the linear dependence between Y_t and X_{t+1}. However, when α is relatively small like 0.1 or 0.2, the reverse seems to be the case. This suggests that when the coefficient is small, the linear method has less capability in detecting causality than the nonlinear one.

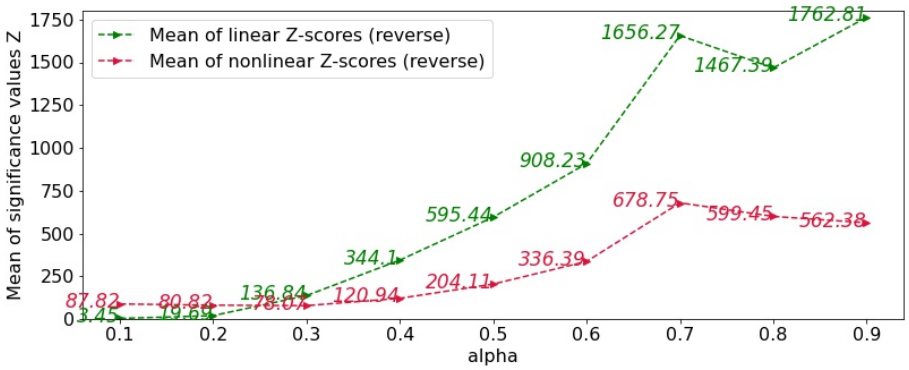

Figure 12: Mean of Z-scores for $TE^{(L)}_{Y\to X}$ for coupled logistic maps (varying alpha while T=1, N=1000 and $\epsilon = 0.9$)

5.2.3 Experiments of varying epsilon

When varying the coefficient ϵ in the coupling function, similar patterns can be observed in Figure 13 compared with Figure 11, the nonlinear method can always obtain significant results and also outperforms the linear one in most cases.

When analyzing for $TE_{Y\to X}$, the nonlinear measure performs better than the linear one when ϵ is smaller than 0.4, the opposite holds true when ϵ is larger. When $\epsilon = 0.1$, the $\overline{Z}_{TE}$ from the linear measure is not significant although it is supposed to be, while the $\overline{Z}_{TE}$ from the nonlinear measure is significant. The reason beyond it might be the incapability of the linear model with handling weak dependence. Besides, when $\epsilon = 0$, equation (32) tells that there should be no linear causality in this case, which accords with the fact that $\overline{Z}_{TE}$ is insignificant. However, the result from the nonlinear method shows that there still exists nonlinear Granger causality from Y_t to X_{t+1} when $\epsilon = 0$. This actually answers the question that is left in the previous section about whether Y_t nonlinearly causing X_{t+1} or not. Therefore, it can be concluded that Y_t is a nonlinear Granger cause for X_{t+1} according to the empirical results.

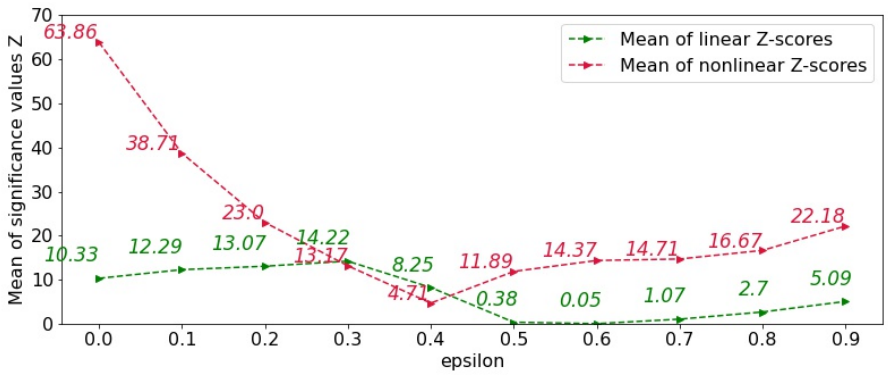

Figure 13: Mean of Z-scores for $TE^{(L)}_{X\to Y}$ for coupled logistic maps (varying epsilon while T=1, N=1000 and $\alpha = 0.4$)

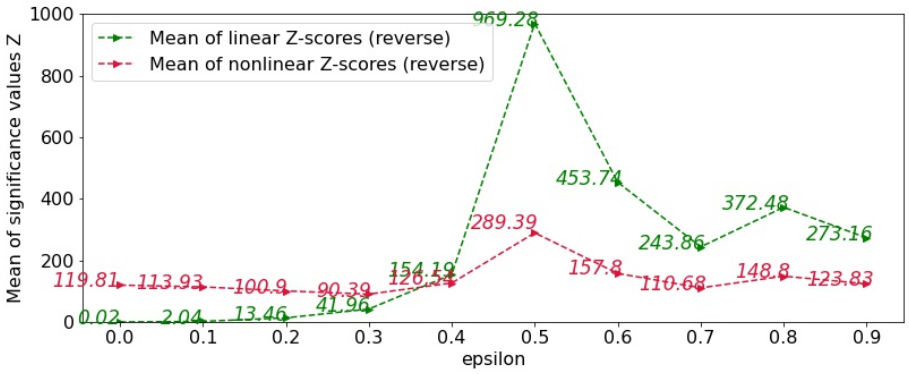

Figure 14: Mean of Z-scores for $TE^{(L)}_{Y\to X}$ for coupled logistic maps (varying epsilon while T=1, N=1000 and $\alpha = 0.4$)

5.3 Results for Ternary Wiener Processes

The ternary wiener process is the first three-variable distribution that is discussed in this paper. The additional statistic, when compared with two-variable distributions, is the mean of z-scores for the difference between the conditional and unconditional transfer entropies $(TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y})$ after running 100 experiments. The notation of this statistic is $\overline{Z}_{HOTE}$.

5.3.1 Experiments of varying N

When varying N from 100 to 500, similarly there is an uptrend showing in both linear and nonlinear cases and all points in Figure 15 are significant. It suggests that the information of Z is useful for improving both linear and nonlinear regression models.

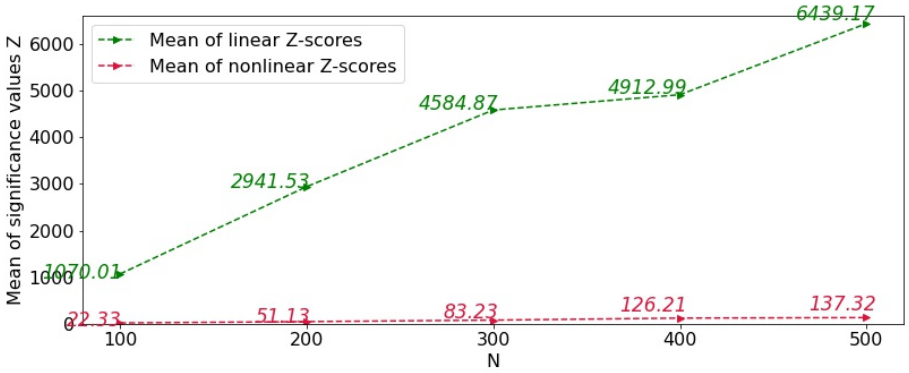

Figure 15: Mean of Z-scores for $TE^{(L)}_{X\to Y|Z}$ for ternary wiener processes (varying N while T=1, $\alpha = \phi = \beta = 0.5$ and lag=5)

To explain for it, we need to recall the equations for calculating $TE^{(L)}_{X\to Y|Z}$:

$$Y_t = c + dY_{t-L} + eX_{t-L} + u'_{Y,t}$$
$$Y_t = f + gY_{t-L} + hX_{t-L} + iZ_{t-L} + u''_{Y,t}$$

$$TE^{(L)}_{X\to Y|Z} = \frac{1}{2}\ln\left(\frac{var(u'_{Y,t})}{var(u''_{Y,t})}\right)$$

When generating the process, the value of Z_{t-L} is considered into the calculation for Y_t, which suggests that there should be a causal link from Z_{t-L} to Y_t. Therefore, the regression model, which includes Z_{t-L}, should outperform the other one, meaning that $u''_{Y,t}$ should be less than $u'_{Y,t}$. In this case, $TE^{(L)}_{X\to Y|Z}$ will be larger than zero and then $\overline{Z}_{TE}$ is significant. Meanwhile, the curve of the linear methods in Figure 15 is always above the nonlinear one, which is verifiable because the ternary wiener process is a linear distribution.

However, when reproducing the calculation of $\overline{Z}_{TE}$ for $TE^{(L)}_{Y\to X|Z}$, all values in Figure 16 do not satisfy the requirement of significance. It means that including Z into the regression model does not help.

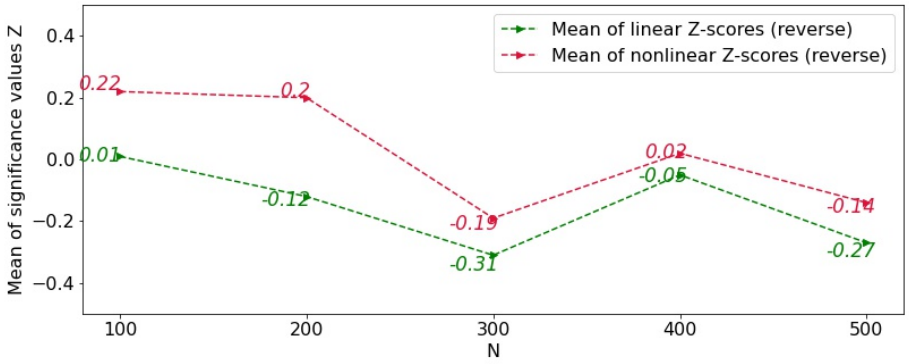

Figure 16: Mean of Z-scores for $TE^{(L)}_{Y\to X|Z}$ for ternary wiener processes (varying N while T=1, $\alpha = \phi = \beta = 0.5$ and lag=5)

To understand it, let's start with the calculation of $TE^{(L)}_{Y\to X|Z}$:

$$X_t = c + dY_{t-L} + eX_{t-L} + u'_{X,t}$$
$$X_t = f + gY_{t-L} + hX_{t-L} + iZ_{t-L} + u''_{X,t}$$

$$TE^{(L)}_{Y\to X|Z} = \frac{1}{2}\ln\left(\frac{var(u'_{X,t})}{var(u''_{X,t})}\right)$$

When generating the process, the past information of Z is never used in the calculation for the current X. Therefore, including the term of Z_{t-L} into the regression model would not bring any favorable effect to the regression accuracy, which means that $u'_{X,t}$ and $u''_{X,t}$ will be quite similar. In this case, $TE^{(L)}_{Y\to X|Z}$ will be close to zero and therefore, the results of $\overline{Z}_{TE}$ will be insignificant.

As for $\overline{Z}_{HOTE}$, all results are significant proving that $TE^{(L)}_{X\to Y|Z}$ significantly differs from $TE^{(L)}_{X\to Y}$. In Figure 17, all numbers are positive and larger than 2.58 indicating that $TE^{(L)}_{X\to Y|Z}$ is significantly larger than $TE^{(L)}_{X\to Y}$. Referencing the equations for the calculations of these two TE, $TE^{(L)}_{X\to Y|Z} > TE^{(L)}_{X\to Y}$ is equivalent to $\ln\left(\frac{var(u'_{Y,t})}{var(u''_{Y,t})}\right) > \ln\left(\frac{var(u_{Y,t})}{var(u'_{Y,t})}\right)$, therefore $(var(u'_{Y,t}))^2 > var(u_{Y,t})var(u''_{Y,t})$. This inequality holds because the accuracy of the model's prediction is improved resulting from a $var(u''_{Y,t})$ smaller than $var(u'_{Y,t})$ when incorporating the lagged values of Z into regression. All the evidence mentioned above proves the effectiveness of the model.

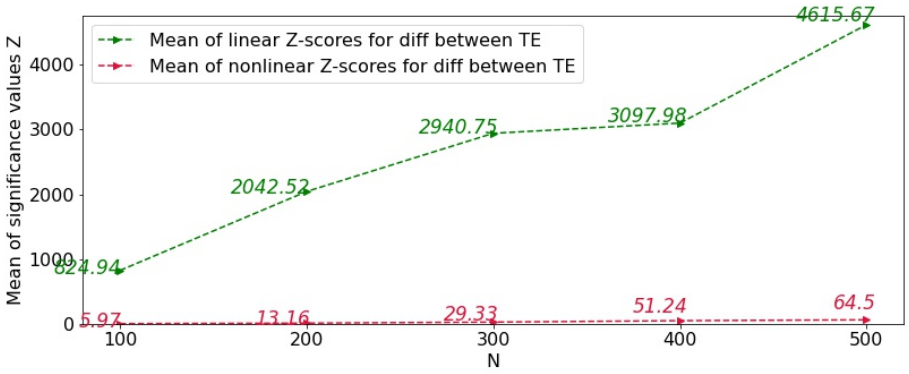

Figure 17: Mean of Z-scores for difference $(TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y})$ for ternary wiener processes (varying N while T=1, $\alpha = \phi = \beta = 0.5$ and lag=5)

5.3.2 Experiments of varying alpha

When varying the coefficient α between X_{t-L} and Y_t from 0.3 to 0.7, both curves in Figure 18 reach their peaks when alpha is equal to 0.5. Both methods are able to obtain significant results for all values of alpha while the curve for the linear method is always above, which demonstrates the strength of the model.

Similarly, no enhancement within the regression is detected from adding Z into the models (Figure 19) and all $\overline{Z}_{HOTE}$ are significant (Figure 20) indicating that incorporating Z effectively helps to forecast Y.

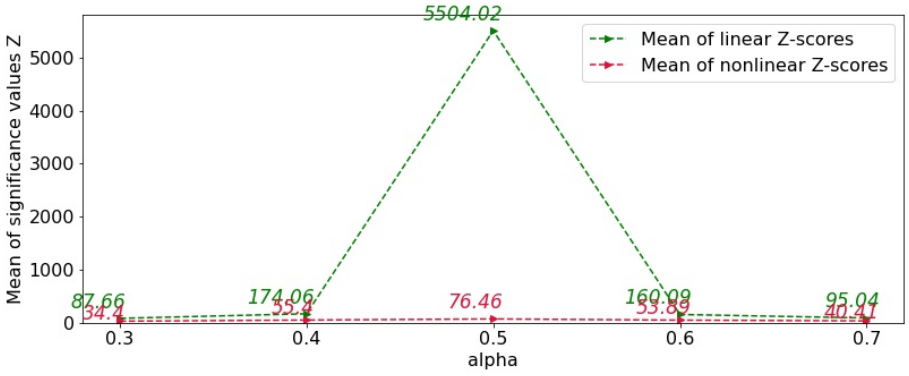

Figure 18: Mean of Z-scores for $TE^{(L)}_{X\to Y|Z}$ for ternary wiener processes (varying α while T=1, N=300, $\phi = \beta = 0.5$ and lag=5)

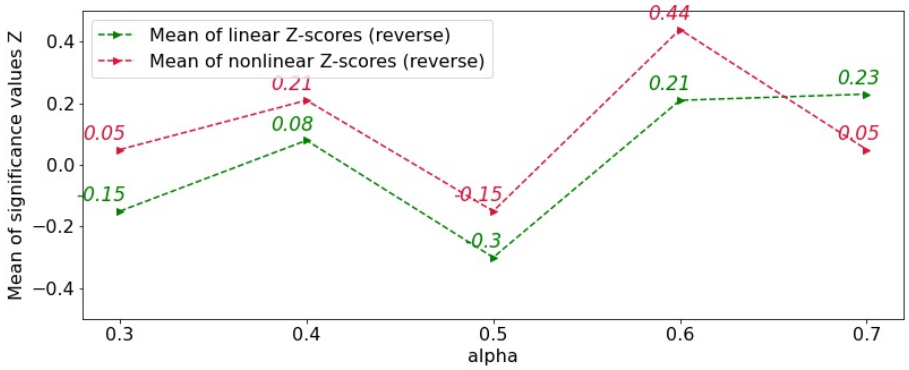

Figure 19: Mean of Z-scores for $TE^{(L)}_{Y\to X|Z}$ for ternary wiener processes (varying α while T=1, N=300, $\phi = \beta = 0.5$ and lag=5)

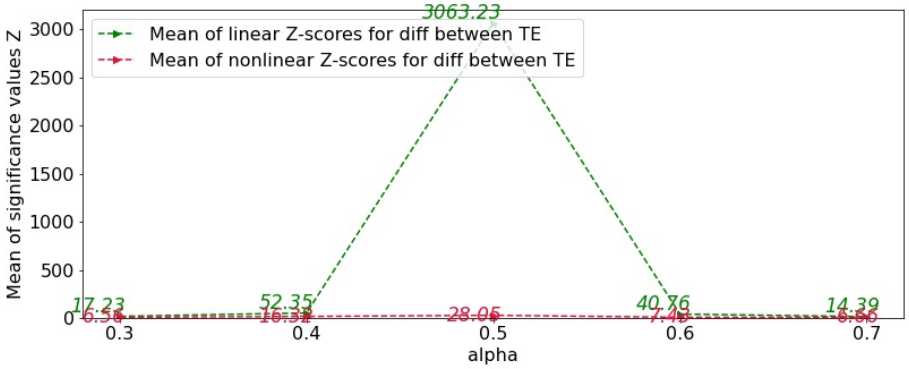

Figure 20: Mean of Z-scores for difference $(TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y})$ for ternary wiener processes (varying α while T=1, N=300, $\phi = \beta = 0.5$ and lag=5)

5.3.3 Experiments of varying phi

When varying the coefficient ϕ between Z_t and X_t from 0.3 to 0.7, in Figure 21, the curve of linear measures goes into a valley when $\phi = 0.5$. However, no matter which value it takes, the model can always detect significant improvement from the regression by incorporating the values of Z.

From Figure 22, all results are insignificant proving that there is no significant change in the regression for X by considering Z into the model. Besides, all $\overline{Z}_{HOTE}$ are significant in Figure 23, indicating that incorporating Z effectively helps forecasting Y.

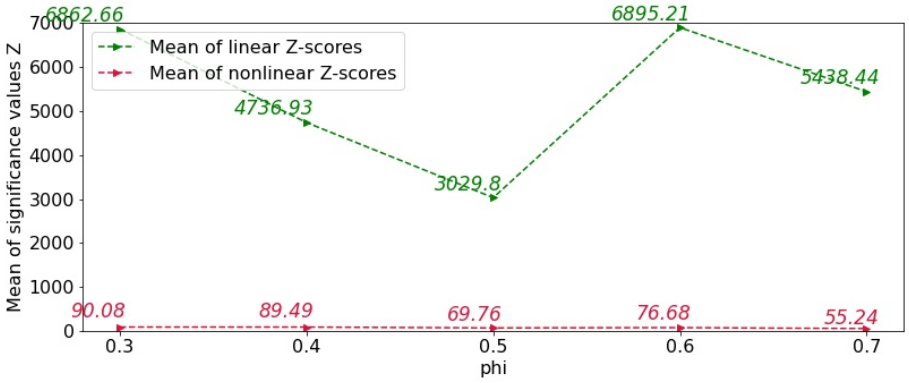

Figure 21: Mean of Z-scores for $TE^{(L)}_{X\to Y|Z}$ for ternary wiener processes (varying ϕ while T=1, N=300, $\alpha = \beta = 0.5$ and lag=5)

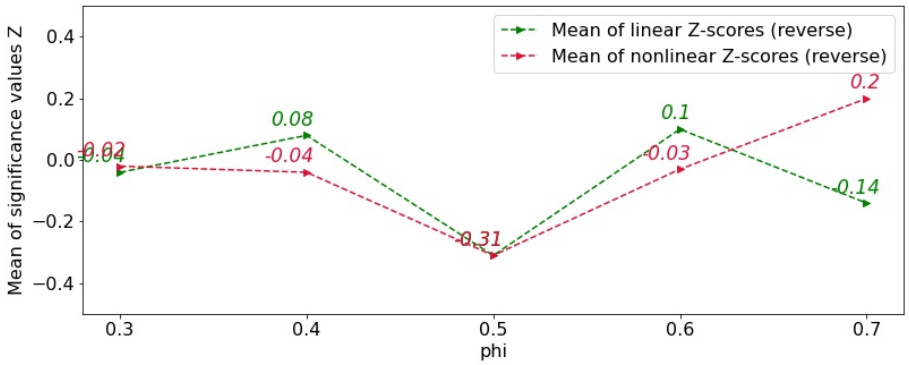

Figure 22: Mean of Z-scores for $TE^{(L)}_{Y\to X|Z}$ for ternary wiener processes (varying ϕ while T=1, N=300, $\alpha = \beta = 0.5$ and lag=5)

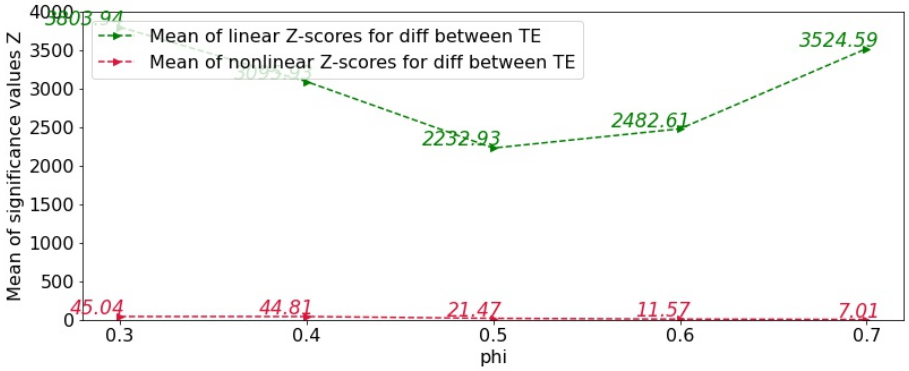

Figure 23: Mean of Z-scores for difference ($TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y}$) for ternary wiener processes (varying ϕ while T=1, N=300, $\alpha = \beta = 0.5$ and lag=5)

5.3.4 Experiments of varying beta

When varying the coefficient β between Z_{t-L} and Y_t from 0.3 to 0.7, all graphs show exactly the same conclusions as the previous examples, indicating the effectiveness of the models again.

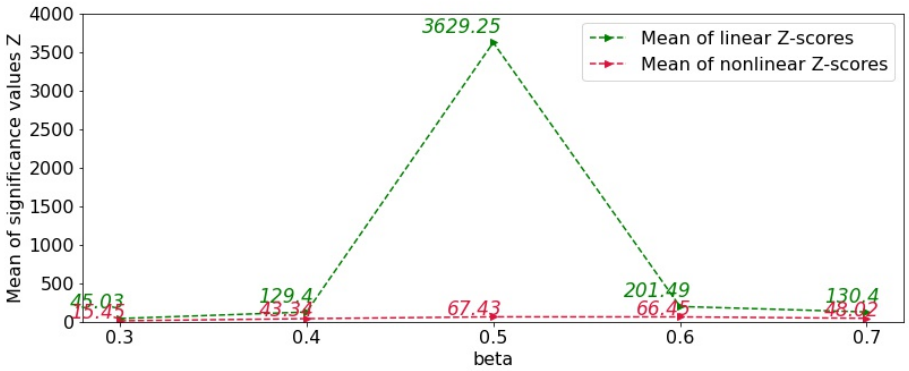

Figure 24: Mean of Z-scores for $TE^{(L)}_{X\to Y|Z}$ for ternary wiener processes (varying β while T=1, N=300, $\alpha = \phi = 0.5$ and lag=5)

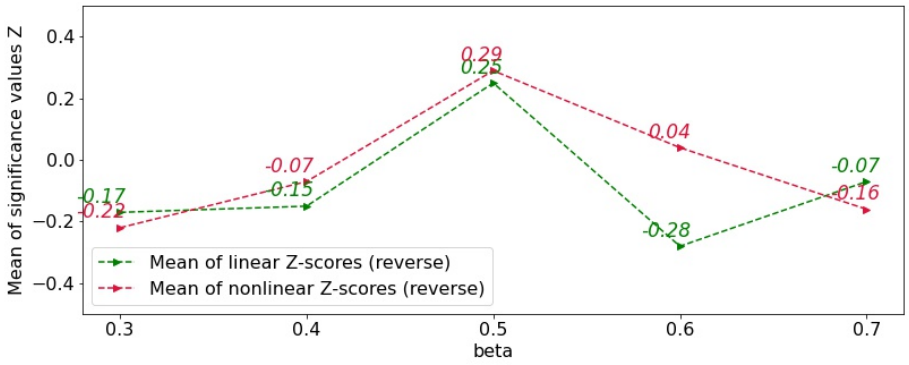

Figure 25: Mean of Z-scores for $TE^{(L)}_{Y\to X|Z}$ for ternary wiener processes (varying β while T=1, N=300, $\alpha = \phi = 0.5$ and lag=5)

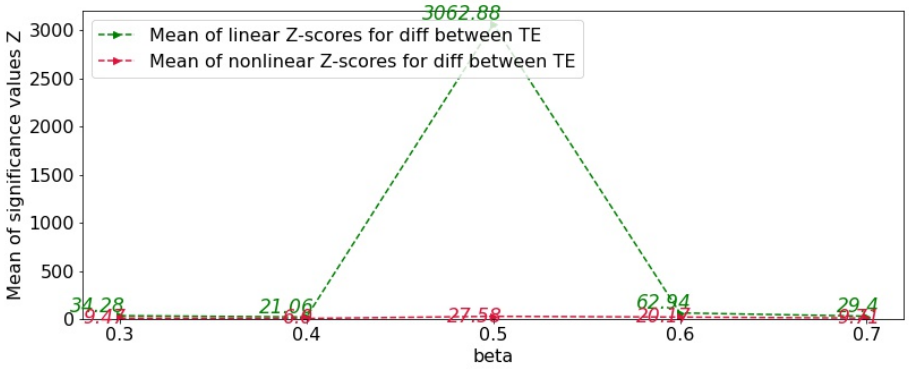

Figure 26: Mean of Z-scores for difference ($TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y}$) for ternary wiener processes (varying β while T=1, N=300, $\alpha = \phi = 0.5$ and lag=5)

5.4 Results for Ternary Logistic Maps

5.4.1 Experiments of varying N

As for ternary logistic maps, when varying N from 500 to 900, all experiments can obtain significant results proving the advantage of including Z for forecasting Y as shown in Figure 27. The ternary logistic map is theoretically a nonlinear distribution so the curve of nonlinear results is supposed to be above the linear one. However, the model gives a different situation. There are several potential reasons beyond it: (1) this paper only adopts a very simple MLP

model for the nonlinear modeling and does not further investigate solutions for improving its performance; (2) the values of X, Y and Z are all in range of (0,1), there is only little change in values between time steps. In this case, it brings difficulty to neural networks for understanding the patterns. These will be some valuable aspects for further study.

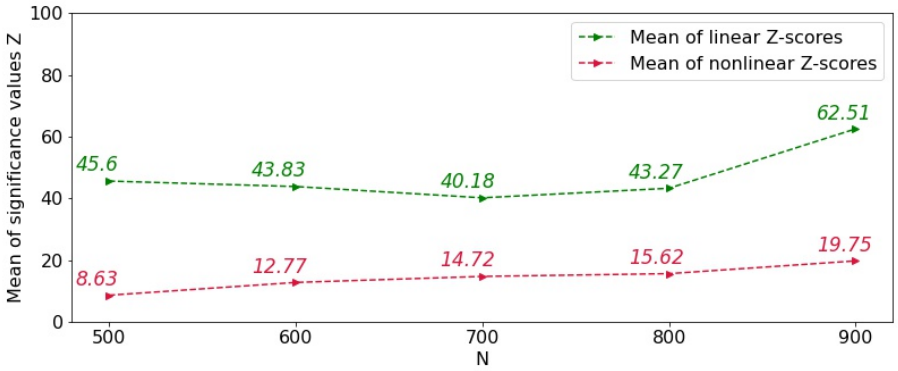

Figure 27: Mean of Z-scores for $TE^{(L)}_{X\to Y|Z}$ for ternary logistic maps (varying N while T=1, $\alpha = 0.4$, $\phi = 0.9$ and lag=5)

Results in Figure 28 tells that Z can not boost the prediction ability of forecasting X for the models. When generating the ternary logistic maps, the time series Z is randomly-generated standard normal distribution, which independent to X. Therefore, including Z_{t-L} into the regression is useless for improving the model's accuracy.

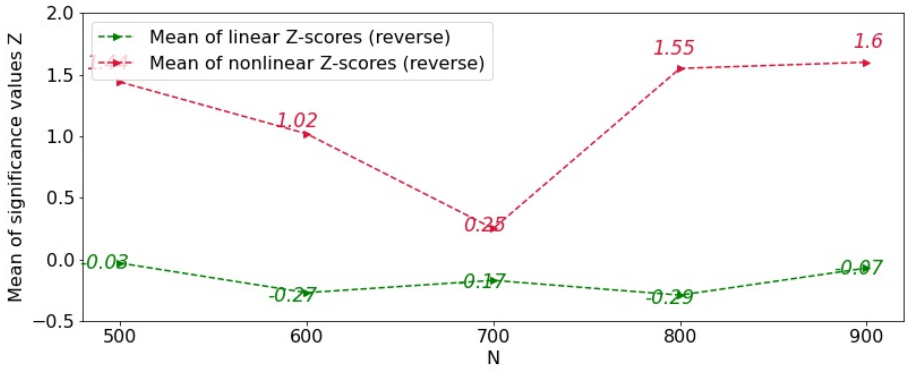

Figure 28: Mean of Z-scores for $TE^{(L)}_{Y\to X|Z}$ for ternary logistic maps (varying N while T=1, $\alpha = 0.4$, $\phi = 0.9$ and lag=5)

Afterwards, all $\overline{Z}_{HOTE}$ in Figure 29 are significant indicating that the lagged values of Z play significant roles in forecasting Y. All conclusions above verify the abilities of the models in detecting Granger causality.

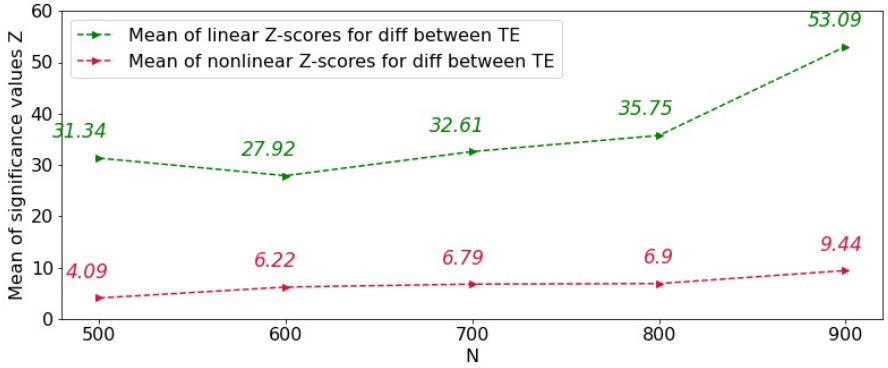

Figure 29: Mean of Z-scores for difference ($TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y}$) for ternary logistic maps (varying N while T=1, $\alpha = 0.4$, $\phi = 0.9$ and lag=5)

5.4.2 Experiments of varying alpha

When varying the coupling strength α from 0.2 to 0.6, all values of $\overline{Z}_{TE}$ in Figure 30 are much greater than 2.58 meaning that the model of predicting Y is improved when it is given the information of Z. Same as before, there is no enhancement after inputting Z into the model for X (Figure 31) and adding lagged Z into the regression model for Y effectively decreases the fitting errors (Figure 32).

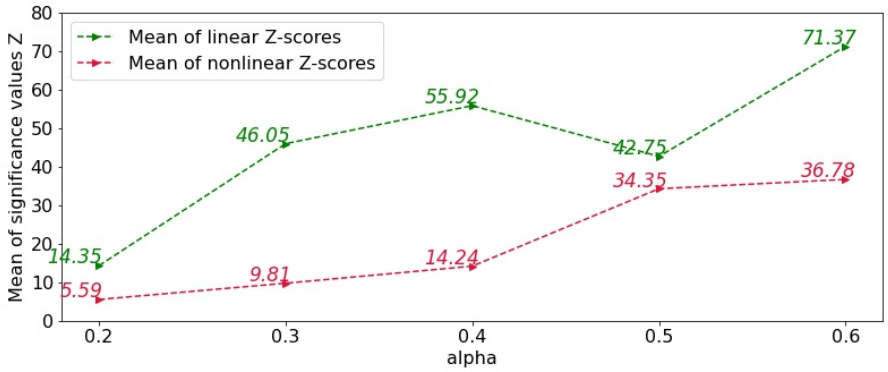

Figure 30: Mean of Z-scores for $TE^{(L)}_{X\to Y|Z}$ for ternary logistic maps (varying α while T=1, N=700, $\epsilon = 0.9$ and lag=5)

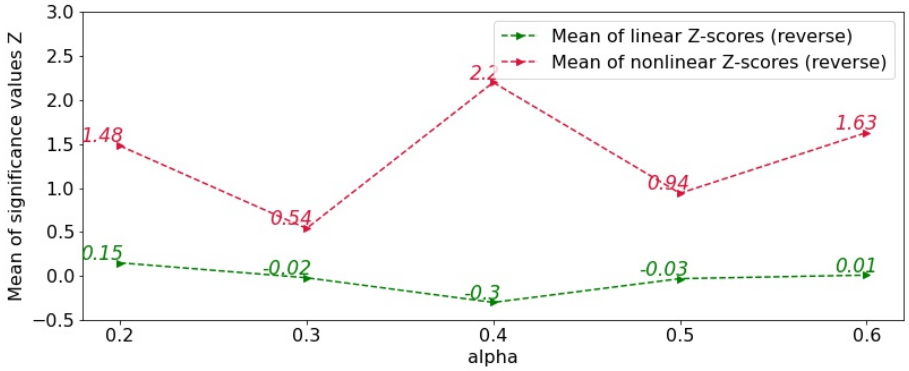

Figure 31: Mean of Z-scores for $TE^{(L)}_{Y\to X|Z}$ for ternary logistic maps (varying α while T=1, N=700, $\epsilon = 0.9$ and lag=5)

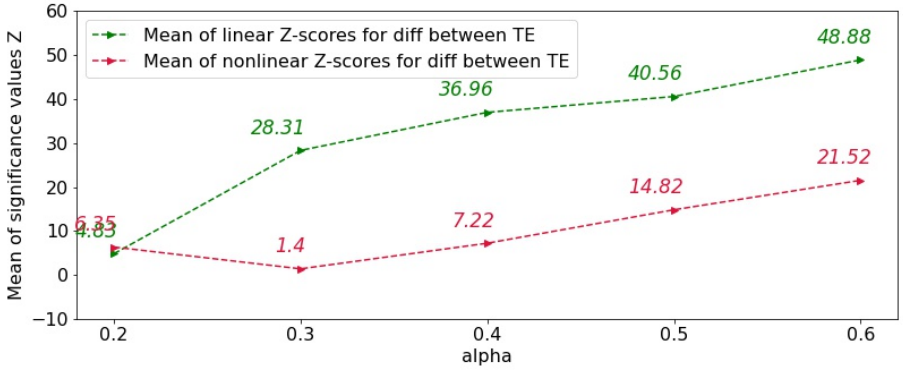

Figure 32: Mean of Z-scores for difference ($TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y}$) for ternary logistic maps (varying α while T=1, N=700, $\epsilon = 0.9$ and lag=5)

5.4.3 Experiments of varying epsilon

When varying the coefficient ϵ in the coupling function from 0.5 to 0.9, all results in Figure 33, Figure 34 and Figure 35 match the conclusions from previous examples.

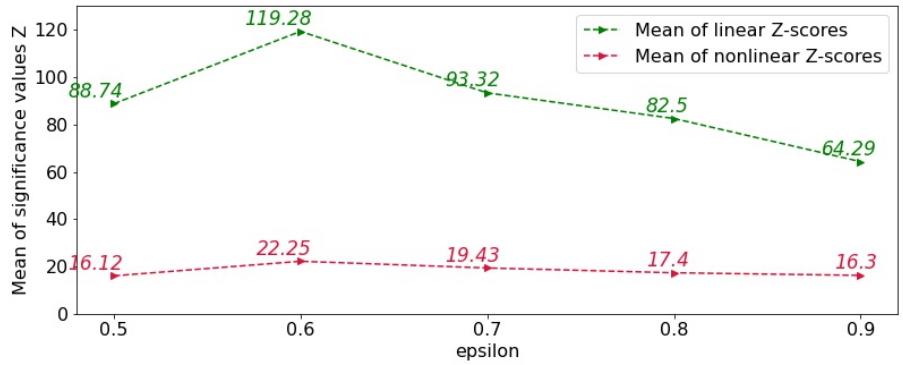

Figure 33: Mean of Z-scores for $TE^{(L)}_{X\to Y|Z}$ for ternary logistic maps (varying ϵ while T=1, N=700, $\alpha = 0.4$ and lag=5)

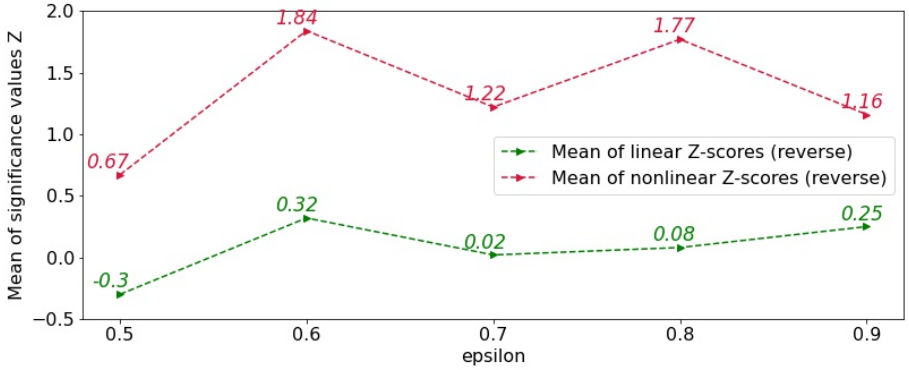

Figure 34: Mean of Z-scores for $TE^{(L)}_{Y\to X|Z}$ for ternary logistic maps (varying ϵ while T=1, N=700, $\alpha = 0.4$ and lag=5)

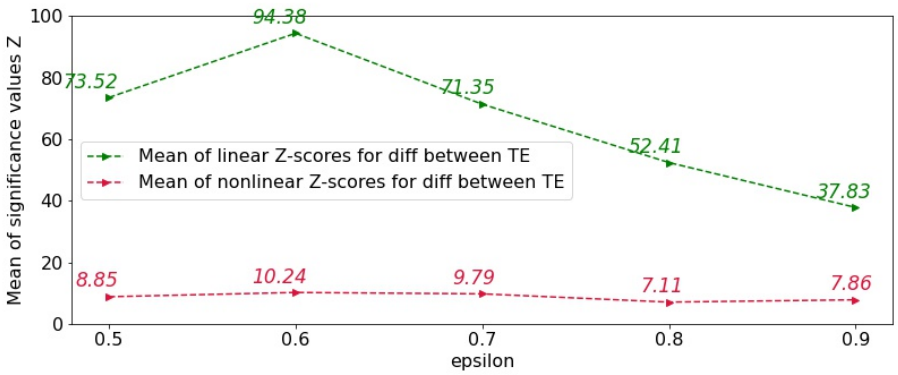

Figure 35: Mean of Z-scores for difference ($TE^{(L)}_{X\to Y|Z} - TE^{(L)}_{X\to Y}$) for ternary logistic maps (varying ϵ while T=1, N=700, $\alpha = 0.4$ and lag=5)

5.5 Reasons of Default Value Choices

The reason why those default values are chosen is due to the purpose of avoiding the required time for running the model to be too long while ensuring the model to have enough information for capturing the causality. For example, for coupled logistic maps, the default value of N is 1000. Although it takes a long time to run the model for 1000 time steps, it is still worthy because the model might not be able to detect significant Granger causality if the number of data points is not large enough. However, as for ternary logistic maps, the significance values Z computed from the models when N=700 are already much larger than 2.58 and increasing N does not make a huge difference to Z, therefore, keeping N=700 as the default value is reasonable.

6. CONCLUSION

Studying the causal links between things is a vital process for people to understand, explain, predict, decide and control them in the world. However, causality is notoriously difficult to understand due to its complexity.[28] Even the definition of causality has kept philosophers arguing for over two thousand years and has not yet been resolved. The studies of causality had not made great progress until Granger proposed the measure called Granger causality to quantify causality. Granger causality has been widely adopted in research in various fields and received popular recognition. However, this method has a fatal limitation of being only applicable for linear dependence. which causes it to be incapable to capture the causality in the real world as most of the real cases are nonlinear.

Therefore, to deal with this problem, a non-parametric measure called transfer entropy was used to replace the traditional Granger causality in this paper. Regarding the detailed calculation for transfer entropy, the vector autoregressive models and neural networks were employed to measure the linear and nonlinear causality respectively, and an open-resource Python package was developed for users to conveniently build up these models. In this paper, we studied the causality for not only two-variable cases but also three-variable cases. The performance of the models was comprehensively analyzed by adopting sensitivity tests for parameters. The results suggested that the proposed models were effective in detecting both linear and nonlinear Granger causality for the majority of cases and also successfully investigated the effects on transfer entropy of incorporating a third variable into the regression models. However, regarding the ternary logistic map, there were some peculiar outcomes that lacked further understanding and explanation. For example, although the models detected significant causality from X to Y, the results from linear methods were unexpectedly larger than the ones from nonlinear methods. As the ternary logistic map is very complex, it remains a question why this situation occurred. Further analysis for it could focus on studying the behaviors of the ternary logistic map and adjusting the structure of the MLP for improvements.

Undoubtedly, this paper has made remarkable contributions to the current research, especially the novel idea of detecting nonlinear Granger causality using neural networks. To date, there is quite limited literature about this topic, therefore, the work in this paper could provide important reference value. However, due to the time limits, there are still some unsolved problems remaining for future improvement, which have been listed in the last section Future Scope.

7. FUTURE SCOPE

Although the proposed model in this paper has been proved to be effective for either linear or nonlinear distribution, however, there are still some perspectives that have not yet been covered and will be ideal choices for future research.

(1) Dataset Choices: In this paper, there are only four synthetic data distributions involved. It was quite challenging to design ternary processes for fitting into the models, therefore, it is worth exploring more ternary choices for validating the effectiveness of the models in the future. Although the model is applicable to the four distributions, it will be even better if it can be applied to real-world datasets for practical purposes. However, real-world datasets will be more complex and always contain noises or outliers, which means that the same set-up of the model might not work anymore and further adjustments will be required.

(2) Trend Explanation: In this paper, our main focus is on whether the results are significant or not. However, a more specific and detailed explanation for the curve trends should also be paid attention to. For example, there are some cases where the curve reaches a peak or goes down into a valley. The reasons beyond these trends are worth investigating to see whether they are explicable and reasonable.

(3) Neural Network: In this paper, the MLP structure is pretty simple with only one hidden layer and 100 neurons in it. Besides, only one type of activation functions called Rectified Linear Unit (ReLU) is adopted and there is no parameter adjustment in this paper. This limitation causes the problem mentioned in the result analysis for ternary logistic maps before, therefore it is worth trying different neural networks, comparing their performance and choosing the best one in the future.

(4) Python Package: In this paper, there is a useful and open Python package developed for users. Users can try all the examples introduced in this paper and even have a chance to self-specify the values for any parameter. However, when implementing the model in Python, the time efficiency of running models is not an element considered. In some cases, if the number of experiments or time steps is large, it might take about 30 mins for the computer to output the results. In the future, more actions of improving the running speed should be taken. For instance, users should be provided choices to use GPU instead of CPU, which can achieve sustainable speedup.[24]

REFERENCES

[1] F. Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain". In: *Psychological Review* (1958), pp. 65–386.

[2] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3 (1969), pp. 424–438. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912791.

[3] Marvin Minsky and Seymour Papert. "Perceptrons". In: (1969).

[4] Seppo Linnainmaa. "Taylor expansion of the accumulated rounding error". In: *BIT Numerical Mathematics* 16.2 (1976), pp. 146–160.

[5] R. Ashley, C. W. J. Granger, and R. Schmalensee. "Advertising and Aggregate Consumption: An Analysis of Causality". In: *Econometrica* 48.5 (1980), pp. 1149–1167. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912176.

[6] C.W.J. Granger. "Testing for causality: A personal viewpoint". In: *Journal of Economic Dynamics and Control* 2 (1980), pp. 329–352. ISSN: 0165-1889. DOI: https://doi.org/10.1016/0165-1889(80)90069-X. URL: https://www.sciencedirect.com/science/article/pii/016518898090069X.

[7] Richard G. Sheehan and Robin Grieves. "Sunspots and Cycles: A Test of Causation". In: *Southern Economic Journal* 48.3 (1982), pp. 775–777. ISSN: 00384038. URL: http://www.jstor.org/stable/1058669.

[8] John F. Geweke. "Measures of Conditional Linear Dependence and Feedback Between Time Series". In: *Journal of the American Statistical Association* 79.388 (1984), pp. 907–915. ISSN: 01621459, 1537274X. URL: http://www.jstor.org/stable/2288723.

[9] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[10] George Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[11] Thomas Schreiber. "Measuring Information Transfer". In: *Phys. Rev. Lett.* 85 (2 July 2000), pp. 461–464. DOI: 10.1103/PhysRevLett.85.461. URL: https://link.aps.org/doi/10.1103/PhysRevLett.85.461.

[12] David Hume. *A treatise of human nature*. Courier Corporation, 2003.

[13] Robert M May. "Simple mathematical models with very complicated dynamics". In: *The Theory of Chaotic Attractors*. Springer, 2004, pp. 85–93.

[14] Umberto Triacca. "Is Granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature?" In: *Theoretical and Applied Climatology* 81 (July 2005), pp. 133–135. DOI: 10.1007/s00704-004-0112-1.

[15] M.D. Weir and J. Hass. *Thomas' Calculus*. Pearson Addison Wesley, 2008. ISBN: 9788131718674. URL: https://books.google.co.uk/books?id=ps8zduMZW5AC.

[16] Lionel Barnett, Adam B. Barrett, and Anil K. Seth. "Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables". In: *Phys. Rev. Lett.* 103 (23 Dec. 2009), p. 238701. DOI: 10.1103/PhysRevLett.103.238701. URL: https://link.aps.org/doi/10.1103/PhysRevLett.103.238701.

[17] Tamás Szabados. "An elementary introduction to the Wiener process and stochastic integrals". In: *arXiv preprint arXiv:1008.1510* (2010).

[18] A Attanasio and U Triacca. "Detecting human influence on climate using neural networks based Granger causality". In: *Theoretical and Applied Climatology* 103.1 (2011), pp. 103–107.

[19] Daniel W Hahs and Shawn D Pethel. "Distinguishing anticipation from causality: Anticipatory bias in the estimation of information flow". In: *Physical review letters* 107.12 (2011), p. 128701.

[20] Ramon Quiza and J. Davim. "Computational Methods and Optimization". In: Jan. 2011, pp. 177–208. ISBN: 978-1-84996-449-4. DOI: 10.1007/978-1-84996-450-0.

[21] Katerina Schindlerova. "Equivalence of Granger Causalityand Transfer Entropy: A Generalization." In: 2011.

[22] Raul Vicente et al. "Transfer Entropy–a Model-Free Measure of Effective Connectivity for the Neurosciences". In: *J. Comput. Neurosci.* 30.1 (Feb. 2011), pp. 45–67. ISSN: 0929-5313. DOI: 10.1007/s10827-010-0262-3. URL: https://doi.org/10.1007/s10827-010-0262-3.

[23] Claude Lemaréchal. "Cauchy and the gradient method". In: *Doc Math Extra* 251.254 (2012), p. 10.

[24] Vincent Boyer and Didier El Baz. "Recent Advances on GPU Computing in Operations Research". In: *2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum*. 2013, pp. 1778–1787. DOI: 10.1109/IPDPSW.2013.45.

[25] Gabjin Oh et al. "An information flow among industry sectors in the Korean stock market". In: *Journal of the Korean Physical Society* 65 (2014), pp. 2140–2146.

[26] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural Networks* 61 (2015), pp. 85–117. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2014.09.003. URL: https://www.sciencedirect.com/science/article/pii/S0893608014002135.

[27] N. Ahmad Aziz. "Transfer entropy as a tool for inferring causality from observational studies in epidemiology". In: *bioRxiv* (2017). DOI: 10.1101/149625. eprint: https://www.biorxiv.org/content/early/2017/06/14/149625.full.pdf. URL: https://www.biorxiv.org/content/early/2017/06/14/149625.

[28] Mukesh Dalal, Amy Sliva, and David Blumstein. "Complex Causality: Computational Formalisms, Mental Models, and Objective Truth". In: *International Conference on Applied Human Factors and Ergonomics*. Springer. 2017, pp. 108–120.

[29] Deniz Gencaga. "Transfer Entropy". In: *Entropy* 20 (Apr. 2018), p. 288. DOI: 10.3390/e20040288.

[30] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019.

[31] Brian Lindner et al. "Comparative analysis of Granger causality and transfer entropy to present a decision flow for the application of oscillation diagnosis". In: *Journal of Process Control* 79 (2019), pp. 72–84. ISSN: 0959-1524. DOI: https://doi.org/10.1016/j.jprocont.2019.04.005. URL: https://www.sciencedirect.com/science/article/pii/S095915241830516X.

[32] Rana Pratap Maradana et al. "Innovation and economic growth in European Economic Area countries: The Granger causality approach". In: *IIMB Management Review* 31.3 (2019), pp. 268–282. ISSN: 0970-3896. DOI: https://doi.org/10.1016/j.iimb.2019.03.002. URL: https://www.sciencedirect.com/science/article/pii/S097038961630101X.

[33] Zac Keskin and Tomaso Aste. "Information-theoretic measures for nonlinear causality detection: application to social media sentiment and cryptocurrency prices". In: *Royal Society Open Science* 7 (Sept. 2020). DOI: 10.1098/rsos.200863.

[34] Toan Luu Duc Huynh. "The effect of uncertainty on the precious metals market: New insights from Transfer Entropy and Neural Network VAR". In: *Resources Policy* 66 (2020), p. 101623. ISSN: 0301-4207. DOI: https://doi.org/10.1016/j.resourpol.2020.101623. URL: https://www.sciencedirect.com/science/article/pii/S0301420719309365.

[35] Ali Shojaie and Emily B. Fox. "Granger Causality: A Review and Recent Advances". In: *Annual Review of Statistics and Its Application* 9.1 (2022), null. DOI: 10.1146/annurev-statistics-040120-010930. eprint: https://doi.org/10.1146/annurev-statistics-040120-010930. URL: https://doi.org/10.1146/annurev-statistics-040120-010930.