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.

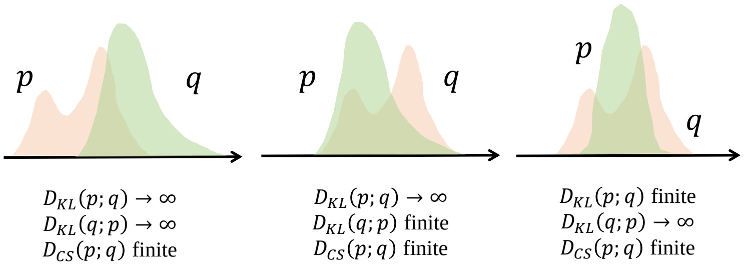

Figure 1: KL divergence is infinite even though there is an overlap between supp(p) and supp(q), but neither is a subset of the other. CS divergence does not have such constraint.

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 $\mathcal{T}_{\text{CS}}$:

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

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):

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

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

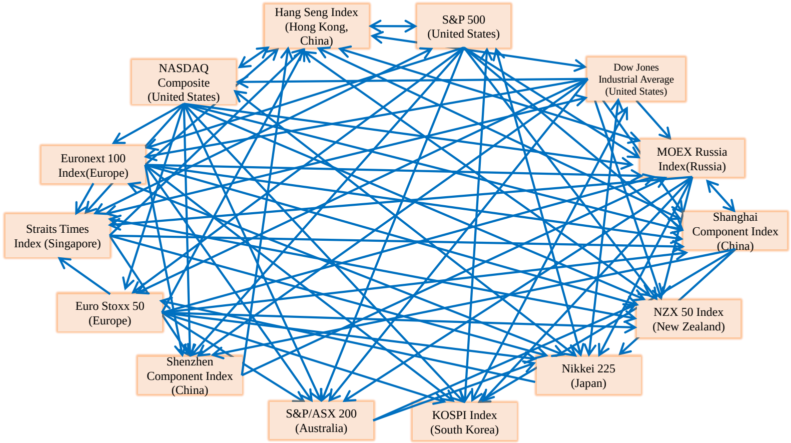
Hang Seng Index (Hong Kong, China)
S&P 500 (United States)
NASDAQ Composite (United States)
Dow Jones Industrial Average (United States)
Euronext 100 Index(Europe)
MOEX Russia Index(Russia)
Straits Times Index (Singapore)
Shanghai Component Index (China)
Euro Stoxx 50 (Europe)
NZX 50 Index (New Zealand)
Shenzhen Component Index (China)
Nikkei 225 (Japan)
S&P/ASX 200 (Australia)
KOSPI Index (South Korea)

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.

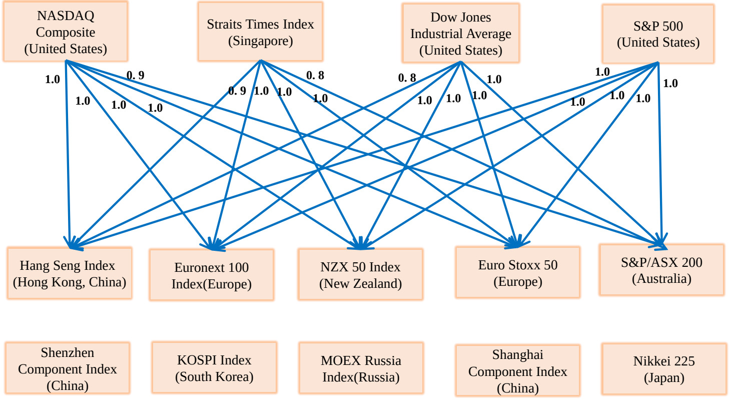
NASDAQ Composite (United States)
Straits Times Index (Singapore)
Dow Jones Industrial Average (United States)
S&P 500 (United States)
1.0
0. 9
1.0
1.0
1.0
0. 9
1.0
1.0
0. 8
1.0
0. 8
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
Hang Seng Index (Hong Kong, China)
Euronext 100 Index(Europe)
NZX 50 Index (New Zealand)
Euro Stoxx 50 (Europe)
S&P/ASX 200 (Australia)
Shenzhen Component Index (China)
KOSPI Index (South Korea)
MOEX Russia Index(Russia)
Shanghai Component Index (China)
Nikkei 225 (Japan)

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.

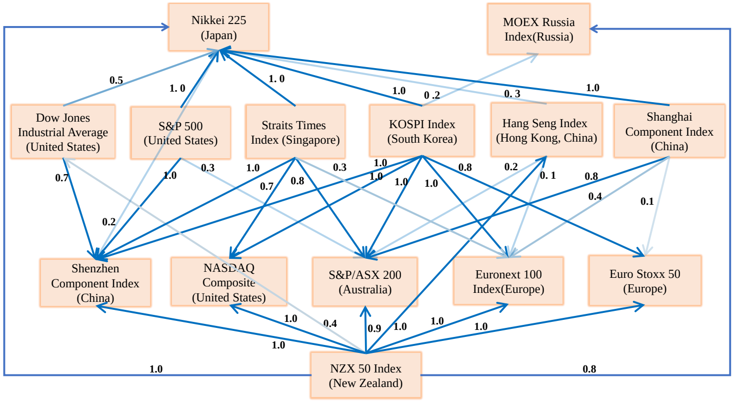

Figure 4: CCS causal network

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.

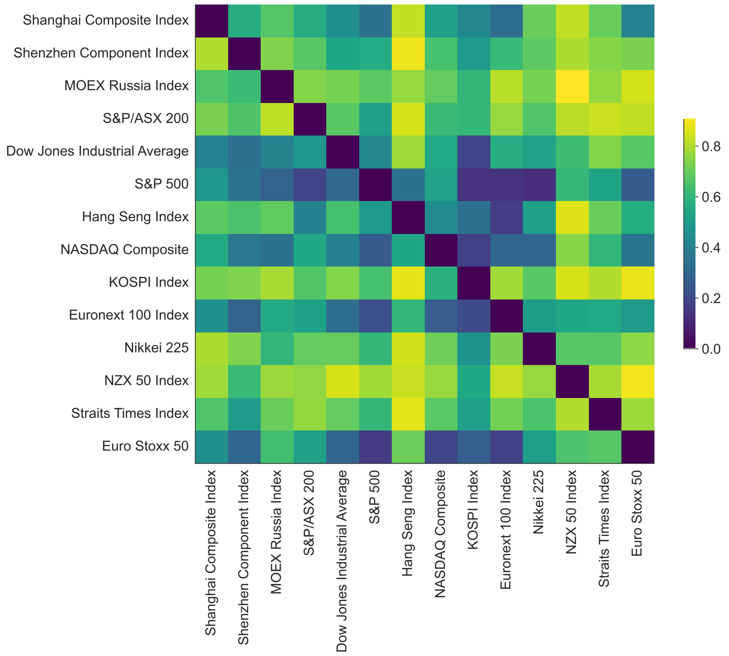

Figure 5: KGC causal network

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.

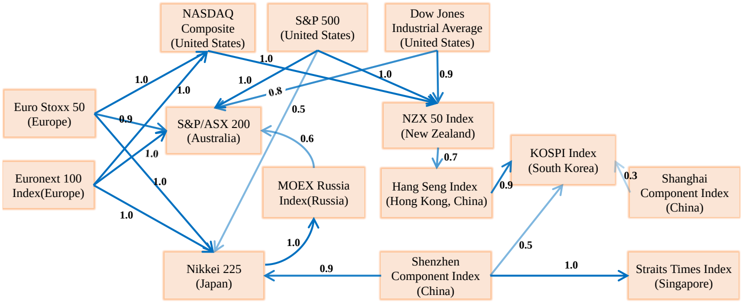
NASDAQ Composite (United States)
S&P 500 (United States)
Dow Jones Industrial Average (United States)
Euro Stoxx 50 (Europe)
Euronext 100 Index(Europe)
S&P/ASX 200 (Australia)
NZX 50 Index (New Zealand)
KOSPI Index (South Korea)
Shanghai Component Index (China)
MOEX Russia Index(Russia)
Hang Seng Index (Hong Kong, China)
Nikkei 225 (Japan)
Shenzhen Component Index (China)
Straits Times Index (Singapore)
1.0
1.0
1.0
1.0
1.0
0.9
0.8
0.9
0.5
0.6
1.0
1.0
1.0
0.7
0.9
0.3
1.0
0.5
0.9
1.0

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.

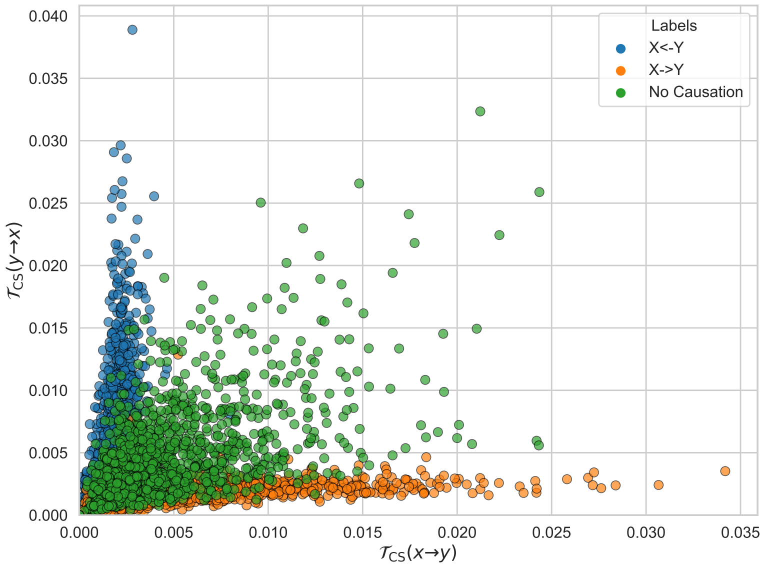

Figure 7: Training data set $\mathcal{T}_{\mathrm{CS}}$ distribution

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 $\mathcal{T}_{\text{CS}}$ 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 $\mathcal{T}_{\text{CS}}$ 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 $\mathcal{T}_{\text{CS}}$ value permutation distribution and compare the original $\mathcal{T}_{\text{CS}}$ 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 $\eta$.
Ensure: Test decision (Is $H_0 : C(x \to y)$ significant or not?).
Construct $\{y_{t+1}, x_t^m, y_t^n\}_{t=1}^T$ ($T$ represents the total number of observations) from $\{x_t\}$ and $\{y_t\}$.
Compute $C(x \to y) = D(p(y_{t+1}|y_t^n); p(y_{t+1}|y_t^n, x_t^m))$ with $\mathcal{T}_{\text{CS}}$.
for $m = 1$ to $P$ do
    Construct a pair of surrogate time series $x_{\text{m}}^{\text{surr}}$ and $y_{\text{m}}^{\text{surr}}$.
    Compute $C(x_{\text{m}}^{\text{surr}} \to y_{\text{m}}^{\text{surr}})$ with $\mathcal{T}_{\text{CS}}$.
end for
if $\frac{1+\sum_{m=1}^{P} \mathbf{1}\{C(x \to y) \leq C(x_m^{\text{surr}} \to y_m^{\text{surr}})\}}{1+P} \leq \eta$ then then
    $C(x \to y)$ is not significantly large.
else
    $C(x \to y)$ is significantly large.
end if
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).