Supplementary Material: Proof of the Empirical Estimator for 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

In our paper *Cauchy-Schwarz Divergence Transfer Entropy*, we introduce a novel formulation of Transfer Entropy (TE) based on the Cauchy-Schwarz (CS) divergence, accompanied by a closed-form estimator for this measure. In this proof, we provide a detailed and rigorous derivation of the empirical estimator for this newly proposed formulation.

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

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

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):

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

3 Conclusions

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.