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. Finally, we provide a detailed explanation of the nonlinear data generation strategy used to test the classifier-based approach employing CS-TE, along with a description of the generation, testing, and performance of linear data.

1 Proofs

1.1 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 $\mathcal{T}_{\mathrm{CS}}$:

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

1.2 Estimation

For the first term in Eq.(1), we have:

$$\begin{aligned}&\int p(X_{-1}, Y, Y_{-1})p(X_{-1}, Y_{-1})p(Y, Y_{-1})\, dX_{-1}\, dY\, dY_{-1} \\ &\quad = \mathbb{E}_{p(X_{-1}, Y, Y_{-1})}\left(p(Y_{-1})p(X_{-1}, Y_{-1})p(Y, Y_{-1})\right).\end{aligned} \tag{2}$$

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.(2) can be approximated using a Monte Carlo estimator:

$$\frac{1}{N}\sum_{t=1}^{N} p(y_{t-})p(x_{t-}, y_{t-})p(y_{t+1}, y_{t-}). \tag{3}$$

Further, by using Gaussian kernels for $p(x_{t-}, y_{t-}), p(y_{t+1}, y_{t-}), p(y_{t-})$, Eq.(3) can be expressed

as Eq.(4):

$$
\begin{aligned}
&\approx \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{N(\sqrt{2\pi}\sigma)^{d_{y_{t-}}}}\sum_{i=1}^{N}\exp\left(-\frac{\|y_{j-1}-y_{i-1}\|_2^2}{2\sigma^2}\right)\right)\\
&\cdot\left(\frac{1}{N(\sqrt{2\pi}\sigma)^{d_{x_{t-}}+d_{y_{t-}}}}\sum_{i=1}^{N}\exp\left(-\frac{\|x_{j-1}-x_{i-1}\|_2^2}{2\sigma^2}\right)\exp\left(-\frac{\|y_{j-1}-y_{i-1}\|_2^2}{2\sigma^2}\right)\right)\\
&\cdot\left(\frac{1}{N(\sqrt{2\pi}\sigma)^{d_{y_{t+1}}+d_{y_{t-}}}}\sum_{i=1}^{N}\exp\left(-\frac{\|y_{j+1}-y_{i+1}\|_2^2}{2\sigma^2}\right)\exp\left(-\frac{\|y_{j-1}-y_{i-1}\|_2^2}{2\sigma^2}\right)\right).
\end{aligned}
\tag{4}
$$

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:

$$
\begin{aligned}
&\int p(X_{-1},Y,Y_{-1})p(X_{-1},Y_{-1})p(Y,Y_{-1})\,dX_{-1}\,dY\,dY_{-1}\\
&=\frac{1}{N^4(\sqrt{2\pi}\sigma)^{d_{x_{t-}}+d_{y_{t+1}}+3d_{y_{t-}}}}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}M_{ji}\right)\left(\sum_{i=1}^{N}K_{ji}M_{ji}\right)\left(\sum_{i=1}^{N}L_{ji}M_{ji}\right).
\end{aligned}
\tag{5}
$$

Similarly, For the second and third terms of Eq.(2), we can apply the same pattern to obtain Eq.(6) and Eq.(7).

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

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

Finally, by combining Eq.(5), Eq.(6), and Eq.(7) and eliminating the normalization constant term, we obtain the empirical estimator for Eq.(1):

$$
\begin{aligned}
&\widehat{D}_{\mathrm{CS}}((p(X_{-1},Y,Y_{-1})p(Y_{-1});p(X_{-1},Y_{-1})p(Y,Y_{-1}))\\
&=-2\log\left(\sum_{j=1}^{N}\left(\left(\sum_{i=1}^{N}M_{ji}\right)\left(\sum_{i=1}^{N}K_{ji}M_{ji}\right)\left(\sum_{i=1}^{N}L_{ji}M_{ji}\right)\right)\right)\\
&\quad+\log\left(\sum_{j=1}^{N}\left(\left(\sum_{i=1}^{N}K_{ji}L_{ji}M_{ji}\right)\left(\sum_{i=1}^{N}M_{ji}\right)^2\right)\right)\\
&\quad+\log\left(\sum_{j=1}^{N}\left(\frac{\left(\sum_{i=1}^{N}K_{ji}M_{ji}\right)^2\left(\sum_{i=1}^{N}L_{ji}M_{ji}\right)^2}{\left(\sum_{i=1}^{N}K_{ji}L_{ji}M_{ji}\right)}\right)\right).
\end{aligned}
\tag{8}
$$

1.3 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.(8) 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.

2 Causal Network Base on Different Methods

2.1 Causal Network Base on GC

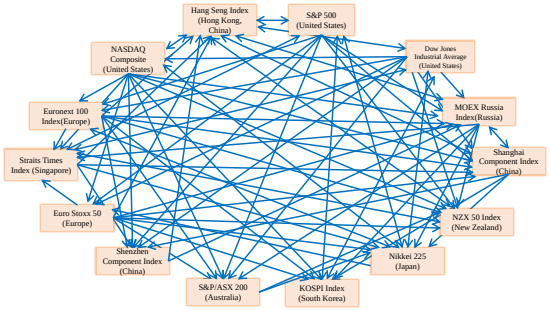

Figure 1: GC causal network

The GC causal network in Fig. 1 revealed numerous significant causal relationships. 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.

2.2 Causal Network Base on TE

According to the results of the TE, in Fig. 2, 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.

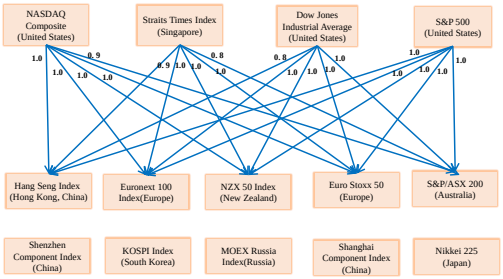

Figure 2: TE causal network

2.3 Causal Network Base on CCS

From the results of CCS in Fig. 3, 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.

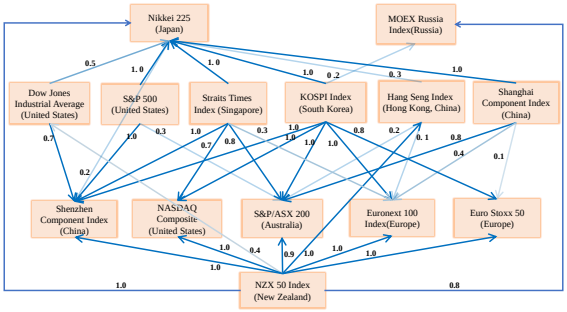

Figure 3: CCS causal network

2.4 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. 4. 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. 4, 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.

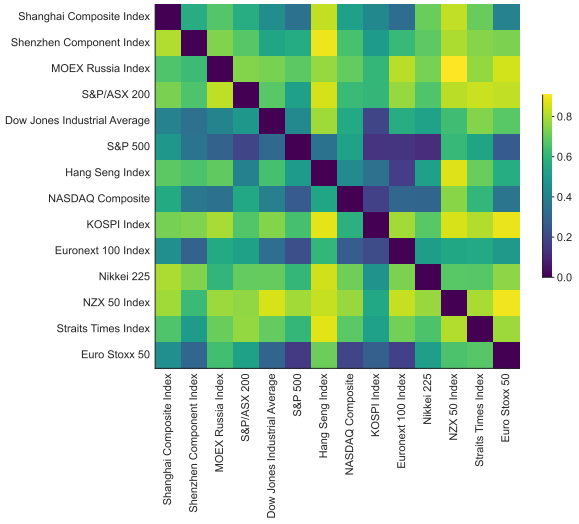

Figure 4: KGC causal network

2.5 Causal Network Base on Our Proposed Method

In Fig. 5, 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.

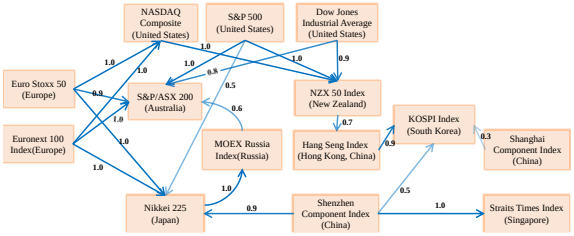

Figure 5: Out proposed method causal network

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.

2.6 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.

3 Nonlinear data synthesis

3.1 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*.

3.2 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 $x \to y$:

$$x_t = 0.5x_{t-1} + 0.9N_x \tag{9}$$

$$y_t = 1.5\exp\left(-(x_{t-1} + x_{t-2})\right) + 0.7\cos\left(y_{t-1}^2\right) + 0.2N_y \tag{10}$$

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

$$y_t = 1.2y_{t-1} + 0.3N_y \tag{11}$$

$$x_t = -1.5\exp(-(y_{t-1} + y_{t-2})^2) + 0.7\cos(x_{t-1}^2) + 0.2N_x \tag{12}$$

- For *No Causation*:

$$x_t = 0.5x_{t-1} + 0.9N_x \tag{13}$$

$$y_t = 1.5\cos\left(y_{t-1}^2\right) + 2.5N_y \tag{14}$$

3.3 Nonlinear test result

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

4 Linear data synthesis

4.1 Linear training data

In our experiments, we also tested the linear causal discrimination ability of the classifier combined with CS-TE. We similarly used the VAR model to synthesize 7,500 time series pairs of length 256 with causal labels. For example, Eq. (15) was used to generate data for the relationship $X \rightarrow Y$:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \frac{1}{p}\sum_{\tau=1}^{p} \begin{bmatrix} a_\tau & 0 \\ c_\tau & d_\tau \end{bmatrix} \begin{bmatrix} x_{t-\tau} \\ y_{t-\tau} \end{bmatrix} + \begin{bmatrix} N_x \\ N_y \end{bmatrix} \tag{15}$$

4.2 Linear test data

Similarly, we used the training data to train the classifier and attempted to modify the data generation strategy to test its performance. We generated 300 pairs of series and evenly distributed causal labels based on Eq.(15) by adjusting the noise variance $p \in \{0.5, 1.0, 1.5, 2.0\}$, lag order $\in \{1, 2, 3\}$.

4.3 Linear test result

The classifier achieved an accuracy of 0.90 on the linear test dataset.