		Clarity of presentation and structure	Justification of methodology, approach and parameter choices	Content, results, and validation	Consistency of language, code and mathematical notation	Critical interpretation of results and further challenges
		10%	30%	30%	15%	15%
2	80	Excellet presentation lear ad easy to read	The methodology is well introoduced and clearly justified	Results are clar and well persented	Notation is good and consitent	Conclusions and discussins are welll supported by the findings

COMP0047 Data Science Coursework:

Machine Learning Methods for Estimating Transfer Entropy and Conditional Transfer Entropy

Contents

1 Introduction and Motivation

As a group, we sought to use the "GlobalTradeAtlas" dataset to answer the question: Which countries or regions hold the most influence in the global trade of the commodities in the dataset? In my opinion, the idea of "influence" implies causality. But how do we use statistical and computational methods to infer causality? This is the central topic of this project.

1.1 Inferring Causality

Based on the ideas of Wiener (1956), Granger Causality (GC) was the first method developed to quantify causal effects between two time series (Granger, 1969). Informally, we say that "X causes Y if the prediction of Y is enhanced by X". Formally, we say that "X causes Y if Y is not independent of the past of X after conditioning on the past of Y" (Peters *et al.*, 2017).

$$y_{t+1}|\mathbf{y}_{\text{past}} \neq y_{t+1}|\mathbf{x}_{\text{past}}, \mathbf{y}_{\text{past}} \implies y_{t+1} \not\perp\!\!\!\perp \mathbf{x}_{\text{past}}|\mathbf{y}_{\text{past}} \implies X \to Y \tag{1}$$

We can also look from the other direction, and say "X does *not* cause Y if there is no difference between a prediction of Y based on Y and a prediction of Y based on both X and Y":

$$y_{t+1}|\mathbf{y}_{\mathbf{t}-} = y_{t+1}|\mathbf{x}_{\text{past}}, \mathbf{y}_{\text{past}} \implies y_{t+1} \perp\!\!\!\perp \mathbf{x}_{\text{past}}|\mathbf{y}_{\text{past}} \implies X \not\to Y \tag{2}$$

But how can we uncover if $X \not\to Y$ or $X \to Y$ is preferred? The classic GC test uses a linear vector autoregressive (VAR) model. Given data for X and Y, we formulate two alternative regression models for Y, one based on the past of only Y, and one with both X and Y:

$$y_{t+1} = \sum_{i=1}^{k} a_i y_{t-i} + \epsilon_1$$

$$y_{t+1} = \sum_{i=1}^{k} a_i y_{t-i} + \sum_{i=1}^{k} b_i x_{t-i} + \epsilon_2$$

We can then use the residual sum of squares of the two models to calculate an F-statistic. We reject the null hypothesis that $X \not\to Y$ if the observed F value exceeds the $(1-\alpha)\%$ quantile of an F-distribution with appropriately selected degrees of freedom.

While this remains a very widespread method [and we in fact used it in our group presentation], the use of vector autoregressive models restricts GC to detecting linear causality, and causality in the mean (Runge, 2018). There is however a more general metric that can handle arbitrary functional relationships, and effectively estimate the *strength* of causal interactions, not just their presence. This is known as *transfer entropy*.

1.2 Upcoming Content

In the remaining part of this text, we first rigorously define transfer entropy. The subsequent chapter provides an overview of the history of existing computation methods to estimate it. Then, we introduce AGM-TE, a novel machine learning method for estimating transfer entropy that I am developing as part of my MSc research project. In the final chapter, we expand the method to handle *conditional transfer entropy*, and apply it to the commodities data.

2 An Introduction to Transfer Entropy

Schreiber (2000) introduced the concept of *transfer entropy* (TE) to quantify the strength of causal interactions. The most fundamental advantage of transfer entropy over previous conceptualizations of causality (such as Granger causality) is that it can detect any kind of change to the distribution of target variable, and therefore does not require assumptions about the functional nature of the causal relationship between variables.

2.1 Defining Transfer Entropy

Let us consider a random variable Y, which is evolving according to a *discrete time Markov process* of order k. This means that at each timestep t, the conditional probability of finding Y in state y_{t+1} at the next timestep is completely determined by the past k values of y up to the current time t, that is, $p(y_{t+1}|y_t, y_{t-1}, ..., y_{t-k+1}) = p(y_{t+1}|y_t, y_{t-1}, ..., y_{t-k+1}, y_{t-k})$. Such a system can be said to posses the *generalized Markov property*. If we use the shorthand notation $\mathbf{y_t} = \{y_t, y_{t-1}, ..., y_{t-k+1}\}$, this can be expressed more compactly as $p(y_{t+1}|\mathbf{y_t}) = p(y_{t+1}|\mathbf{y_t}, y_{t-k})$. We can expand the concept by also considering the effect of conditioning on the past k values of the random variable X, which we similarly denote $\mathbf{x_t}$. If, after conditioning on the past of Y, also conditioning on $\mathbf{x_t}$ does not influence the future probabilities of Y, we can say that:

$$p(y_{t+1}|\mathbf{y_t}) = p(y_{t+1}|\mathbf{x_t}, \mathbf{y_t}) \tag{3}$$

The *Transfer entropy* from X to Y (denoted $T_{X\to Y}$) is quantifying the expected excess surprise from incorrectly assuming that the distribution of Y is $p(y_{t+1}|\mathbf{y_t})$ and therefore not affected by X, when the actual distribution is $p(y_{t+1}|\mathbf{y_t}, \mathbf{x_t})$. In the original definition by Schreiber (2000), this is formulated as a KL divergence between the $P(y_{t+1}|\mathbf{x_t}, \mathbf{y_t})$ and $P(y_{t+1}|\mathbf{y_t})$ distributions:

$$T_{X\to Y} := \mathbb{E}\left[D_{KL}(P(y_{t+1}|\mathbf{x_t}, \mathbf{y_t})||P(y_{t+1}|\mathbf{y_t}))\right] \tag{4}$$

An equivalent definition is to conceptualize transfer entropy as conditional mutual information (CMI). The corresponding formula takes the form:

$$T_{X\to Y} := \mathcal{I}(y_{t+1}; \mathbf{x_t}|\mathbf{y_t}) \tag{5}$$

This implies that transfer entropy is the mutual information between y_{t+1} and $\mathbf{x_t}$, when accounting for the effects of $\mathbf{y_t}$ on y_{t+1}. As both CMI and KL divergences are by definition non-negative, this means that the minimum value of $T_{X\to Y}$ is 0.

Finally, TE can also be defined as an expression in terms of the difference between the entropies of two conditional distributions, $p(y_{t+1}|y_t)$ and $p(y_{t+1}|x_t, y_t)$ as:

$$\mathcal{T}_{X\to Y} := \mathcal{H}(y_{t+1}|\mathbf{y_t}) - \mathcal{H}(y_{t+1}|\mathbf{x_t}, \mathbf{y_t}) \tag{6}$$

This means that TE is the reduction in the uncertainty of y_{t+1} when it is conditioned on the joint of $\mathbf{x_t}$ and $\mathbf{y_t}$, compared to only conditioning on $\mathbf{y_t}$, or in other words, measuring the additional information on Y available in the past of X, that is not available in the past of Y.

As work in this project relates to a novel method for estimating TE, it is important to introduce the rich history of computational approaches dedicated to inferring TE from observational data.

3 A Short Review of Existing TE Estimation Methods

3.1 kNN Approaches to Computing TE

The first batch of practical methods for TE inference rely on k nearest neighbour (kNN) approaches to estimate information theoretic quantities. These approaches [such as Vejmelka and Paluš (2008)] are based on the approach of Frenzel and Pompe (2007), which itself is a generalization of the KSG method of Kraskov *et al.* (2004) for estimating MI. The fundamental idea here is to utilise the fact that $\mathcal{I}(y_{t+1}; \mathbf{x_t}|\mathbf{y_t})$ can be decomposed into joint entropies as:

$$\mathcal{I}(y_{t+1}; \mathbf{x_t}|\mathbf{y_t}) = \mathcal{H}(y_{t+1}, \mathbf{y_t}) + \mathcal{H}(\mathbf{y_t}, \mathbf{x_t}) - \mathcal{H}(\mathbf{y_t}) - \mathcal{H}(y_{t+1}, \mathbf{y_t}, \mathbf{x_t}) \tag{7}$$

and therefore all relevant entropies can be estimated directly from data without a model.

While these methods were the first to appear, and thus still very widespread, they suffer from poor scaling with increased data dimensionality. Specifically, theory indicates that the performance of KSG approaches [on which all kNN methods for estimating TE are based] degrades exponentially with increasing d (Sricharan *et al.*, 2013; Gao *et al.*, 2018; Zhao and Lai, 2020), a phenomenon that cleary be observed in the empirical results of Mukherjee *et al.* (2019).

3.2 TE Estimation via Variational Lower Bounds on the CMI

Machine learning (ML) researchers are often faced with the problem of estimating mutual information (MI) in high dimensional data, such as between images. This has led to the emergence of so-called *variational methods* to estimate MI. These have since been generalized to estimate conditional mutual information (CMI). We will use the phrase CMI, and the generic $\mathcal{I}(X; Y|Z)$ notation for CMI [rather than $\mathcal{I}(y_{t+1}; \mathbf{x_t}|\mathbf{y_t})$] throughout this section to better conform to existing literature on this topic, but keep in mind that this subsumes TE as a specific case.

The CMI $\mathcal{I}(X; Y|Z)$ can be expressed as a KL divergence between p, the observed joint distribution $p(x, y, z)$ and a hypothetical distribution q, which is constructed under the assumption of conditional independence as $p(x, z)p(y|z)$, as:

$$\mathcal{I}(X; Y|Z) := D_{KL}(p(x, y, z)||p(x, z)p(y|z)) = D_{KL}(p||q) \tag{8}$$

This means that ML methods for estimating CMI require two components: a system that can estimate KL divergences [this is called a *recognition model*, and denoted r], and a *generative model*[denoted q] to generate samples from the hypothetical $p(x, z)p(y|z)$ distribution.

The seminal work of Belghazi *et al.* (2018) introduced the idea of using the Donsker and Varadhan (1983) variational lower bound on the KL divergence [denoted $DV(p||q)$] to build a recognition model. For probability distributions p and q over $\mathcal{X}$ with a finite KL divergence, and for the class of functions $f(s)$:$\mathcal{X} \to \mathbb{R} \in \mathcal{F}$ bounded in expectation, $DV(p||q)$ is:

$$D_{KL}(p||q) \geq DV(p||q) := \sup_{f \in \mathcal{F}} \mathbb{E}_{s \sim p(s)}[f(s)] - \log(\mathbb{E}_{s \sim q(s)}[\exp(f(s))]) \tag{9}$$

The main innovation of Belghazi *et al.* (2018) is to define $\mathcal{F}$ to be the set of functions that can be approximated by a neural network $r_\theta(s)$ with a fixed architecture. This means that the variational problem in $\mathcal{F}$ is reduced to an optimization problem over θ, with better values of θ leading to an increase in the lower bound.

Mukherjee *et al.* (2019) then further improved this approach by showing that if r computes the log likelihood ratio $r(s) = \log \frac{p(s)}{q(s)}$, then the DV lower bound is in fact equal to the KL divergence. As Lopez-Paz and Oquab (2016) showed that the log likelihood ratio can be computed using a classifier which is trained to predict if a given sample s was drawn from p or q, Mukherjee *et al.* (2019) train $r(s)$ to distinguish between samples drawn from the true joint p, and samples drawn from the model q [which approximates $p(x, z)p(y|z)$]. For q, the authors test a variety of established approaches, including Conditional Variational Autoencoders (CVAEs) (Sohn *et al.*, 2015) and simple kNN based permutation. This approach was named CCMI, for classifier-based conditional mutual information. Mondal *et al.* (2020) improve on this idea further with C-MI-GAN [which gets its name from Generative Adversarial Networks]. Rather than treating the problem of learning the recognition model r and the generator q separately [as in CCMI], they define a joint loss function L, which can be used to directly estimate CMI.

When comparing their machine learning based approaches with the traditional kNN method in synthetic data for which the ground truth is known, Mukherjee *et al.* (2019) and Mondal *et al.* (2020) find that CCMI and C-MI-GAN are able to scale to variables with significantly higher dimensionality, while requiring a smaller number of samples. In fact C-MI-GAN is able to perform well even in 100 dimensional data (Fig. 1).

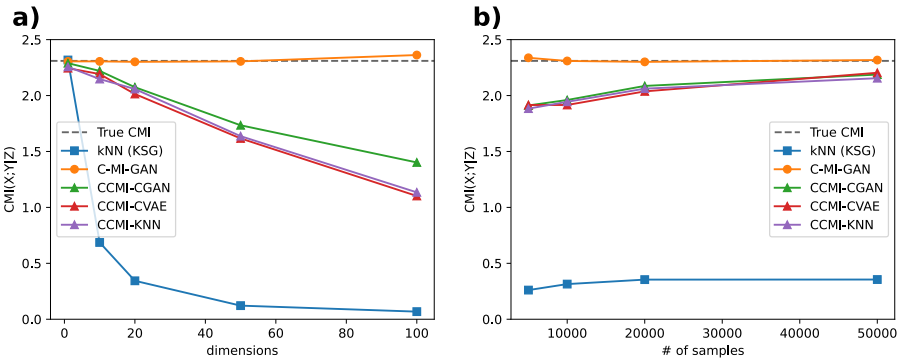

Figure 1: Test results comparing the performance of kNN (KSG) approaches [square marker], CCMI [triangle markers] (with sub-methods are differentiated by the type of generative model), and C-MI-GAN [circle marker], in estimating conditional mutual information. **a)** Performance as a function of data dimensionality in a synthetic system with known ground truth. **b)** Performance in a 20-dimensional synthetic dataset as a function of the number of data samples. Source: modified from Mondal *et al.* (2020) Figure 2 a-b.

While the variational restatement of the CMI estimation problem has resulted in significant improvements in the handling of high-dimensional data compared to kNN-KSG methods, some issues remain. Song and Ermon (2020) show that existing variational estimators are highly biased and that the variance of their estimates grows exponentially with the underlying true CMI. Concerningly, McAllester and Stratos (2020) prove that any distribution-free high confidence lower bound on the KL divergence cannot be larger than $\mathcal{O}(\ln N)$, where N is the number of samples. As the Donsker-Varadhan bound is a special case of this, the CMI estimators of Mukherjee *et al.* (2019) and Mondal *et al.* (2020) are expected to suffer from extremely slow convergence. This also means that applications of CMI estimators to transfer entropy, such as by Zhang *et al.* (2019) will also be prone to underestimation, especially if the true TE is large.

3.3 Estimation using Cross Entropy

McAllester and Stratos (2020) propose a solution: Rather than using a lower bound, they proved that using an upper bound on the CMI will lead to convergence at the rate of $1/\sqrt{N}$.

To derive this upper bound, they use a decomposition of CMI into two conditional entropies:

$$\mathcal{I}(X;Y|Z) := \mathcal{H}(X|Z) - \mathcal{H}(X|Y,Z) \tag{10}$$

To start the estimation process, they learn two models, $Q(X|Z)$ and $Q(X|Y,Z)$. These models can be learned with any method that outputs valid probability distributions over X. For example, if X is discrete, then training a classifier to predict $P(x)$ given an input from Z yields $Q(X|Z)$, and training a second classifier to predict $P(x)$ given both Y and Z yields $Q(X|Y,Z)$.

Then, they exploit the fact that for an empirical data distribution P, and model distribution Q, the *cross entropy* $H(P,Q) := -\mathbb{E}_{x\sim P(X)}[\log Q(x)]$ can be formulated as:

$$H(P,Q) := \mathcal{H}(P) + D_{KL}(P||Q) \tag{11}$$

which means that $H(P,Q)$ is an upper bound on $\mathcal{H}(P)$, the entropy of our data distribution.

This is useful because the cross entropy between discrete data and a model probability mass function [and also between continuous data and certain parametric families] can be trivially calculated. This in turn facilitates the use of the cross entropy as a training objective [loss function]. Training procedures that minimize $H(P,Q)$ will therefore decrease $D_{KL}(P||Q)$ [make Q similar to P], and thereby make the cross entropy an increasing tight upper bound on $\mathcal{H}(P)$.

This means that after training $Q(X|Z)$ and $Q(X|Y,Z)$, we can estimate $\mathcal{I}(X;Y|Z)$ as:

$$\hat{\mathcal{I}}(X;Y|Z) = H(P(X|Z), Q(X|Z)) - H(P(X|Y,Z), Q(X|Y,Z)) \tag{12}$$

Such an estimation strategy was first implemented in the program NJEE by Shalev *et al.* (2022), who also proved that the method is strongly consistent. Garg *et al.* (2022) developed DETE, which used this approach to estimate transfer entropy. They find that their method outperforms existing KSG and variational methods, especially with more delayed causation.

3.4 Outlook

In our opinion however, even these modern approaches make a very serious assumption, which is potentially unwarranted in many situations. Specifically, Shalev *et al.* (2022) and Garg *et al.* (2022) use feedforward neural networks to learn $Q(X|Z)$ and $Q(X|Y,Z)$. The use of feedforward models limits the functional relationships between the variables to be constant in time [this is also true for all of the previous estimation methods discussed here]. This modelling choice relies on an assumption of stationarity, which is not valid for many dynamical systems.

Fortunately, this is not an inherent conceptual limitation of TE, which has been successfully extended to dynamical systems which violate the assumption of stationarity [see Darmon and Rapp (2017); Yin *et al.* (2020)]. To our knowledge however, no machine learning methods have been developed to deal with such non-stationary systems.

We therefore believe that our method AGM-TE, which uses an underlying model of a dynamical system to parametrize continuous probability distributions, could better help estimate transfer entropy in many real world systems of interest.

4 Estimation of Transfer Entropy using AGM-TE

4.1 Problem Statement

Our fundamental objective is to estimate the transfer entropy $\mathcal{T}_{X\to Y}$ of a dynamical system with variables X and Y. We intend to accomplish this using the method of McAllester and Stratos (2020), which means that we express $\mathcal{T}_{X\to Y}$ by decomposing it into conditional entropies as:

$$\mathcal{T}_{X\to Y} := \mathcal{H}(y_{t+1}|\mathbf{y_t}) - \mathcal{H}(y_{t+1}|\mathbf{y_t}, \mathbf{x_t})$$

We then intend to estimate this using the following difference of cross entropies:

$$\hat{\mathcal{T}}_{X\to Y} = H(p_{\psi_t}(y_{t+1}|\mathbf{y_t}), q_{\phi_t}(y_{t+1}|\mathbf{y_t})) - H(p_{\psi_t}(y_{t+1}|\mathbf{y_t}, \mathbf{x_t}), q_{\phi_t}(y_{t+1}|\mathbf{y_t}, \mathbf{x_t})) \tag{13}$$

Let us unpack this dense formula. Recall that the bold subscript $\mathbf{t}$, such as in $\mathbf{y_t}$ and $\mathbf{x_t}$, denotes the k past values of the variable. y_{t+1} is the upcoming observation of y if we are currently at time t. Then, $p_{\psi_t}(y_{t+1}|\mathbf{y_t})$ and $p_{\psi_t}(y_{t+1}|\mathbf{y_t}, \mathbf{x_t})$ represent the "true" [target] distributions over y_{t+1}, given the past k values of only y, or both x and y. The parametrisation by ψ_t is meant to convey the fact that the parameters ψ of the conditional probability distributions that describe relationships between y_{t+1}, $\mathbf{y_t}$, and $\mathbf{x_t}$ change over time, and therefore the same "inputs" do not always lead to the same "outputs" [unlike in a stationary system].

Our goal is to construct the *approximate generative models*, $q_{\phi_t}(y_{t+1}|\mathbf{y_t})$ and $q_{\phi_t}(y_{t+1}|\mathbf{y_t}, \mathbf{x_t})$. These models should take as input either just $\mathbf{y_t}$, or both $\mathbf{y_t}$ and $\mathbf{x_t}$, and yield the parameters ϕ_t needed to specify a valid probability distribution over y_{t+1}. Given these probability distributions and the data, we can calculate the cross entropies in Eq. 13. We therefore call this approach AGM-TE, which stands for *approximate generative model* estimate of *transfer entropy*

4.2 The Approximate Generative Model

4.2.1 Model Architecture

In this section, which details the model architecture, we will, for the sake of brevity, restrict ourselves to the case of modelling $y_{t+1}|\mathbf{y_t}$. However, all the ideas described here apply equally to $y_{t+1}|\mathbf{y_t}, \mathbf{x_t}$, which will be reintroduced later.

Our dataset $\mathbf{Y}$ is single sample of *consecutive* observations of y from $p_{\psi_t}(y_{t+1}|\mathbf{y_t})$ over time $t \in 1 : T$. For the sake of generality, $y \in \mathbb{R}^d$ is a d dimensional vector, so $\mathbf{Y}$ is a $T \times d$ real-valued matrix. Given these restrictions, we must ask: What properties of $p_{\psi_t}(y_{t+1}|\mathbf{y_t})$ can we expect to estimate directly from $\mathbf{Y}$? Are there properties that we can estimate using our model? We should then select the parameters for our model such that it can express inferable properties of $\mathbf{Y}$, while minimizing further assumptions.

As $\mathbf{Y}$ consists of point mass observations derived from a non-stationary system, we expect to have one sample for most combinations of y_{t+1} and $\mathbf{y_t}$ values in the dataset. However, this single sample is enough to [crudely] estimate $\mathbb{E}[p_{\psi_t}(y_{t+1}|\mathbf{y_t})]$. This means that $\mathbf{Y}$ can be used to constrain the expected value of $q_{\phi_t}(y_{t+1}|\mathbf{y_t})$, which we denote $\hat{y}_{t+1}$.

$$\mathbb{E}[q_{\phi_t}(y_{t+1}|\mathbf{y_t})] := \hat{y}_{t+1} \approx y_{t+1}$$

We can then use the *model mean* $\hat{y}_{t+1}$ and the *observed* value of y_{t+1} to calculate the squared prediction error, and thereby constrain the variance of $q_{\phi_t}(y_{t+1}|\mathbf{y_t})$ as:

$$\mathrm{Var}(q_{\phi_t}(y_{t+1}|\mathbf{y_t})) \approx \mathbb{E}[(y_{t+1} - \hat{y}_{t+1})^2]$$

This way, our variance is providing a measure of model uncertainty. For cases where the mean and variance of a distribution are the only parameters which can be reasonably resolved, the *principle of maximum entropy* (Jaynes, 1957) tells us that to minimize further assumptions, we should use a multivariate Gaussian as the parametric family.

To parametrise a multivariate Gaussian, we need a vector of means and a covariance matrix. This means that the number of parameters that our model needs to estimate scales as $\mathcal{O}(d) + \mathcal{O}(d^2)$ [d is the dimensionality of y]. However, if we limit the covariance matrix to be diagonal, then the number of parameters now scales as $\mathcal{O}(d + d) = \mathcal{O}(d)$. This should mean that an increase in d only requires a corresponding linear increase to T to maintain identifiability.

As described in the problem statement, our model $q_{\phi_t}(y_{t+1}|\mathbf{y_t})$ should take in as input $\mathbf{y_t}$, to yield the parameters ϕ, which we have now established as $\phi = \{\mu, \sigma^2\}$. The selection of parameters for the multivariate Gaussian from the input $\mathbf{y_t}$ is accomplished by some model g_θ. To minimize our assumptions about the properties of the dynamical systems that we seek to model, we use Recurrent Neural Networks (RNNs), a class of models which are universal approximators of dynamical systems (Schäfer and Zimmermann, 2006; Chen *et al.*, 2023).

Another advantage of RNNs is that when given a sequence of inputs $y_t, y_{t+1}, y_{t+2}, ...$ they operate on one element at a time, but retain a compressed representation of the history of inputs. This means that an RNN g_θ can correctly take into account the past k values of y, that is $\mathbf{y_t}$, while only requiring a single y_t value for each computation. This means that while we explicitly drop the use of the bold $\mathbf{t}$ when discussing the model to better represent the actual computation it performs, our procedure nonetheless retains long term information about the history of y_t.

4.2.2 Learning the Approximate Generative Model

Our approximate generative model $q_{\phi_t}(y_{t+1}|\mathbf{y_t})$, is a multivariate diagonal Gaussian parametrised by recurrent neural network g_θ. For each y_t in $\mathbf{Y}$, the RNN g_θ computes the parameters:

$$\mu_{t+1}, \sigma^2_{t+1} = g_\theta(y_t)$$

Therefore $q_{\phi_t}(y_{t+1}|\mathbf{y_t})$ can be described as the following composition,

$$p_{\psi_t}(y_{t+1}|\mathbf{y_t}) \approx q_{\phi_t}(y_{t+1}|\mathbf{y_t}) := \mathcal{N}(\mu_{t+1}, \sigma^2_{t+1}) := \mathcal{N}(g_\theta(y_t)) \tag{14}$$

Training therfore involves the optimization of θ, such that the μ_{t+1} and σ_{t+1} parameters produced by g_θ in response to the input y_t maximize the probability of observing each y_{t+1}, and thus $\mathbf{Y}$ overall. For a multivariate diagonal Gaussian with parameters μ_t and σ_t, the negative log likelihood (NLL) of observing a specific y_t is given by:

$$-\log \mathcal{L}(\mu_t, \sigma^2_t|y_t) = \frac{1}{2}\left[d\log(2\pi) + \sum_{j=1}^{d}\log\sigma_j^2 + \sum_{j=1}^{d}\frac{(y_j - \mu_j)^2}{\sigma_j^2}\right] \tag{15}$$

We then average the NLL across timesteps $t \in 1:T$ in the training data, such that the final loss that we minimise [using stochastic gradient descent] is:

$$\text{Loss}(\theta, \mathbf{Y}) := \frac{1}{T}\sum_{t=1}^{T} -\log \mathcal{L}(\mu_{t+1}, \sigma^2_{t+1}|y_{t+1}) := \frac{1}{T}\sum_{t=1}^{T} -\log \mathcal{L}(g_\theta(y_t)|y_{t+1}) \tag{16}$$

4.2.3 Estimating Transfer Entropy using the Approximate Generative Models

Our loss [Eq. 16] is the expectation of the NLL. This in fact, is also the empirical estimate of the *cross entropy* between the true $p_{\psi_t}(y_{t+1}|\mathbf{y_t})$ and our model $q_{\phi_t}(y_{t+1}|\mathbf{y_t})$ (Goodfellow *et al.*, 2016). As we established in the previous chapter, minimization of the cross entropy will make the upper bound converge to the entropy of the data distribution. This means that as training progresses, $Loss(\theta, \mathbf{Y})$ will become an increasingly accurate estimate of $\mathcal{H}(\psi_t(y_{t+1}|\mathbf{y_t})$.

If we now reintroduce the problem of modelling $y_{t+1}|\mathbf{y_t}, \mathbf{x_t}$, and denote the new parametrising neural network as $g_{\theta_2}(y_t, x_t)$, and our previously discussed model of $y_{t+1}|\mathbf{y_t}$ as $g_{\theta_1}(y_t)$, then we can show that the following formula:

$$\hat{\mathcal{T}}_{X\to Y} = \left[\frac{1}{T}\sum_{t=1}^{T} -\log \mathcal{L}(g_{\theta_1}(y_t)|y_{t+1})\right] - \left[\frac{1}{T}\sum_{t=1}^{T} -\log \mathcal{L}(g_{\theta_2}(y_t, x_t)|y_{t+1})\right] \tag{17}$$

will form the estimate of transfer entropy which we defined in Eq. 13.

Furthermore, we can show that because the two cross entropies are being subtracted, the difference between the true and estimated transfer entropy is:

$$\mathcal{T}_{X\to Y} - \hat{\mathcal{T}}_{X\to Y} = D_{KL}(p_{\psi_t}(y_{t+1}|\mathbf{y_t})||q_{\phi_t}(y_{t+1}|\mathbf{y_t})) - D_{KL}(p_{\psi_t}(y_{t+1}|\mathbf{y_t}, \mathbf{x_t})||q_{\phi_t}(y_{t+1}|\mathbf{y_t}, \mathbf{x_t}))$$

As KL divergences are by definition non-negative, the model errors counteract each other, leading to a smaller overall error bias for $\hat{\mathcal{T}}_{X\to Y}$ (Garg *et al.*, 2022).

Taken together, this means that given enough training data and sufficient time to converge during training, AGM-TE can be expected to produce good estimates of transfer entropy.

4.3 Validation in Synthetic Data

4.3.1 The Synthetic Bivariate Linear Gaussian Data Generating Model

The time series $\mathbf{Y}$ and $\mathbf{X}$ are drawn from a bivariate linear Gaussian model described using the equations below. This can also be represented as the *time series graph* (TSG) in Fig. 2.

$$x_{t+1} = b_x x_t + \mathcal{E} \sim \mathcal{N}(0, \sigma_x^2) + \lambda y_t$$
$$y_{t+1} = b_y y_t + \mathcal{E} \sim \mathcal{N}(0, \sigma_y^2)$$

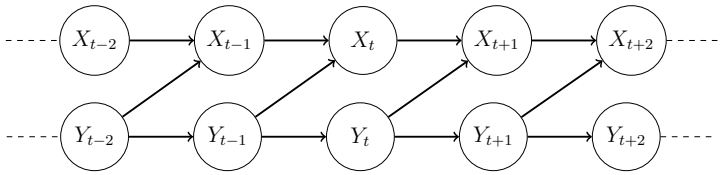

Figure 2: Time series graph of the bivariate linear Gaussian system.

From the equations and the causal graph it is clear that $\mathcal{T}_{X\to Y}$ is always 0. This system is useful for validation, because Edinburgh *et al.* (2021) derived an analytic formula for how $\mathcal{T}_{Y\to X}$ increases with the coupling parameter λ. This formula will be used as the ground truth.

4.3.2 Validating Theoretical Convergence Properties

In the first series of tests, we wish to see if we can expect the minimization of NLL over training to cause the TE estimates to converge to the true $\mathcal{T}_{X\to Y}$ and $\mathcal{T}_{Y\to X}$ values, as predicted by the theory. We set $\lambda = 0.7$ and train 50 replicates of AGM-TE on a dataset of $T = 20000$. If we plot the within-sample estimates of transfer entropy across the replicates (Fig. 3), we can clearly see that the TE estimates become increasingly accurate over training.

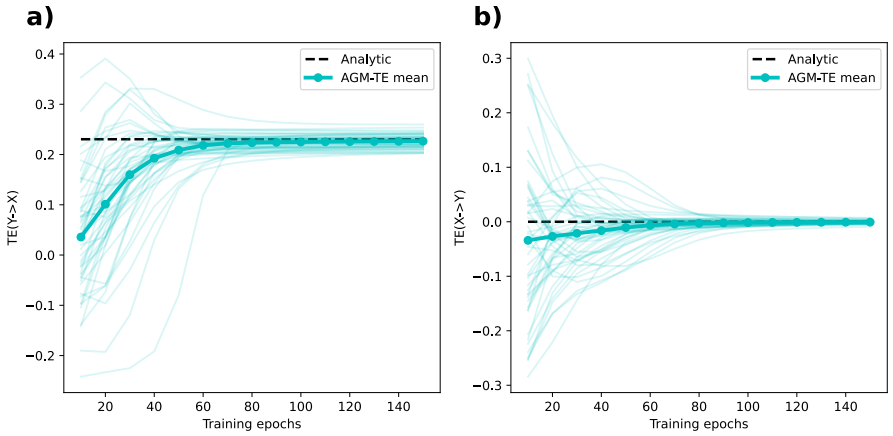

Figure 3: As training progresses, the average over 50 replicate $\mathcal{T}_{X\to Y}$ and $\mathcal{T}_{Y\to X}$ estimates from AGM-TE converges to the analytical ground truth. Faint blue lines are the estimates of independent replicates.

4.3.3 Comparing Sample Efficiency and Accuracy to kNN Approaches

How will AGM-TE and kNN methods compare if we only have data from $T = 2000$? How does the accuracy scale across different TE values? To find out, we compare over a range of 10 equally spaced λ (coupling strength) values from 0 to 1. For each λ, we computed means and standard deviations across 20 independent replicates of an AGM-TE model with 2 neurons trained for 250 epochs, and 100 independent replicates of the kNN approach. We find that the TE estimates produced by AGM-TE are not only less biased the kNN approach [which consistently overestimates TE in this dataset], but also have less variance, matching the ground truth analytic $\mathcal{T}_{X\to Y}$ and $\mathcal{T}_{Y\to X}$ values in a highly consistent manner (Fig. 4).

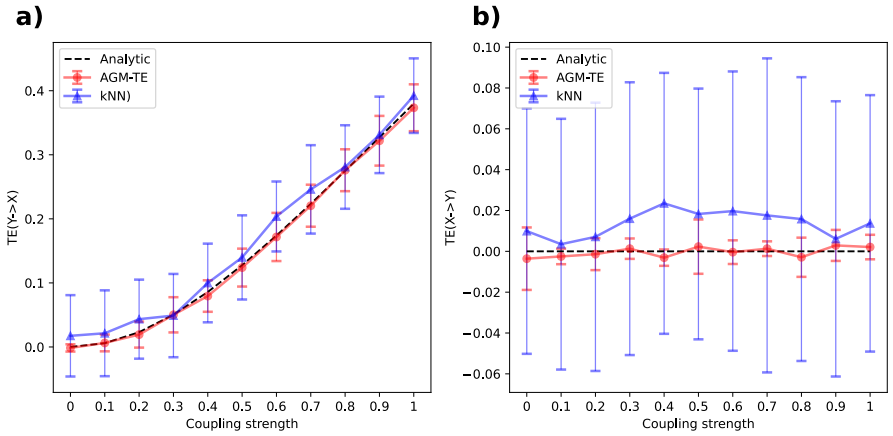

Figure 4: Comparison of the analytic ground truth with kNN and AGM-TE estimates across a range of λs. Error bars signify $\pm$ 1 standard deviation across replicates. **a)** estimates of $\mathcal{T}_{Y\to X}$ **b)** estimates of $\mathcal{T}_{X\to Y}$

5 Expanding AGM-TE to Conditional Transfer Entropy

5.1 Introduction to Conditional Transfer Entropy

An extremely strong assumption of all the approaches discussed (including AGM-TE) is that of *causal sufficiency*. This means that we assume that the causal relationship between X and Y is fully described considering only those variables (Peters *et al.*, 2017). But what if we have access to measurements from a third variable, Z? How can ignoring the effect of Z lead to erroneous conclusions about causal relationships between X and Y?

First, consider a case where Z affects both X and Y. This would be denoted $X \leftarrow Z \rightarrow Y$, and is known as a *fork*. If the information from Z reaches X before Y, we will have $\mathcal{T}_{X \to Y} > 0$, despite the fact that there is no causal interaction between X and Y. Second, we can consider a case where the causal diagram is $X \rightarrow Z \rightarrow Y$, which is referred to as a *chain*. Here, we will have $\mathcal{T}_{X \to Y} > 0$, despite the fact that X does not directly cause Y.

To avoid these issues, which are expected to be common in real-world systems, conditional transfer entropy (CTE), denoted $\mathcal{T}_{X \to Y|Z}$ was proposed (James *et al.*, 2016; Shahsavari Baboukani *et al.*, 2020). The conditioning on Z should help reduce cases where transfer entropy is positive, but no actual causal relationship exists.

As with basic TE, one of the ways to formulate CTE is using a difference in conditional entropies:

$$T_{X \to Y|Z} := \mathcal{H}(y_{t+1}|\mathbf{y_t}, \mathbf{z_t})) - \mathcal{H}(y_{t+1}|\mathbf{x_t}, \mathbf{y_t}, \mathbf{z_t})) \tag{18}$$

This can be interpreted as saying that CTE is the reduction in the uncertainty of y_{t+1} when it is conditioned on the joint of $\mathbf{x_t}$, $\mathbf{y_t}$, and $\mathbf{z_t}$ compared to only conditioning on the joint of $\mathbf{z_t}$ and $\mathbf{y_t}$. In other words, CTE is measuring the additional information on Y available in the past of X, that is not available in the past of Y and Z.

Based on this decomposition and interpretation, the estimation of CTE by AGM-TE will require two models, $q_{\phi_t}(y_{t+1}|\mathbf{y_t}, \mathbf{z_t})$ and $q_{\phi_t}(y_{t+1}|\mathbf{x_t}, \mathbf{y_t}, \mathbf{z_t})$.

5.2 Validating in Synthetic Data

As with classical transfer entropy, we first seek to establish the validity of our approach in synthetic data. As the synthetic systems for this chapter do not have analytically tractable TEs or CTEs, we will use the sample hungry, but established KNN approach as a reference. As KNN approaches work by estimating joint entropies, we decompose the CTE as:

$$T_{X \to Y|Z} = [\mathcal{H}(y_{t+1}, \mathbf{y_t}, \mathbf{z_t}) - \mathcal{H}(\mathbf{y_t}, \mathbf{z_t})] - [\mathcal{H}(y_{t+1}, \mathbf{y_t}, \mathbf{x_t}, \mathbf{z_t}) - \mathcal{H}(\mathbf{y_t}, \mathbf{x_t}, \mathbf{z_t})] \tag{19}$$

5.2.1 The Synthetic Trivariate Fork Model

The first system in which we test the proposed CTE estimation approach is a forking model. This is a trivariate linear Gaussian system described by the following equations and the time series graph in Fig. 5.

$$\begin{aligned} x_{t+1} &= b_x x_t + \mathcal{E} \sim \mathcal{N}(0, \sigma_x^2) + \lambda y_t \\ y_{t+1} &= b_y y_t + \mathcal{E} \sim \mathcal{N}(0, \sigma_y^2) \\ z_{t+1} &= b_z z_t + \mathcal{E} \sim \mathcal{N}(0, \sigma_z^2) + \lambda y_{t-1} \end{aligned}$$

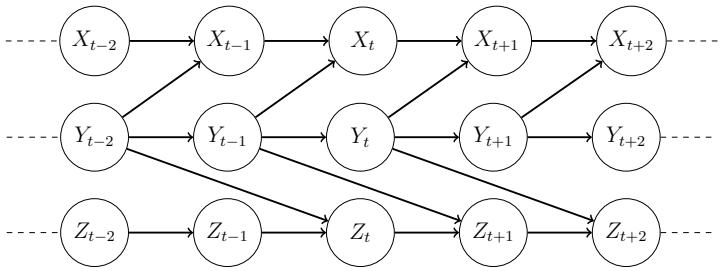

Figure 5: Time series graph of the trivariate fork system.

As the influence of y_t arrives in x_{t+1} before arriving in z_{t+2}, x_{t+1} carries information which reduces uncertainty in z_{t+2}. This actually means that a correct method for estimating $\mathcal{T}_{X\to Z}$ *should* yield a TE > 0, despite the fact that clearly $X \not\to Z$.

In this system, we tested in a high sample size environment ($T = 10000$), and used 100 replicates for kNN, and 20 replicates for AGM-TE models with 3 neurons trained for 500 epochs. We expect that correct systems should yield $\mathcal{T}_{X\to Z} > 0$. We know for a fact that conditioning on Y should reduce $T_{X\to Z|Y}$ to 0. We therefore looked at the percentage decrease due to conditioning:

method	$\mathcal{T}_{X\to Z}$	$\mathcal{T}_{X\to Z\|Y}$	decrease
kNN	0.0317	0.0290	8.51%
AGM-TE	0.0296	0.0030	89.86%

We see that despite starting from similar $\mathcal{T}_{X\to Z}$ estimates, conditioning on Y is able to reduce the $T_{X\to Z|Y}$ estimates of AGM-TE more than 10x compared to the kNN approach.

5.2.2 The Synthetic Trivariate Chain Model

The second system in which we test our approach is described by the following set of equations and the time series graph in Fig. 6.

$$
\begin{aligned}
x_{t+1} &= b_x x_t + \mathcal{E} \sim \mathcal{N}(0, \sigma_x^2) \\
y_{t+1} &= b_y y_t + \mathcal{E} \sim \mathcal{N}(0, \sigma_y^2) + \lambda x_t \\
z_{t+1} &= b_z z_t + \mathcal{E} \sim \mathcal{N}(0, \sigma_z^2) + \lambda y_t
\end{aligned}
$$

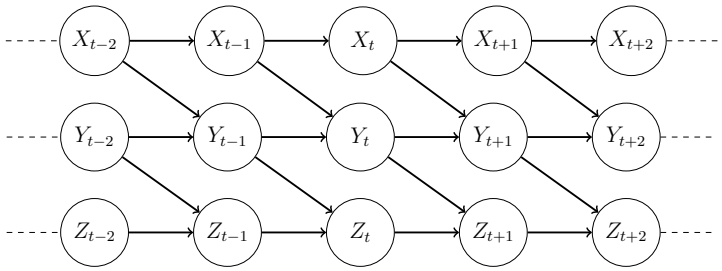

Figure 6: Time series graph of the trivariate chain system.

We expect the influence of x_t to arrive in z_{t+2} through y_{t+1}, causing $\mathcal{T}_{X\to Z}$ to be > 0, despite the fact that X is not directly causing Z. We should however, expect $T_{X\to Z|Y}$ to be 0.

As we intend to use our estimators in empirical data with very few samples, we looked at how the mean squared error of $T_{X\to Z|Y}$ [across 20 AGM-TE and 100 kNN replicates] changes as if we restrict our data sample sizes to $T = \{2000, 1000, 500, 200, 100\}$. The results in Fig. 7 clearly show that AGM-TE is able to reasonably estimate CTE, even in very data-poor environments.

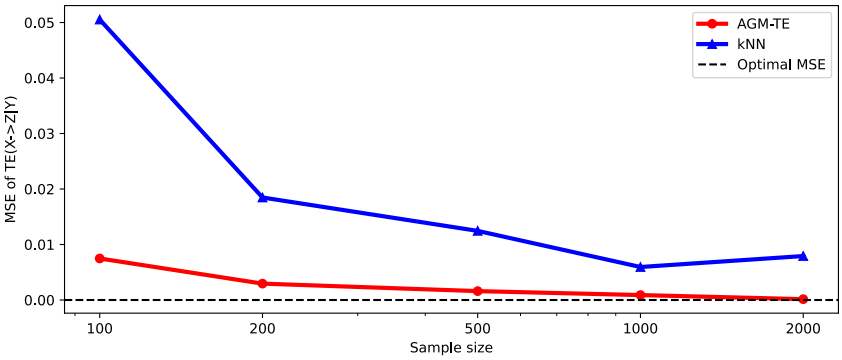

Figure 7: Comparison of the MSE of kNN and AGM-TE estimates of $T_{X \to Z|Y}$ across a range of small sample sizes. The correct value of $T_{X \to Z|Y}$ is 0, and the most optimal MSE is also 0. These results indicate that for low dimensional systems, AGM-TE is able to estimate CTE even at small sample sizes.

5.3 Empirical Testing in the Iron and Steel Dataset

5.3.1 Introduction to the Dataset, Preprocessing, and the Problem Statement

The "Iron and Steel" dataset contains monthly data on the amount of iron and steel exports between various countries from 01/2013 to 10/2024. To avoid problems due to currency conversion and inflation, we used the tonnage data. We selected the 10 largest exporters, and 10 largest importers by total volume, as traffic from these largest exporters and largest importers alone accounts for 37% of the total volume by tonnage recorded in the dataset.

Unfortunately, there were many months where data for a given exporter-importer pair was missing. As most of these cases were after 2022, we discarded data after 12/2021. In the remaining cases, we attempted to substitute the data from the most recent month. If data for a given exporter-importer pair was missing for 2 or more consecutive months, we removed them from the dataset. This missing data affected 1 major exporter, and 1 major importer.

This leaves us with 9 datasets containing the export volumes of Russia, China, Japan, Brazil, South Korea, Germany, India, Belgium, and France into the United States, Italy, China, Turkey, Germany, Taiwan, Thailand, Mexico, and South Korea. For cases where top exporters are also top importers, the datasets were 108×8, and 108×9 in other cases.

For a given pair of countries X and Y, (where both X and Y are one of the 9 top exporters) we are interested in $T_{X \to Y|Z}$ and $T_{Y \to X|Z}$, where Z represents the data from the other 7 exporters. That is, we are interested in whether the exports from one country can help predict [reduce uncertainty in] the future exports of another country, while conditioning on the effect of all other exporting countries. This operationalizes our notion of global "influence".

5.3.2 Analysis Procedure

As Z is formed by concatenating results from the remaining 7 countries, it has $7\times$ higher dimensionality than X and Y. To avoid overwhelming the CTE estimators, we train an autoencoder to reduce the dimensionality of Z. An autoencoder has an *encoder*, which reduces the dimensionality of the data, and a *decoder* which then attempts to reconstruct the original from the lower dimensional representation. The model is trained to minimize the difference between the input and the reconstructed output. By using the output of the encoder we can reduce Z down to 8 dimensions. This same Z is shared for a given pair of $T_{X \to Y|Z}$ and $T_{Y \to X|Z}$ estimates.

Then, we perform 30 replicate AGM-TE analyses for $T_{X \to Y|Z}$. To reduce the effects of outliers, we discard the 5 smallest and 5 largest results, yielding 20 values. We calculate the mean, and the standard error of the mean (SEM). To make our results maximally conservative [that is, minimize CTE estimates], and reduce false positives, we subtract the SEM from the mean, and then clip the values to be non-negative. An identical procedure is completed for $T_{Y \to X|Z}$

5.3.3 Results and Interpretation

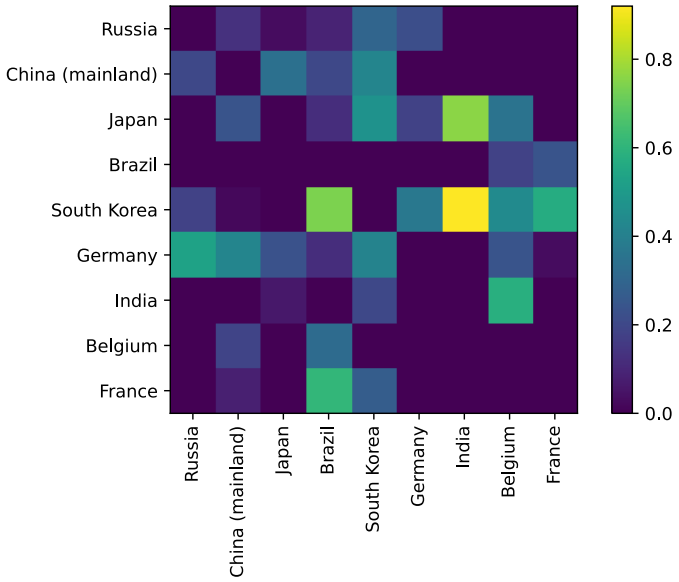

Figure 8: Conditional transfer entropies between the volumes of nine large exporters. Colour indicates the strength of the inferred $T_{\text{row country} \to \text{column country}|\text{all other countries}}$

Fig 8 shows the results from our analyses. One way to aid interpretation is to sum over rows of this matrix of transfer entropies to yield an estimate of a given exporters "overall influence" on the global market. In our analyses, this total is the biggest for South Korea (3.24), followed by Japan (2.11), Germany (1.97), and China (1.16). This is reassuring, as these four countries were the top four exporters of iron and steel worldwide.

Unfortunately, the "overall influence" scores do not correlate with total monetary values, another valid measure of influence, and in fact, the ordering is opposite, as the order is China ($69.8B), Germany ($36.9B), Japan ($35.7B), and South Korea ($29B).

This may be due to multiple factors. First, we used tonnage, rather than value to fit our model. Second, we excluded exports from the 175 exporters representing the remaining 63% of the market. However, the biggest issue is the small amount [108 samples] of temporal data. While we showed in synthetic datasets that AGM-TE can work as few as 100 samples, that was for one, not eight dimensional variables, and if model performance scales linearly with d, we would in fact need samples from 800 time points. For this dataset, our SEM was around 30% of our raw means, which indicates very high uncertainty in our CTE estimates across replicates.

Based on these results, I would *not* consider applying AGM-TE to datasets with low temporal resolution, and limited sample sizes. Perhaps it is more suited to analyse high frequency data.

References

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. 2018. Mine: Mutual information neural estimation.

Chen, X., Tao, Y., Xu, W., and Yau, S. S.-T. 2023. Recurrent neural networks are universal approximators with stochastic inputs. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10): 7992–8006.

Darmon, D. and Rapp, P. E. 2017. Specific transfer entropy and other state-dependent transfer entropies for continuous-state input-output systems. *Physical Review E*, 96(2): 022121.

Donsker, M. D. and Varadhan, S. R. S. 1983. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2): 183–212.

Edinburgh, T., Eglen, S. J., and Ercole, A. 2021. Causality indices for bivariate time series data: A comparative review of performance. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(8).

Frenzel, S. and Pompe, B. 2007. Partial mutual information for coupling analysis of multivariate time series. *Physical Review Letters*, 99(20): 204101.

Gao, W., Oh, S., and Viswanath, P. 2018. Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory*, 64(8): 5629–5661.

Garg, S., Gupta, U., Chen, Y., Datta Gupta, S., Adler, Y., Schneider, A., and Nevmyvaka, Y. 2022. Estimating transfer entropy under long ranged dependencies. *The 38th Conference on Uncertainty in Artificial Intelligence*.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*. MIT Press.

Granger, C. W. J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3): 424.

James, R. G., Barnett, N., and Crutchfield, J. P. 2016. Information flows? a critique of transfer entropies. *Physical Review Letters*, 116(23): 238701.

Jaynes, E. T. 1957. Information theory and statistical mechanics. *Physical Review*, 106(4): 620–630.

Kraskov, A., Stögbauer, H., and Grassberger, P. 2004. Estimating mutual information. *Physical Review E*, 69(6): 066138.

Lopez-Paz, D. and Oquab, M. 2016. Revisiting classifier two-sample tests.

McAllester, D. and Stratos, K. 2020. Formal limitations on the measurement of mutual information.

Mondal, A. K., Bhattacharya, A., Mukherjee, S., AP, P., Kannan, S., and Asnani, H. 2020. C-mi-gan : Estimation of conditional mutual information using minmax formulation.

Mukherjee, S., Asnani, H., and Kannan, S. 2019. Ccmi : Classifier based conditional mutual information estimation.

Peters, J., Janzing, D., and Bernhard, S. 2017. *Elements of causal inference*. The MIT Press, Cambridge, Massachusetts.

Runge, J. 2018. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7).

Schreiber, T. 2000. Measuring information transfer. *Physical Review Letters*, 85(2).

Schäfer, A. M. and Zimmermann, H. G. 2006. *Recurrent Neural Networks Are Universal Approximators*, pages 632–640. Springer Berlin Heidelberg.

Shahsavari Baboukani, P., Graversen, C., Alickovic, E., and Østergaard, J. 2020. Estimating conditional transfer entropy in time series using mutual information and nonlinear prediction. *Entropy*, 22(10): 1124.

Shalev, Y., Painsky, A., and Ben-Gal, I. 2022. Neural joint entropy estimation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13.

Sohn, K., Yan, X., and Lee, H. 2015. Learning structured output representationusing deep conditional generative models. *Advances in Neural Information Processing Systems 28 (NIPS 2015)*.

Song, J. and Ermon, S. 2020. Understanding the limitations of variational mutual information estimators.

Sricharan, K., Wei, D., and Hero, A. O. 2013. Ensemble estimators for multivariate entropy estimation. *IEEE Transactions on Information Theory*, 59(7): 4374–4388.

Stokes, P. A. and Purdon, P. L. 2017. A study of problems encountered in granger causality analysis from a neuroscience perspective. *Proceedings of the National Academy of Sciences*, 114(34).

Vejmelka, M. and Paluš, M. 2008. Inferring the directionality of coupling with conditional mutual information. *Physical Review E*, 77(2): 026214.

Wiener, N. 1956. *The theory of Prediction*. Modern Mathematics for the Engineer. McGraw-Hill.

Yin, Y., Zhang, J., and Duan, X. 2020. Information transfer with respect to relative entropy in multi-dimensional complex dynamical systems. *IEEE Access*, 8: 39464–39478.

Zhang, J., Simeone, O., Cvetkovic, Z., Abela, E., and Richardson, M. 2019. Itene: Intrinsic transfer entropy neural estimator.

Zhao, P. and Lai, L. 2020. Analysis of knn information estimators for smooth distributions. *IEEE Transactions on Information Theory*, 66(6): 3798–3826.