Cauchy-Schwarz Divergence Transfer Entropy

Zhaozhao Ma[⋆,†]
[⋆]*Zhejiang University*
[†]*Georgia Institute of Technology*
zhaozhaoma@gatech.edu

Shujian Yu[‡,§]
[‡]*Vrije Universiteit Amsterdam*
[§]*UiT - The Arctic University of Norway*
s.yu3@vu.nl

***Abstract*—Transfer entropy (TE) is a powerful information-theoretic tool for analyzing causality in time series and complex systems. In this work, we propose a new formulation of TE using the Cauchy-Schwarz (CS) divergence. The resulting CS-TE offers a closed-form estimator and naturally extends to capture more complex causal relationships, such as indirect causation and synergistic effects, beyond just pairwise interactions. We also explore the feasibility of using a classifier, rather than regression models, to perform Granger tests in a supervised way. Lastly, we demonstrate the effectiveness of CS-TE on benchmark simulated data and stock indices from 14 stock markets. The code and supplementary material are available in our project repository: https://github.com/SJYuCNEL/Cauchy-Schwarz-Transfer-Entropy.**

***Index Terms*—Cauchy-Schwarz divergence, transfer entropy, multivariate time series, supervised learning, causality**

I. Introduction

Granger Causality (GC) [1], a foundational concept in time series analysis, is a causal inference method by examining whether one time series provides predictive power for the future values of another time series beyond the information contained in the latter's values. It has been widely used in finance [2], neuroscience [3], and social sciences [4], etc.

According to GC, a time series $\{x_t\}$ "Granger-causes" (or "G-causes") another time series $\{y_t\}$, then:

$$\mathbb{E}\left(p(y_{t+1}|y_{t-})\right) = \mathbb{E}\left(p(y_{t+1}|y_{t-}, x_{t-})\right), \tag{1}$$

where x_{t-} and y_{t-} refer to, respectively, the past observations of x_t and y_t. Usually, $x_{t-} = [x_t, x_{t-\tau}, \cdots, x_{t-(m-1)\tau}]$, which is known as the delay-coordinate embedding or Taken's embedding [5], in which τ is the delay and m is the embedding dimension [6]. Similarly, $y_{t-} = [y_t, y_{t-\tau}, \cdots, y_{t-(n-1)\tau}]$.

To test if the two conditional distributions $p(y_{t+1}|y_{t-})$ and $p(y_{t+1}|y_{t-}, x_{t-})$ are identical, traditional methods use statistical testing to determine whether the two conditional means $\mathbb{E}(y_{t+1}|y_{t-})$ and $\mathbb{E}(y_{t+1}|y_{t-}, x_{t-})$ and are equal. To this end, a regression model will be selected to fit the data. Notable examples include the linear vector autoregression (VAR) models [1], the kernel machines [7], the radial basis function (RBF) neural networks [8], and the more complicated long-short term memory networks (LSTMs) [9]. However, selecting an appropriate regression model and its hyperparameters (e.g., the depth of a neural network) is challenging in practice [10].

An alternative is to use nonlinear, model-agnostic information-theoretic measures such as transfer entropy (TE) [11] and partial transfer entropy (PTE) [12]. Formally, TE quantifies the deviation from $p(y_{t+1}|y_{t-})$ to $p(y_{t+1}|y_{t-}, x_{t-})$ with the expected Kullback-Leibler (KL) divergence:

$$\begin{aligned}
&\mathbb{E}\left[\log\left(\frac{p(y_{t+1}|y_{t-}, x_{t-})}{p(y_{t+1}|y_{t-})}\right)\right] \\
&= \iiint p(y_{t+1}, y_{t-}, x_{t-}) \log\left(\frac{p(y_{t+1}|y_{t-}, x_{t-})}{p(y_{t+1}|y_{t-})}\right) dy dy_{t-} dx_{t-} \\
&= -\mathbb{E}\left(\log p(y_{t-}, x_{t-})\right) + \mathbb{E}\left(\log p(y_{t+1}, y_{t-}, x_{t-})\right) \\
&\quad - \mathbb{E}\left(\log p(y, y_{t-})\right) + \mathbb{E}\left(\log p(y_{t-})\right) \\
&= H(Y_{-1}, X_{-1}) - H(Y, Y_{-1}, X_{-1}) + H(Y, Y_{-1}) - H(Y_{-1}),
\end{aligned} \tag{2}$$

where H denotes Shannon entropy or joint entropy. The primary advantage of such an information-theoretic functional for detecting causality is that, unlike GC, it does not rely on any specific regression model. Interestingly, GC and TE are equivalent when the underlying processes are Gaussian [13].

The authors of [14] generalize TE to Rényi transfer entropy (RTE), by replacing the Shannon entropy terms in Eq. (2) with the Rényi entropy [15], [16]. They show that Rényi entropy selectively highlights certain parts of the underlying empirical distribution while strongly suppressing others, making it a valuable tool for time series analysis, particularly in finance. However, RTE can yield negative values, leading to challenges in interpretation. Additionally, both TE and RTE are difficult to estimate, especially for high-dimensional and noisy data.

This work proposes a new formulation of TE by substituting the KL divergence in Eq. (2) with the Cauchy-Schwarz (CS) divergence [17]–[19]. This substitution immediately offers the advantages of easy estimation, computational stability, and complementary insights into the KL divergence-based TE (for more details, interested readers can refer to the supplementary material). Our main contributions include:

- We propose CS-TE and derive its closed-form estimator;
- We extend CS-TE to multivariate time series and further propose CS divergence-based conditional transfer entropy (CS-CTE) and CS divergence-based joint transfer entropy (CS-JTE). These new measures can identify complex causal relationships, including indirect causality and synergistic effects (see examples in Fig. 1);
- We demonstrate the feasibility of using CS-TE with a classifier (rather than a regressor) to test G-causality;

II. The Cauchy-Schwarz Divergence

The CS divergence, originating from the signal processing community in the 2000s [17], [20], has shown favorable per-

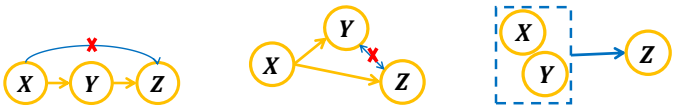

Fig. 1: (a) indirect causality from x to z. (b) x is a confounder to y and z; (c) synergistic effect (x and y produce a causal effect to z greater than the sum of their individual effects).

formance in traditional signal processing applications, including clustering [18] and independent component analysis [21].

Motivated by the renowned Cauchy-Schwarz inequality for square-integrable functions:

$$\left(\int p(x)q(x)dy\right)^2 \leq \int p(x)^2dx \int q(x)^2dx, \tag{3}$$

with equality iff $p(x)$ and $q(x)$ are linearly dependent, the CS divergence defines the distance between $p(x)$ and $q(x)$ by quantifying the gap between the left-hand side and right-hand side of Eq. (3) in terms of the logarithm of a ratio:

$$D_{\text{CS}}(p;q) = -\log\left(\frac{(\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x})^2}{\int p(\mathbf{x})^2d\mathbf{x}\int q(\mathbf{x})^2d\mathbf{x}}\right). \tag{4}$$

Due to the favorable properties of the CS divergence, such as its closed-form expression for mixture-of-Gaussians [22], it has recently garnered renewed attention in deep learning applications. Recent examples include deep multi-view learning [23], intrinsically-motivated reinforcement learning [19], and unsupervised domain adaptation [24], among others.

III. Proposed Method

A. Transfer entropy defined with CS divergence

Following the paradigm of GC, we wish to know if Y and X_{-1} are independent given Y_{-1}, where Y_{-1}, X_{-1}, and Y are random variables. The observations $(y_{t-}, x_{t-}, y_{t+1}), t = 1, 2, \cdots, N$, are identically but not necessarily independently distributed. To test whether $p(Y|Y_{-1}) = p(Y|Y_{-1}, X_{-1})$ holds, we can rewrite the condition as $p(Y_{-1}, Y)/p(Y_{-1}) = p(Y_{-1}, X_{-1}, y)/p(Y_{-1}, X_{-1})$, which is equivalent to:

$$p(X_{-1}, Y_{-1}, Y)p(Y_{-1}) = p(X_{-1}, Y_{-1})p(Y_{-1}, Y), \tag{5}$$

for each vector $(y_{t-1}, x_{t-1}, y_{t+1})$ in the support (Y_{-1}, X_{-1}, Y). By applying the CS divergence on both sides of Eq. (5), we obtain the Cauchy-Schwarz divergence transfer entropy (CS-TE), denoted as $\mathcal{T}_{\text{CS}}$:

$$\begin{aligned}\mathcal{T}_{\text{CS}}(x \to y) &= D_{\text{CS}}(p(X_{-1}, Y, Y_{-1})p(Y_{-1}); p(X_{-1}, Y_{-1})p(Y, Y_{-1})) \\ &= -2\log\left(\int p(X_{-1}, Y, Y_{-1})p(Y_{-1})p(X_{-1}, Y_{-1})p(Y, Y_{-1})\right) \\ &+ \log\left(\int p^2(X_{-1}, Y, Y_{-1})p^2(Y_{-1})\int p^2(X_{-1}, Y_{-1})p^2(Y, Y_{-1})\right).\end{aligned} \tag{6}$$

Proposition 1. *Given N observations $\{\mathbf{x}_{t-}, y_{t+1}, \mathbf{y}_{t-}\}_{t=1}^N$ drawing from an unknown and fixed joint distribution $p(X_{-1}, Y, Y_{-1})$ in which $\mathbf{x}_{t-} \in \mathbb{R}^m$, $y_{t+1} \in \mathbb{R}$, and $\mathbf{y}_{t-} \in \mathbb{R}^n$ refer to, respectively, the past observation of x, the future observation of y and the past observation of y at time index t. Let $K \in \mathbb{R}^{N\times N}$ be the Gram (a.k.a., kernel) matrix for variable X_{-1}, that is, $K_{ji} = \exp\left(-\frac{\|x_{j-1}-x_{i-1}\|_2^2}{2\sigma^2}\right)$, in which σ is the kernel width. Likewise, let $L \in \mathbb{R}^{N\times N}$ and $M \in \mathbb{R}^{N\times N}$ be the Gram matrices for variables Y and Y_{-1}, respectively. The empirical estimator of Eq. (6) is given by:*

$$\begin{aligned}&\widehat{D}_{CS}((p(X_{-1}, Y, Y_{-1})p(Y_{-1}); p(X_{-1}, Y_{-1})p(Y, Y_{-1})) = \\ &-2\log\left(\sum_{j=1}^N\left(\left(\sum_{i=1}^N M_{ji}\right)\left(\sum_{i=1}^N K_{ji}M_{ji}\right)\left(\sum_{i=1}^N L_{ji}M_{ji}\right)\right)\right) \\ &+\log\left(\sum_{j=1}^N\left(\left(\sum_{i=1}^N K_{ji}L_{ji}M_{ji}\right)\left(\sum_{i=1}^N M_{ji}\right)^2\right)\right) \\ &+\log\left(\sum_{j=1}^N\left(\frac{\left(\sum_{i=1}^N K_{ji}M_{ji}\right)^2\left(\sum_{i=1}^N L_{ji}M_{ji}\right)^2}{\left(\sum_{i=1}^N K_{ji}L_{ji}M_{ji}\right)}\right)\right).\end{aligned} \tag{7}$$

Proof. See the supplementary material. □

Remark 1. *We note that the authors of [19] also introduce a formulation of TE using CS divergence (see their Proposition 4). However, there are two major differences: (1) our formulation applies the original CS divergence to both sides of Eq. (5), whereas [19] defines CS-TE based on* conditional *CS (CCS) divergence, which involves double integrals. As a result, the estimators are fundamentally different. (2) The authors of [19] do not discuss multivariate extensions of CS-TE, nor do they explore its application in a supervised learning framework.*

To determine the significance of $\mathcal{T}_{\text{CS}}(x \to y)$, we follow [25], [26] by directly extracting segments of length S from the original data to generate P groups of surrogate data:

$$x_{\text{surr}} = [x_i, x_{i+1}, \cdots, x_{i+S-1}], \quad y_{\text{surr}} = [y_j, y_{j+1}, \cdots, y_{j+S-1}], \tag{8}$$

such that $|i - j|$ is sufficiently large. This way, the surrogate data sequences preserve the same mean, variance, and autocorrelation function as the original data, but the cross-correlations are disrupted. We compare $\mathcal{T}_{\text{CS}}(x \to y)$ with $\mathcal{T}_{\text{CS}}(x_{\text{surr}} \to y_{\text{surr}})$ from the P permutations to compute the p-value. Details are provided in the supplementary material.

B. Extension to multivariate time series

1) Cauchy-Schwarz divergence conditional transfer entropy (CS-CTE): To distinguish direct causation from indirect or common causal relationships, we propose CS-CTE, denoted by $\mathcal{C}_{\text{CS}}$. Given three time series $\{x_t\}$, $\{y_t\}$, and $\{z_t\}$, the central idea behind $\mathcal{C}_{\text{CS}}(x \to y|z)$ is to quantify the direct flow of information from x to y conditioned on z, by evaluating the CS divergence between the distributions $p(Y|Y_{-1}, Z_{-1})$ and $p(Y|Y_{-1}, Z_{-1}, X_{-1})$, which further reduces to the divergence between the joint distributions $p(X_{-1}, Y_{-1}, Z_{-1}, Y)p(Y_{-1}, Z_{-1})$ and $p(X_{-1}, Y_{-1}, Z_{-1})p(Y_{-1}, Z_{-1}, Y)$ by the Bayes's Theorem:

$$\mathcal{C}_{\text{CS}}(x \to y|z) = D_{\text{CS}}(p(X, Y, Z); q(X, Y, Z)), \tag{9}$$

where

$$p(X,Y,Z) = p(X_{-1},Y_{-1},Z_{-1},Y)p(Y_{-1},Z_{-1}),$$
$$q(X,Y,Z) = p(X_{-1},Y_{-1},Z_{-1})p(Y_{-1},Z_{-1},Y)).$$

Comparing with Eq. (9), we observe that each term inside Eq. (5) has an additional variable Z_{-1}. However, the estimator in Eq. (7) still applies. This is because we can concatenate variables Z_{-1} and Y_{-1} (i.e., the past of y_t includes also the past of z_t) and re-evaluate the corresponding Gram matrix M:

$$M_{ji} = \exp\left(-\frac{\|(y_{j-};z_{j-}) - (y_{i-};z_{i-})\|_2^2}{2\sigma^2}\right). \tag{10}$$

2) Cauchy-Schwarz divergence joint transfer entropy (CS-JTE): A synergy effect refers to the combined effect of two or more substances (in this context, time series), which exceeds the sum of their individual effects. To measure the joint effect from $\{x_t, y_t\}$ to z_t, we propose CS-JTE, denoted as $\mathcal{J}_{\text{CS}}$, to quantify the divergence between $p(Z|Z_{-1})$ and $p(Z|Z_{-1},X_{-1},Y_{-1})$, or equivalently between the joint distributions $p(Z,Z_{-1},Y_{-1},X_{-1})p(Z_{-1})$ and $p(Z,Z_{-1})p(Z_{-1},Y_{-1},X_{-1})$:

$$\mathcal{J}_{\text{CS}}(x,y\to z) = D_{\text{CS}}(\tilde{p}(X,Y,Z);\tilde{q}(X,Y,Z)), \tag{11}$$

where

$$\tilde{p}(X,Y,Z) = p(X_{-1},Y_{-1},Z_{-1},Z)p(Z_{-1}),$$
$$\tilde{q}(X,Y,Z) = p(Z_{-1},Z)p(X_{-1},Y_{-1},Z_{-1})).$$

To estimate Eq. (11), one can concatenate X_{-1} and Y_{-1}.

Remark 2. *Both $\mathcal{C}_{CS}(x\to y|z)$ and $\mathcal{J}_{CS}(x,y\to z)$ can be simply extended to more than three time series. If the causality between x and y is indirect, taking one additional series or more than one additional series in the causality chain as the conditional one(s) will not make the results any different [6].*

C. Detect causality in a supervised way

Traditional regression-based Granger causality tests rely heavily on selecting and constructing an appropriate regression model. On the other hand, TE and other information-theoretic measures require carefully designed methods (e.g., the permutation test with surrogate data, as mentioned earlier) to infer the distribution of the test statistic under the null hypothesis, which can also be challenging. Motivated by [27], [28], we explore the use of CS-TE to infer causality by employing a classifier to directly determine causal directions without the need for a permutation test.

To this end, we need to construct a training set that consists of M pairs of bivariate time series $S^1, S^2, \cdots, S^M$, where each time series $S^j (j = 1,2,\cdots,M)$ contains T_j observations $\{(x_1^j, y_1^j), (x_2^j, y_2^j), \cdots, (x_{T_j}^j, y_{T_j}^j)\}$ and a corresponding label $l^j \in \{1,-1,0\}$ which implies the ground-truth causal relationships $X \to Y$, $X \leftarrow Y$, or *No Causation*.

We extract a 2d feature $d_j = [\mathcal{T}_{\text{CS}}(x\to y); \mathcal{T}_{\text{CS}}(y\to x)]$ for each S^j. Meanwhile, to enhance feature expressiveness, we follow [28] and use the random Fourier features [29] to project d_j into a D-dimensional space ($D \gg d$) with a randomized map $z : \mathbb{R}^d \mapsto \mathbb{R}^D$:

$$z(d_j) = \sqrt{\frac{2}{D}}\left[\cos(\omega_1^T d_j + b_1), \ldots, \cos(\omega_D^T d_j + b_D)\right], \tag{12}$$

where $\omega_1, \omega_2, \ldots, \omega_D$ are sampled from a Gaussian distribution $\mathcal{N}_D(\mathbf{0}, \mathbf{I})$, and b_i is a bias term randomly sampled from a uniform distribution $U(0, 2\pi)$. We use the training set $\{z(d_j), l^j\}_{j=1}^M$ to construct a classifier and apply it to directly infer the causal relationship of a new pair of test time series.

IV. Experiments

A. The usefulness of CS-CTE and CS-JTE

We first demonstrate the effectiveness of CS-CTE and CS-JTE in modeling complex causal relationships. For all experiments, detailed setups, including hyperparameter tuning, are discussed in the supplementary material.

1) Hénon chaotic maps: We first consider 5 coupled Hénon chaotic maps [30], [31], whose ground truth causal relationship is $x_{i-1} \to x_i$ $(i = 2, \cdots, 5)$:

$$\begin{cases} x_{1,t} = 1.4 - x_{1,t-1}^2 + 0.3x_{1,t-2} \\ x_{i,t} = 1.4 - (Cx_{i-1,t-1} + (1-C)x_{i,t-1})^2 + 0.3x_{i,t-2}. \end{cases} \tag{13}$$

In our simulation, we set the coupling strength $C = 0.3$ and generate $1,024$ samples in 10 independent realizations respectively. As can be seen in Fig. 2, the linear GC fails for nonlinear data, whereas our CS-TE, the kernel Granger causality (KGC) [7], the TE, and the CCS [19] can effectively detect most of pairwise causal directions. Additionally, our CS-CTE successfully removes indirect causations. However, our measure may not be as powerful as TE or CCS in identifying causality specifically from x_1 to x_2.

2) Synergistic causal model: We next consider a synergistic causal model defined as [32], [33]:

$$x_{4,t} = 0.1(x_{1,t-1} + x_{2,t-1}) + \rho x_{2,t-1}x_{3,t-1} + 0.1\epsilon_t, \tag{14}$$

in which $\rho \geq 0$ is the strength of coupling between x_2 and x_3, $x_1 \sim \mathcal{N}(0,1)$, $x_2 \sim \mathcal{N}(0,1)$, $x_3 \sim \mathcal{N}(0,1)$, ϵ is an additive Gaussian noise. The true causal relationship is that x_2 and x_3 synergistically influence x_4, where x_1 is an additional cause.

Our results in Fig. 3 shows that all CS-TE values increase as ρ increases. Moreover, we consistently observe that:

$$\mathcal{J}_{\text{CS}}(x_2, x_3 \to x_4) > \mathcal{T}_{\text{CS}}(x_2 \to x_4) + \mathcal{T}_{\text{CS}}(x_3 \to x_4). \tag{15}$$

which implies the ability of CS-JTE to identify the synergistic effect. More interestingly, conditioning on x_2, we observe that the information flow from x_3 to x_4 increases significantly. That is, x_2 is a suppressor variable for x_3 with respect to the influence on x_4 [32].

B. The feasibility of classifier-based approach

We generate training data for supervised learning by utilizing a p-th order nonlinear vector autoregressive (NVAR) model to synthesize $7,500$ time series pairs of length 256, each

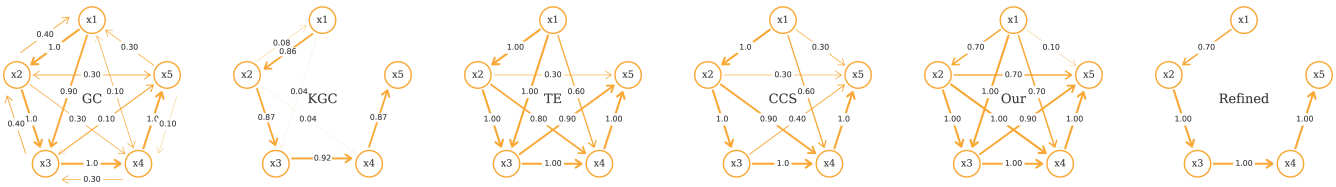

Fig. 2: The causal graph identified by linear Granger causality (GC), kernel Granger causality (KGC), transfer entropy (TE) with kNN estimator, the conditional Cauchy-Schwarz divergence (CCS), and our CS-TE w/wo refinement of CS-CTE. The numerical value on top of each line represents the percentage of successful detections across 10 independent runs.

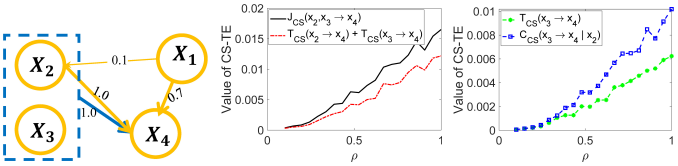

Fig. 3: Left: reconstructed causal graph; Right: values of TE.

TABLE I: Test accuracy for 300 pairs of nonlinear time series

GC	KGC	TE	CCS	CS-TE (permutation)	CS-TE (classifier)
0.80	0.93	0.89	0.65	0.97	0.96

with one of the following true causal relationships: $X \to Y$, $X \leftarrow Y$, and *No Causation*. For example, the following model is used to generate time series with $X \to Y$ [28]:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \frac{1}{p}\sum_{\tau=1}^{p} \begin{bmatrix} a_\tau & 0 \\ c_\tau & d_\tau \end{bmatrix} \begin{bmatrix} \sigma(x_{t-\tau}) \\ y_{t-\tau} \end{bmatrix} + \begin{bmatrix} N_x \\ N_y \end{bmatrix} \tag{16}$$

where σ is a sigmoid function, $p \in \{1, 2\}$, a_τ, d_τ are drawn from the uniform distribution $\mathcal{U}(-1, 1)$, and $c_\tau \in \{-1, 1\}$. N_x and N_y are Gaussian noises with distribution $\mathcal{N}(0, 1)$. Time series with $X \leftarrow Y$, and *No Causation* are generated similarly.

We use the training data to train a random forest classifier with $2,500$ decision trees, each with a maximum depth of 15. The model's performance is tested on 300 pairs of synthetic test data, generated using a different data function but with the same length as the training data. For example, the nonlinear time series test data with $X \to Y$ is generated by:

$$\begin{aligned} x_t &= 0.5x_{t-1} + 0.9N_x \\ y_t &= 1.5\exp\left(-(x_{t-1} + x_{t-2})\right) + 0.7\cos\left(y_{t-1}^2\right) + 0.2N_y, \end{aligned} \tag{17}$$

where N_x and N_y are Gaussian noises with distribution $\mathcal{N}(0, s)$ and $s \in \{0.5, 1.0, 1.5, 2.0\}$. Similarly, we generated nonlinear time series with $X \leftarrow Y$. For time series with *No Causation*, we simply omitted the exponential term in Eq. (17). From Table I, CS-TE achieves the highest accuracy with both the permutation-based and classifier-based approaches. The relatively poorer performance of CCS [19] suggests that TE is better modeled using the classical CS divergence rather than the conditional CS divergence (see also Remark 1).

C. Finaical data

Following [10], we use preprocessed daily closing price data[1] for stock indices from 14 markets via Yahoo Finance, covering the period from January 1, 2016, to December 31, 2019. The indices include the Shanghai Composite Index, Shenzhen Component Index, MOEX Russia Index, S&P/ASX 200, Dow Jones Industrial Average, S&P 500, Hang Seng Index, NASDAQ Composite, KOSPI Index, Euronext 100 Index, Nikkei 225, NZX 50 Index, Straits Times Index, and Euro Stoxx 50.

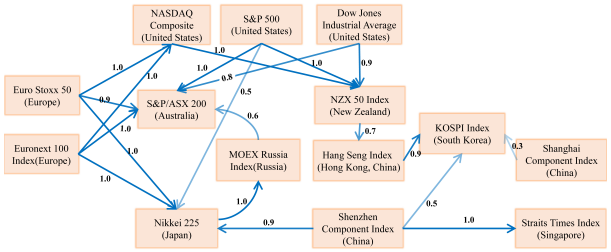

Fig. 4: Causal network of stock indices. The numerical value on top of each line represents the percentage of successful detections across 10 independent runs.

We construct the causal network with CS-TE, and refine it with CS-CTE. Our findings in Fig. 4 align with [14], indicating that the U.S. and European markets are more influential to price fluctuations in the Asia-Pacific sectors (e.g., Australia and Japan). The Chinese stock markets (especially Shanghai) were less affected by external influences [34], likely due to China's stock market policies [35]. The trained classifier verifies the causal relationships and directions in Fig. 4, resulting in an 81% overlap. The networks generated by other methods are presented in the supplementary material. Among these, our results are much more interpretable.

V. Conclusions and Future Work

We developed Cauchy-Schwarz divergence transfer entropy (CS-TE) for time series causal inference and demonstrated its effectiveness in multivariate scenarios and under a supervised learning framework. We plan to explore the possible equivalence between CS-TE and G-causality for Gaussian variables.

[1] https://github.com/cloudy-sfu/GC-significance-test.

REFERENCES

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

[2] R. Marschinski and H. Kantz, "Analysing the information flow between financial time series: An improved estimator for transfer entropy," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 30, pp. 275–281, 2002.

[3] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger causality analysis in neuroscience and neuroimaging," *Journal of Neuroscience*, vol. 35, no. 8, pp. 3293–3297, 2015.

[4] A. Shojaie and E. B. Fox, "Granger causality: A review and recent advances," *Annual Review of Statistics and Its Application*, vol. 9, no. 1, pp. 289–319, 2022.

[5] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*. Springer, 2006, pp. 366–381.

[6] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, "Analyzing multiple nonlinear time series with extended granger causality," *Physics letters A*, vol. 324, no. 1, pp. 26–35, 2004.

[7] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear granger causality," *Physical review letters*, vol. 100, no. 14, p. 144103, 2008.

[8] A. Wismüller, A. M. Dsouza, M. A. Vosoughi, and A. Abidin, "Large-scale nonlinear granger causality for inferring directed dependence from short multivariate time-series data," *Scientific reports*, vol. 11, no. 1, p. 7817, 2021.

[9] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. B. Fox, "Neural granger causality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4267–4279, 2021.

[10] Y. Lu, Y. Lee, H. Feng, J. Leung, A. Cheung, K. Dost, K. Taskova, and T. Lacombe, "Interpretability meets generalizability: A hybrid machine learning system to identify nonlinear granger causality in global stock indices," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2023, pp. 322–334.

[11] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[12] V. A. Vakorin, O. A. Krakovska, and A. R. McIntosh, "Confounding effects of indirect connections on causality estimation," *Journal of neuroscience methods*, vol. 184, no. 1, pp. 152–160, 2009.

[13] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for gaussian variables," *Physical review letters*, vol. 103, no. 23, p. 238701, 2009.

[14] P. Jizba, H. Kleinert, and M. Shefaat, "Rényi's information transfer between financial time series," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 10, pp. 2971–2989, 2012.

[15] A. Rényi, "On measures of entropy and information," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, vol. 4. University of California Press, 1961, pp. 547–562.

[16] S. Sarbu, "Rényi information transfer: Partial rényi transfer entropy and partial rényi mutual information," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5666–5670.

[17] J. C. Principe, D. Xu, Q. Zhao, and J. W. Fisher, "Learning from examples with information theoretic criteria," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 26, pp. 61–77, 2000.

[18] R. Jenssen, J. C. Principe, D. Erdogmus, and T. Eltoft, "The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels," *Journal of the Franklin Institute*, vol. 343, no. 6, pp. 614–629, 2006.

[19] S. Yu, H. Li, S. Løkse, R. Jenssen, and J. C. Príncipe, "The conditional cauchy-schwarz divergence with applications to time-series data and sequential decision making," *arXiv preprint arXiv:2301.08970*, 2023.

[20] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[21] D. Xu, J. C. Principe, J. Fisher, and H.-C. Wu, "A novel measure for independent component analysis (ica)," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 2. IEEE, 1998, pp. 1161–1164.

[22] K. Kampa, E. Hasanbelliu, and J. C. Principe, "Closed-form cauchy-schwarz pdf divergence for mixture of gaussians," in *The 2011 International Joint Conference on Neural Networks*. IEEE, 2011, pp. 2578–2585.

[23] D. J. Trosten, S. Lokse, R. Jenssen, and M. Kampffmeyer, "Reconsidering representation alignment for multi-view clustering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1255–1265.

[24] W. Yin, S. Yu, Y. Lin, J. Liu, J.-J. Sonke, and S. Gavves, "Domain adaptation with cauchy-schwarz divergence," in *The 40th Conference on Uncertainty in Artificial Intelligence*.

[25] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1-4, pp. 77–94, 1992.

[26] P. Duan, F. Yang, S. L. Shah, and T. Chen, "Transfer zero-entropy and its application for capturing cause and effect relationship between variables," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 3, pp. 855–867, 2014.

[27] D. Lopez-Paz, K. Muandet, and B. Recht, "The randomized causation coefficient." *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 2901–2907, 2015.

[28] Y. Chikahara and A. Fujino, "Causal inference in time series via supervised learning." in *IJCAI*, 2018, pp. 2042–2048.

[29] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Advances in neural information processing systems*, vol. 20, 2007.

[30] D. Kugiumtzis, "Direct-coupling information measure from nonuniform embedding," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 87, no. 6, p. 062918, 2013.

[31] H. Li, S. Yu, and J. Principe, "Causal recurrent variational autoencoder for medical time series generation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 7, 2023, pp. 8562–8570.

[32] S. Stramaglia, J. M. Cortes, and D. Marinazzo, "Synergy and redundancy in the granger causal analysis of dynamical networks," *New Journal of Physics*, vol. 16, no. 10, p. 105003, 2014.

[33] W. Zhou, S. Yu, and B. Chen, "Causality detection with matrix-based transfer entropy," *Information Sciences*, vol. 613, pp. 357–375, 2022.

[34] Y. Tang, J. J. Xiong, Y. Luo, and Y.-C. Zhang, "How do the global stock markets influence one another? evidence from finance big data and granger causality directed network," *International Journal of Electronic Commerce*, vol. 23, no. 1, pp. 85–109, 2019.

[35] Q. Zheng and L. Song, "Dynamic contagion of systemic risks on global main equity markets based on granger causality networks," *Discrete Dynamics in Nature and Society*, vol. 2018, no. 1, p. 9461870, 2018.

Supplementary Material

Zhaozhao Ma[*,†]
[*] *Zhejiang University*
[†] *Georgia Institute of Technology*
zhaozhaoma@gatech.edu

Shujian Yu[‡,§]
[‡] *Vrije Universiteit Amsterdam*
[§] *UiT - The Arctic University of Norway*
s.yu3@vu.nl

Abstract

In the supplementary material accompanying our paper *Cauchy-Schwarz Divergence Transfer Entropy*, we provide comprehensive support for our proposed methodology through rigorous derivations, comparative analyses, and detailed explanations. First, we present a detailed and rigorous derivation of the empirical estimator for our novel formulation of Transfer Entropy (TE) based on the Cauchy-Schwarz (CS) divergence. Second, We also present the causal networks constructed using four methods—linear Granger causality (GC), transfer entropy (TE) with a k-nearest neighbors (kNN) estimator, conditional Cauchy-Schwarz divergence (CCS), and kernel Granger causality (KGC)—demonstrating that our method offers significantly greater interpretability compared to the others. Third, we provide a detailed explanation of the nonlinear data generation strategy and the specific selection of kernel size used to test the classifier-based approach employing CS-TE. Furthermore, we use the trained classifier to test the causal relationships in our proposed method's causal network, achieving a coincidence rate of 81%. Finally, we provide a detailed description of the permutation test.

1 The Motivation of CS Divergence

1.1 Motivation in terms of definition

Firstly, although both the Kullback-Leibler (KL) divergence and the CS divergence can be employed to measure the difference or similarity between two entities (such as probability distributions or vectors), the CS divergence is considerably more stable than the KL divergence in that it relaxes the constraints on the supports of the distributions [3]. For any two densities p and q, $D_{\mathrm{KL}}(p;q)$ has finite values only if $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$ (note that, $p(x)\log\left(\frac{p(x)}{0}\right) \to \infty$); whereas $D_{\mathrm{KL}}(q;p)$ has finite values only if $\mathrm{supp}(q) \subseteq \mathrm{supp}(p)$. In contrast, $D_{\mathrm{CS}}(p;q)$ is symmetric and always yields finite values unless the supports of p and q have no overlap, i.e., $\mathrm{supp}(p) \cap \mathrm{supp}(q) = \emptyset$. Please see Fig. 1 for an illustration.

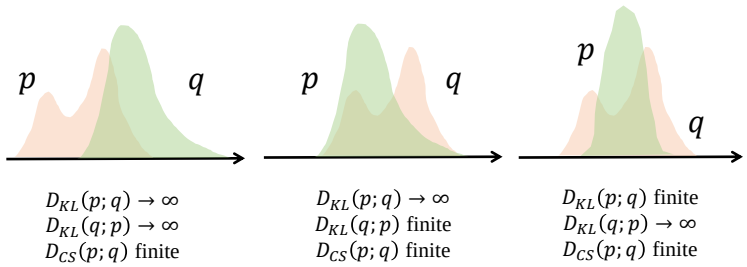

Figure 1: KL divergence is infinite even though there is an overlap between $\mathrm{supp}(p)$ and $\mathrm{supp}(q)$, but neither is a subset of the other. CS divergence does not have such constraint.

Second, both the CS divergence ($\alpha = 2$) and the KL divergence ($\alpha = 1$) can be viewed as special cases of the generalized Rényi's divergence, defined by Lutwark et al.[1]:

$$D_\alpha(p;q) = \log\left(\frac{\left(\int q(x)^{\alpha-1}p(x)\right)^{\frac{1}{1-\alpha}}\left(\int q(x)^\alpha\,dx\right)^{\frac{\alpha}{1-\alpha}}}{\left(\int p(x)^\alpha\,dx\right)^{\frac{1}{\alpha(1-\alpha)}}}\right). \tag{1}$$

One should note that varying values of α emphasize different aspects of the underlying data distribution (e.g., the mode, the tails, etc.)[2]. From this perspective, the CS divergence-based

TE offers complementary insights to the KL divergence-based TE. Specifically, there are scenarios where causality can be better detected using alternative values of α rather than restricting to 1 (i.e., the KL divergence).

1.2 Motivation in terms of estimation

The KL divergence is notoriously difficult to estimate in practice. Consequently, most existing studies that apply KL divergence-based TE to biomedical or financial signals resort to discretizing the data before computing the discrete KL divergence. However, discretization often leads to a loss of information. Additionally, determining the appropriate bin size and number of bins for different types of signals is challenging.

In contrast, our paper develops closed-form estimators for both the CS divergence-based TE and its multivariate extensions. Our approach eliminates the need for discretization and provides an elegant and insightful closed-form expression.

1.3 Motivation in terms of extension

Finally, we emphasize that our study extends beyond the mere substitution of KL divergence with CS divergence. Additionally, we explore the generalization of TE to scenarios involving more than two variables. Our multivariate extensions, including joint TE and conditional TE, represent significant advancements in this field. Extending KL divergence-based TE to multivariate contexts is nontrivial and does not benefit from the availability of closed-form estimators.

2 Proofs

2.1 Definition

In our paper, we rigorously define the Cauchy-Schwarz divergence transfer entropy (CS-TE) for any arbitrary pair of time series $\{x_t\}$ and $\{y_t\}$, establishing a precise mathematical framework for quantifying causal relationships between them. We obtain the Cauchy-Schwarz divergence transfer entropy (CS-TE), denoted as $\mathcal{T}_{\mathrm{CS}}$:

$$\begin{aligned}\mathcal{T}_{\mathrm{CS}}(x \to y) &= D_{\mathrm{CS}}(p(X_{-1}, Y, Y_{-1})p(Y_{-1}); p(X_{-1}, Y_{-1})p(Y, Y_{-1})) \\ &= -2\log\left(\int p(X_{-1}, Y, Y_{-1})p(Y_{-1})p(X_{-1}, Y_{-1})p(Y, Y_{-1})\right) \\ &\quad + \log\left(\left(\int p^2(X_{-1}, Y, Y_{-1})p^2(Y_{-1})\right)\left(\int p^2(X_{-1}, Y_{-1})p^2(Y, Y_{-1})\right)\right).\end{aligned} \tag{2}$$

2.2 Estimation

For the first term in Eq.(2), we have:

$$\begin{aligned}&\int p(X_{-1}, Y, Y_{-1})p(X_{-1}, Y_{-1})p(Y, Y_{-1})\, dX_{-1}\, dY\, dY_{-1} \\ &\quad = \mathbb{E}_{p(X_{-1}, Y, Y_{-1})}\left(p(Y_{-1})p(X_{-1}, Y_{-1})p(Y, Y_{-1})\right).\end{aligned} \tag{3}$$

Given N observations $\{\mathbf{x}_{t-}, y_{t+1}, \mathbf{y}_{t-}\}_{t=1}^{N}$ drawing from an unknown and fixed joint distribution $p(X_{-1}, Y, Y_{-1})$ in which $\mathbf{x}_{t-} \in \mathbb{R}^m$, $y_{t+1} \in \mathbb{R}$, and $\mathbf{y}_{t-} \in \mathbb{R}^n$ refer to, respectively, the past observation of x, the future observation of y and the past observation of y at time index t. Eq.(3) can be approximated using a Monte Carlo estimator:

$$\frac{1}{N}\sum_{t=1}^{N} p(y_{t-})p(x_{t-}, y_{t-})p(y_{t+1}, y_{t-}). \tag{4}$$

Further, by using Gaussian kernels for $p(x_{t-}, y_{t-}), p(y_{t+1}, y_{t-}), p(y_{t-})$, Eq.(4) can be expressed

as Eq.(5):

$$
\begin{aligned}
&\approx \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{N(\sqrt{2\pi}\sigma)^{d_{y_{t-}}}}\sum_{i=1}^{N}\exp\left(-\frac{\|y_{j-1}-y_{i-1}\|_2^2}{2\sigma^2}\right)\right)\\
&\cdot\left(\frac{1}{N(\sqrt{2\pi}\sigma)^{d_{x_{t-}}+d_{y_{t-}}}}\sum_{i=1}^{N}\exp\left(-\frac{\|x_{j-1}-x_{i-1}\|_2^2}{2\sigma^2}\right)\exp\left(-\frac{\|y_{j-1}-y_{i-1}\|_2^2}{2\sigma^2}\right)\right)\\
&\cdot\left(\frac{1}{N(\sqrt{2\pi}\sigma)^{d_{y_{t+1}}+d_{y_{t-}}}}\sum_{i=1}^{N}\exp\left(-\frac{\|y_{j+1}-y_{i+1}\|_2^2}{2\sigma^2}\right)\exp\left(-\frac{\|y_{j-1}-y_{i-1}\|_2^2}{2\sigma^2}\right)\right).
\end{aligned}
\tag{5}
$$

Where σ represents the bandwidth of the Gaussian kernel, and $d_{x_{t-}}$, $d_{y_{t-}}$, and $d_{y_{t+1}}$ denote the dimensions of x_{t-}, y_{t-}, and y_{t+1}, respectively, or more precisely, their embedding dimensions.$d_{x_{t-}} = m$, $d_{y_{t-}} = n$, $d_{y_{t+1}} = 1$.

Let $K \in \mathbb{R}^{N\times N}$ be the Gram (a.k.a., kernel) matrix for variable X_{-1}, $K_{ji} = \exp\left(-\frac{\|x_{j-1}-x_{i-1}\|_2^2}{2\sigma^2}\right)$. Likewise, let $L \in \mathbb{R}^{N\times N}$ and $M \in \mathbb{R}^{N\times N}$ be the Gram matrices for variables Y and Y_{-1}, respectively. We can obtain:

$$
\begin{aligned}
&\int p(X_{-1},Y,Y_{-1})p(X_{-1},Y_{-1})p(Y,Y_{-1})\,dX_{-1}\,dY\,dY_{-1}\\
&=\frac{1}{N^4(\sqrt{2\pi}\sigma)^{d_{x_{t-}}+d_{y_{t+1}}+3d_{y_{t-}}}}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}M_{ji}\right)\left(\sum_{i=1}^{N}K_{ji}M_{ji}\right)\left(\sum_{i=1}^{N}L_{ji}M_{ji}\right).
\end{aligned}
\tag{6}
$$

Similarly, For the second and third terms of Eq.(3), we can apply the same pattern to obtain Eq.(7) and Eq.(8).

$$
\begin{aligned}
\int p^2(X_{-1},Y,Y_{-1})p^2(Y_{-1})\,dX_{-1}\,dY\,dY_{-1} &= \mathbb{E}_{p(X_{-1},Y,Y_{-1})}\left(p(X_{-1},Y,Y_{-1})p^2(Y_{-1})\right)\\
&=\frac{1}{N^4(\sqrt{2\pi}\sigma)^{d_{x_{t-}}+d_{y_{t+1}}+3d_{y_{t-}}}}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}K_{ji}L_{ji}M_{ji}\right)\left(\sum_{i=1}^{N}M_{ji}\right)^2.
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
\int p^2(X_{-1},Y_{-1})p^2(Y,Y_{-1})\,dX_{-1}\,dY\,dY_{-1} &= \mathbb{E}_{p(X_{-1},Y,Y_{-1})}\left(\frac{p^2(X_{-1},Y_{-1})p^2(Y,Y_{-1})}{p(X_{-1},Y,Y_{-1})}\right)\\
&=\frac{1}{N^4(\sqrt{2\pi}\sigma)^{d_{x_{t-}}+d_{y_{t+1}}+3d_{y_{t-}}}}\sum_{j=1}^{N}\left(\frac{\left(\sum_{i=1}^{N}K_{ji}L_{ji}M_{ji}\right)^2\left(\sum_{i=1}^{N}L_{ji}M_{ji}\right)^2}{\left(\sum_{i=1}^{N}K_{ji}L_{ji}M_{ji}\right)}\right).
\end{aligned}
\tag{8}
$$

Finally, by combining Eq.(6), Eq.(7), and Eq.(8) and eliminating the normalization constant term, we obtain the empirical estimator for Eq.(2):

$$
\begin{aligned}
&\widehat{D}_{\mathrm{CS}}((p(X_{-1},Y,Y_{-1})p(Y_{-1});p(X_{-1},Y_{-1})p(Y,Y_{-1}))\\
&=-2\log\left(\sum_{j=1}^{N}\left(\left(\sum_{i=1}^{N}M_{ji}\right)\left(\sum_{i=1}^{N}K_{ji}M_{ji}\right)\left(\sum_{i=1}^{N}L_{ji}M_{ji}\right)\right)\right)\\
&\quad+\log\left(\sum_{j=1}^{N}\left(\left(\sum_{i=1}^{N}K_{ji}L_{ji}M_{ji}\right)\left(\sum_{i=1}^{N}M_{ji}\right)^2\right)\right)\\
&\quad+\log\left(\sum_{j=1}^{N}\left(\frac{\left(\sum_{i=1}^{N}K_{ji}M_{ji}\right)^2\left(\sum_{i=1}^{N}L_{ji}M_{ji}\right)^2}{\left(\sum_{i=1}^{N}K_{ji}L_{ji}M_{ji}\right)}\right)\right).
\end{aligned}
\tag{9}
$$

2.3 Conclusion

The core idea of this proof is to approximate the joint probability density function using Gaussian kernel density estimation and to derive the final model formula Eq.(9) by summing over the similarities between samples. This formula can be further extended to derive the CS divergence-based conditional transfer entropy and CS divergence-based joint transfer entropy, and its feasibility makes it applicable to classifiers.

3 Causal Network Base on Different Methods

3.1 Causal Network Base on GC

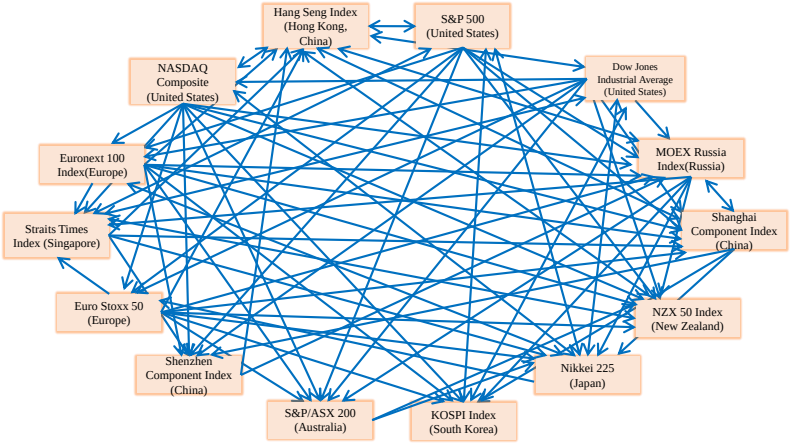

Figure 2: GC causal network

It is evident that GC is overly sensitive in causal relationship detection, leading to a large number of causal relationships between different stock indices in Fig. 2. Moreover, some phenomena that are clearly inconsistent with the patterns of the financial market have emerged, such as: The MOEX Russia Index exhibits causality on several indices, including the Shanghai Composite Index, S&P/ASX 200, Hang Seng Index KOSPI Index, Nikkei 225 Index, NZX 50 Index, Straits Times Index, and Euro Stoxx 50. Despite global market interconnections, the Russian market is relatively small with limited global influence, making it unusual for the Russian stock market to have direct causality effects on such a wide range of international indices, particularly those in the Asia-Pacific region.

3.2 Causal Network Base on TE

According to the results of the TE, in Fig. 3, the Straits Times Index exhibits strong causal relationships with the S&P/ASX 200, Hang Seng Index, Euronext 100 Index, NZX 50 Index, and Euro Stoxx 50. Although Singapore is a major financial hub, it is unusual for its index to exert such strong causal influence on European markets and other key indices. European indices like the Euro Stoxx 50 and Euronext 100 showed no causal relationships with any other indices. Given the prominent role of European markets in global finance, the absence of detectable causal relationships is unexpected. Likewise, it is surprising that Asian indices such as the Nikkei 225, KOSPI Index, Shanghai Composite Index, and Shenzhen Component Index displayed no causal relationships with other indices. These markets often react to global events and, due to overnight trading and global investor sentiment, typically influence other markets in turn. The U.S. indices (Dow Jones Industrial Average, S&P 500, and NASDAQ Composite) exhibited the same level of causal relationships with the same set of indices. While these indices are correlated, it is unusual for them to exert identical causal influence on the same indices unless they are capturing identical information.

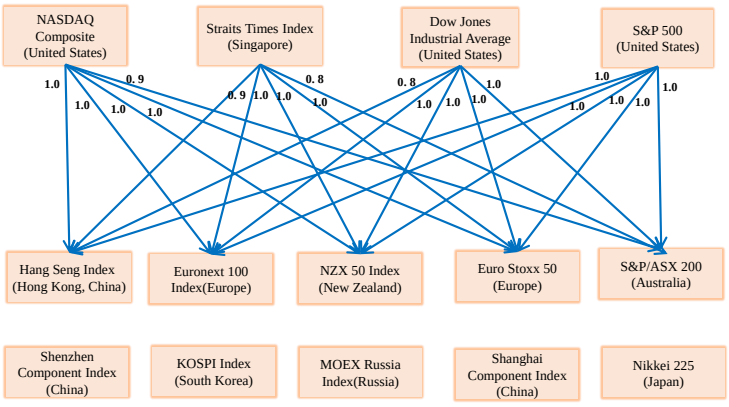

Figure 3: TE causal network

3.3 Causal Network Base on CCS

From the results of CCS in Fig. 4, the following issues are observed: The NZX 50 Index appears to influence nine major global markets, which is illogical considering the relative size of NZX. The NASDAQ Composite lacks causal relationships, and as a major global technology index, it should be expected to influence other markets. The Euronext 100 Index and Euro Stoxx 50 Index, which are significant European markets, do not show any causal relationships with other indices. The Straits Times Index has a causality value of 0.7 with the NASDAQ Composite and exhibits strong causal relationships with other indices. The Dow Jones Industrial Average and S&P 500 show high causality values with the Shenzhen Component Index, given China's capital controls and limited direct exposure to U.S. markets, such strong causality may be overstated.

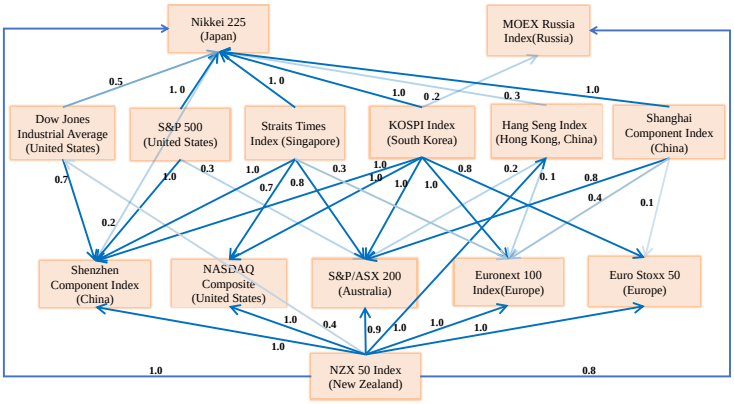

Figure 4: CCS causal network

3.4 Causal Network Base on KGC

Due to the characteristics of the KGC method, numerical causality exists between any pair of stock indices determined by this method. For the sake of readability, a matrix heatmap is used, as shown in Fig. 5. In the figure, the value of each cell represents the causal strength from the row index to the column index.

Based on the results of KGC in Fig. 5, the following prominent issues are observed: The universally high causality values are problematic, as it is unlikely that all global stock indices would exhibit strong causal influences on one another simultaneously. Smaller markets, such as the NZX 50 Index, show high values with major indices like the Dow Jones Industrial Average. Symmetrical causality between indices: Observation many indices exhibit high values in both directions, suggesting mutual causality. While some bidirectional influences are possible, widespread symmetrical causality indicates potential methodological issues.

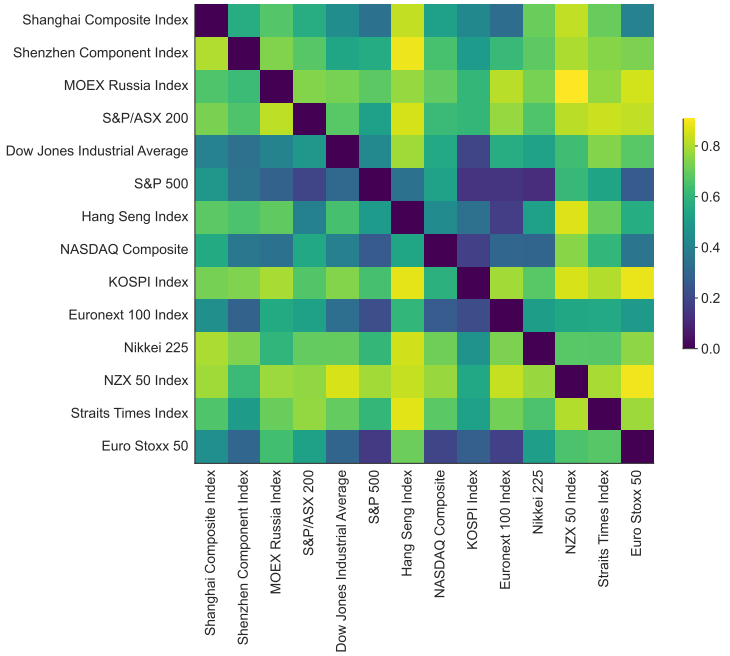

Figure 5: KGC causal network

3.5 Causal Network Base on Our Proposed Method

In Fig. 6, the causal graph derived from our proposed method highlights the influence of major U.S. indices on global markets. As the global financial center, fluctuations in the U.S. markets often impact other markets worldwide. While U.S. markets close after the Asia-Pacific markets, news and economic data released during U.S. trading hours can affect investor sentiment in the Asia-Pacific region, which is then reflected in the opening prices of their markets the following day.

The Shenzhen Component Index influences the KOSPI Index, Nikkei 225, and Straits Times Index, reflecting China's strong economic ties with neighboring Asian countries. Due to interconnected trade and supply chains, Asian markets frequently respond to changes in China's economy.

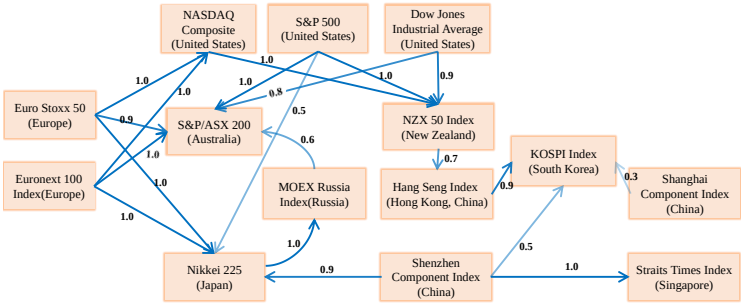

Figure 6: Out proposed method causal network

The Hang Seng Index shows a causal relationship with the KOSPI Index, with a value of 0.9. This is likely due to the significant trade relations and investment flows between Hong Kong and South Korea. As a major financial hub, fluctuations in the Hang Seng Index can influence investor sentiment in regional markets. The NASDAQ Composite displays causal relationships with both the KOSPI Index and the NZX 50 Index. The NASDAQ, which focuses on technology companies, is closely followed by Asian markets like South Korea, where companies such as Samsung play a key role in the tech industry. Investors may adjust their global portfolios based on the NASDAQ's performance.

The Euronext 100 Index significantly influences the S&P/ASX 200, NASDAQ Composite, and Nikkei 225 indices. European markets close before the U.S. markets and may affect indices like the NASDAQ. Furthermore, fluctuations in European markets often reflect global economic trends, influencing investor sentiment in other regions. The Nikkei 225 shows a causal relationship with the MOEX Russia Index, with a value of 1. As Japan is a major importer of energy commodities, changes in its economic outlook can impact commodity prices, which, in turn, affect the Russian market. Although this influence may be indirect, it is reasonable within certain timeframes.

The MOEX Russia Index shows a causal relationship with the S&P/ASX 200, with a value of 0.6. Both Russia and Australia are major commodity exporters, and fluctuations in the Russian market can affect global commodity prices, subsequently impacting the Australian market. The NZX 50 Index exhibits a causal relationship with the Hang Seng Index, with a value of 0.7. Despite the smaller size of the New Zealand market, its economic indicators can provide insights into the Asia-Pacific region. As the New Zealand market opens earlier, its movements may influence investor sentiment before the Hong Kong market opens.

3.6 Similarity

Notably, despite the significant differences between the causal network derived from our proposed method (OUR) and the results from the four methods (CCS, GC, KGC, and TE) at certain levels, some causal relationships exhibit interpretable similarities upon comparison.

For instance, the causal relationship from the Dow Jones Industrial Average to the S&P/ASX 200 is supported by four methods: OUR (0.8), GC (1), KGC (0.484), and TE (1). Similarly, the relationship from the S&P 500 to the S&P/ASX 200 is supported by five methods: OUR (1), GC (1), KGC (0.185), TE (1), and CCS (0.3). This reflects Australia's role as a resource-based economy with significant trade and financial ties to the United States; therefore, fluctuations in the Dow Jones Industrial Average transmit to the Australian market.

The causal relationship from the Dow Jones Industrial Average to the NZX 50 Index is supported by four methods: OUR (0.9), GC (1), KGC (0.628), and TE (1). The relationship from the S&P 500 to the NZX 50 Index is supported by five methods: OUR (1), GC (1), KGC (0.605), TE (1), and CCS (1). Additionally, the relationship from the NASDAQ Composite to the NZX 50 Index is validated by five methods: OUR (1), GC (1), KGC (0.745), TE (1), and CCS (1). These

three causal relationships, confirmed by five methods, strongly indicate that New Zealand's smaller economic scale makes its financial market more sensitive to the dynamics of major global economies, especially the influences from the United States through foreign exchange markets, trade channels, and investor expectations.

The causal relationship from the NZX 50 Index to the Hang Seng Index is supported by four methods: OUR (0.7), GC (1), KGC (0.827), and CCS (1). This may reflect the regional interconnectedness of Asia-Pacific markets, where investors might adjust their expectations for the Hong Kong market based on the performance of the New Zealand market at specific times.

3.7 Conclusion

Compared to the above four methods, the findings of our proposed method are more consistent with economic reality. It demonstrates more selective causal relationships, with values that align better with actual conditions, and focuses on markets with economic connections. This method more accurately reflects the actual market dynamics, as causal relationships are predominantly unidirectional and concentrated between indices that have reasonable economic links.

4 Classifier Construction and Causal Graph Validation

4.1 Nonlinear Training Data

In our paper, we used a nonlinear vector autoregressive (NVAR) model to synthesize 7,500 time series pairs of length 256, each labeled with a causal label: $X \to Y$, $X \leftarrow Y$, or *No Causation*.

4.2 Nonlinear Test Data

We use the training data to train a random forest classifier. Furthermore, we employed different data generation functions to generate 300 pairs data with the same length of 256 to test the classifier:

- For the causal direction is $x \to y$:

$$x_t = 0.5x_{t-1} + 0.9N_x \tag{10}$$

$$y_t = 1.5\exp\left(-(x_{t-1} + x_{t-2})\right) + 0.7\cos\left(y_{t-1}^2\right) + 0.2N_y \tag{11}$$

- For the causal direction is $x \leftarrow y$:

$$y_t = 1.2y_{t-1} + 0.3N_y \tag{12}$$

$$x_t = -1.5\exp(-(y_{t-1} + y_{t-2})^2) + 0.7\cos(x_{t-1}^2) + 0.2N_x \tag{13}$$

- For *No Causation*:

$$x_t = 0.5x_{t-1} + 0.9N_x \tag{14}$$

$$y_t = 1.5\cos\left(y_{t-1}^2\right) + 2.5N_y \tag{15}$$

4.3 Gaussian Kernel Bandwidth

At the end of the proof, we obtain Eq.(9), in which the calculations of K_{ji}, L_{ji} and M_{ji} all require the inclusion of the Gaussian kernel bandwidth, denoted as σ. The parameter σ controls the smoothness of the kernel function and its sensitivity to the distances between data points. In Eq.(9), σ is used to compute the similarity between each pair of data points, thereby determining their weights in the probability density estimation. Consequently, the selection of σdirectly affects the model's sensitivity to data and the accuracy of the kernel density estimation. In practical applications, the choice of σ typically needs to be adjusted according to the characteristics of the data.

We concatenate Y_-, Y, and X_- into a matrix, where each row corresponds to a joint observation of Y_-, Y, and X_- at a specific time step. Subsequently, the pairwise squared Euclidean distances between all rows are computed. The upper triangular elements of the resulting distance matrix are extracted, and all non-zero distances are flattened into a vector R. The median of the non-zero

elements in R, denoted as D, is then calculated, representing the typical distance between samples. This median distance D provides an appropriate estimate for σ, ensuring that it is aligned with the scale of the data. To further refine σ, a scaling parameter ζ is applied to adjust D to an optimal range, maintaining an appropriate sensitivity of σ to the typical inter-sample distances.

The choice of ζ critically affects the behavior of the kernel function: a smaller ζ results in a smaller σ, making the kernel highly sensitive to small distance variations, whereas a larger ζ yields a larger σ, rendering the kernel overly smooth and less responsive to variations in inter-sample distances.

In the computation process of CS-TE based on nonlinear data, we select ζ=0.05 to enhance the sensitivity of CS-TE to small distance variations, thereby facilitating the classification task.

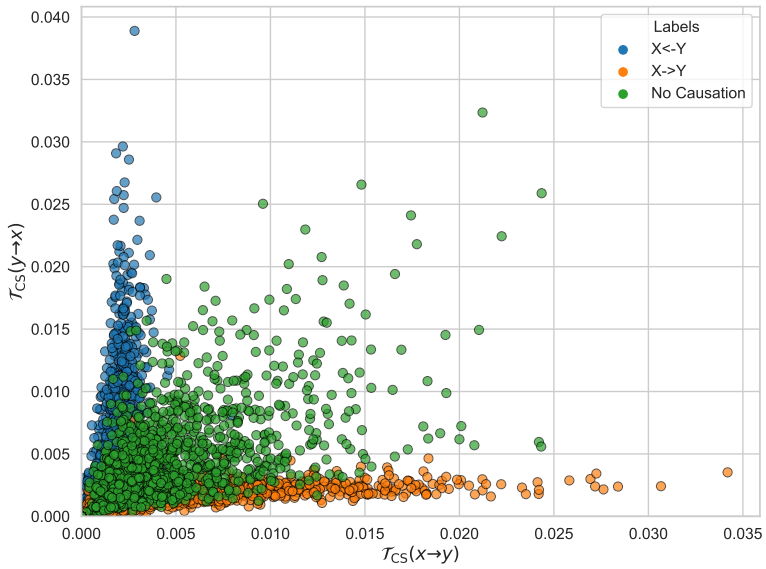

Figure 7: Training data set $\mathcal{T}_{\mathrm{CS}}$ distribution

Taking the nonlinear training data as an example, for each pair of time series labeled with ground truth, $\mathcal{T}_{\mathrm{CS}}$ value pairs are calculated using the Gaussian kernel width method as described, as shown in Fig. 7. Similarly, the test data is processed in the same manner, and both are further projected into a 100-dimensional space using random Fourier features for subsequent model training (for more details, please refer to our paper).

4.4 Nonlinear Test Result

The classifier achieved an accuracy of 0.96 on the test dataset (please refer to our paper).

4.5 Validation of Causal Network Results Using trained classifier

Furthermore, we extract all detected causal relationships, their directions, and the corresponding original time series from the proposed causal network illustrated in Fig. 6. The causal directions depicted in the figure are used as the known ground truth. For the extracted time series, bidirectional $\mathcal{T}_{\mathrm{CS}}$ value pairs are computed and then projected into a higher-dimensional space using random Fourier features. This new dataset is then used as a test set, and the pre-trained classifier is employed to classify the causal types within the test set.

The classifier correctly identified 17 out of the 21 detected causal relationships in the causal graph, achieving an accuracy of 0.81. For details on the four pairs where the model's judgment differs from the causal network, please refer to Table 1.

Index Pair	**True Label**	**Predicted Label**
Shenzhen Component Index - Nikkei 225	X→Y	X←Y
Shenzhen Component Index - Straits Times Index	X→Y	X←Y
Nikkei 225 - MOEX Russia Index	X→Y	No Causation
NZX 50 Index - Hang Seng Index	X→Y	X←Y

Table 1: Causation Table for Index Pairs

Note that due to significant changes in data patterns, ζ needs to be adjusted to ensure that the $\mathcal{T}_{\mathrm{CS}}$ value pairs calculated for the new test set fall within a similar numerical range as those used to train the classifier. Therefore, we employed a grid search to determine that $\zeta = 0.02$.

5 Permutation test

In our paper, to determine the significance of the $\mathcal{T}_{\mathrm{CS}}$ value, we used the permutation test. The permutation test provides a method to assess the significance of a relationship through randomization. It disrupts the temporal dependency between sequences to evaluate the probability of the observed results occurring under conditions of no association. The core idea is to construct a new $\mathcal{T}_{\mathrm{CS}}$ value permutation distribution and compare the original $\mathcal{T}_{\mathrm{CS}}$ value with this distribution, as shown in Algorithm 1.

Algorithm 1 Test the significance of $C(x \to y)$

Require: Two time series $\{x_t\}$ and $\{y_t\}$; Number of permutations P; Significance rate η.
Ensure: Test decision (Is $H_0 : C(x \to y)$ significant or not?).

1: Construct $\{y_{t+1}, x_t^m, y_t^n\}_{t=1}^T$ (T represents the total number of observations) from $\{x_t\}$ and $\{y_t\}$.
2: Compute $C(x \to y) = D(p(y_{t+1}|y_t^n); p(y_{t+1}|y_t^n, x_t^m))$ with $\mathcal{T}_{\mathrm{CS}}$.
3: **for** $m = 1$ to P **do**
4: Construct a pair of surrogate time series $x_{\mathrm{m}}^{\mathrm{surr}}$ and $y_{\mathrm{m}}^{\mathrm{surr}}$.
5: Compute $C(x_{\mathrm{m}}^{\mathrm{surr}} \to y_{\mathrm{m}}^{\mathrm{surr}})$ with $\mathcal{T}_{\mathrm{CS}}$.
6: **end for**
7: **if** $\frac{1+\sum_{m=1}^{P} \mathbf{1}\{C(x \to y) \leq C(x_m^{\mathrm{surr}} \to y_m^{\mathrm{surr}})\}}{1+P} \leq \eta$ then **then**
8: $C(x \to y)$ is not significantly large.
9: **else**
10: $C(x \to y)$ is significantly large.
11: **end if**
12: **return** decision

References

[1] Erwin Lutwak, Deane Yang, and Gaoyong Zhang. "Crame/spl acute/r-Rao and moment-entropy inequalities for Renyi entropy and generalized Fisher information". In: *IEEE Transactions on Information Theory* 51.2 (2005), pp. 473–478.

[2] Josâe C Prâincipe. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer, 2010.

[3] Shujian Yu et al. "Cauchy-Schwarz Divergence Information Bottleneck for Regression". In: *arXiv preprint arXiv:2404.17951* (2024).