Introduction to Transfer Entropy

# Introduction

The book is about *information flow*. We are interested in detecting information flow between two systems, just from *time series*.

Qualitatively, *complex systems* are often described as collections of (generally simple) entities, where the global behaviour is a non-trivial result of the local interactions of the individual elements. They are cellular automata (CA), complex networks, random Boolean networks (RBN), spin systems, and oscillator populations, and flocking behaviour.

# Statistical Preliminaries

Set theory (cardinality, subsets), discrete probabilities (probability spaces), conditional probabilities, independent probabilities, joint probabilities (Bayes’ rule, marginalization), conditional independence.

Time series data and embedding dimensions: time series is ordered values yt, t can be continuous or discrete. Embedding of time series can also be done using time-delay embedding.

Conditional Independence and Markov Processes (Markov process must be *stationary*, i.e., time invariant). Stationarity is important throughout the book!

Vector autoregressive model (VAR(m)), where m is the number of lags. The goal is to understand the *linear* relationships between multiple statistical processes. Estimate matrix A. VAR(1): St+1 = ASt + et+1. Mean and covariance of residual term must be zero.

Moments, mean, (co)variance, Pearson correlation coefficient (measures linear dependence, even if coefficient is 0, that only means there is no linear dependence, but there can certainly be non-linear dependence.

Binomial distribution, sequence of *n* iid 0/1 trials w.p. *p*. Mean is *np*, variance is *npq = np(1-p)*.

Poissoin distribution, used for arrival processes (*stationarity* required, constant arrival intensity *lambda*).

Gaussian distribution, continuous distribution with real line as support. Mean *mu* and sd. *sigma*.

Symmetry and Symmetry Breaking. Concept of a time series having a *fork*. There are two local minima from which you cannot escape easily. You observe the time series for years and think that it is stationary, but suddenly it jumps to the other local minimum and the behavior is completely different (ergodic / non-ergodic).

# Information Theory

## Entropy

Entropy is one of the most alluring and powerful concepts in the history of science and information. chapter. Entropy is the *average uncertainty* in the value of a sample of a variable, equivalent to the *average information required* to determine the value of that sample.

We have relative, or conditional entropy (CE); what is left over when we already know something about a variable. Conversely, mutual information (MI) is essentially how much information is shared by two variables.

Definition of entropy, conditional entropy (the uncertainty left after we have taken into consideration some context.), joint entropy.

## Mutual Information

*Mutual information* (the amount of shared information between X and Y. It is a measure of their statistical dependence. Non-linear form of correlation, always non-negative (its components can be negative). A positive pointwise means “knowing event y increases our expectation of event x occurring”. A negative means that “knowing event y decreases our expectation of event x occurring”, e.g. if we know the animal is a mammal, our expectation that this animal lays eggs decreases. Another example is that an avocado being stoned food, we would expect it to be fruit. However, it is actually a vegetable. Hence, the fact that it is stoned food *misinforms* us about its type.

*Conditional mutual information* is mutual information conditioned on another r.v. Z. e.g. the population of cats and snakes is conditionally independent on mice. Both hunt mice, so if there are mice, there are both more cats and snakes, but they are conditionally independent, since there is only an indirect correlation.

*Multi-information* is Mutual Information extended throughout > 2 dimensions.

*Redundancy:* conditional MI < unconditional MI. If we know Y then we do not get any more information by learning about Z.

*Synergy:* conditional MI > unconditional MI. Conditioning on the other input reveals the synergistic relationship.

## Kullback-Leibler

Similar to Kullback-Leibler divergence of product of marginal distributions (p(x)p(y)) from the joint distribution (p(x,y)). It is the amount of information lost when using q(x) to represent p(x). An example, So, we can ask how much information we get about the variation of temperature from knowing the rainfall.

The cross entropy, G(p : q).

## Entropy of Continuous Processes

Instead of summing, we now integrate over the domain where p(x) > 0. The transition from discrete to continuous is not smooth, and properties can be counter-intuitive, e.g. it can be negative. A linear transformation U = AX changes the differential entropy by a factor + log|A|, so differential entropies may be negative, but is invariant to translations U = X + A does not change anything.

If the mean and variance of a distribution are given, the Shannon entropy is maximal if the distribution is Gaussian. MI of two Gaussians is equal to -1/2log(1-rho^2).

Mutual information has the following interpretations:

1. It tells us if two time sequences share something, a more powerful non-linear measure than correlation.
2. It is the measure of how much information can be transmitted down a noisy channel.
3. Phase transition, e.g. first order or second order transitions. The MI peaks at a second-order phase transition across many systems.

## Numerical Challenges

Calculating entropy is computationally demanding. Numerical estimation of entropy is a very complex topic. There is no best estimator of entropy across all distributions. Estimation of errors and significance is also non-trivial. Naively calculating information from frequency estimates is just that, naive!

Bias is intrinsic to entropy estimation, often underestimating.

### Plug-in / MLE

Not free of bias (underestimating), finite variance. Bias correction by MM, also a maximum bound on the variance. The probability that the error exceeds epsilon decays exponentially with epsilon. As the bias improves, the variance worsens (and vice versa). There is a trade-off between bias and variance in the calculation of entropy. Small datasets are also a problem.

A practical test for checking the robustness of a mutual information estimator is to randomly permute one of the variables. The marginal entropies will not change, but the link between the variables is destroyed, so the MI estimation will be distributed as though those variables had no relationship.

### Kernel Density Estimation

The difficulties of selecting partitions for a direct estimate of the probabilities and subsequently entropies may be circumvented by a statistical technique known as *kernel estimation*. instead of the bins having fixed rigid boundaries we reconsider or adapt the bins each time we consider the PDF for a given sample. Examples are box-kernel estimator, Gaussian Kernel function.

Kernel estimation can measure non-linear relationships and is model-free, though it is sensitive to the parameter choice for resolution r.

### Digamma Entropy Estimator

An alternate way of estimating the differential entropy of a continuous variable from a finite number of samples is the *digamma entropy estimator.* An extension is The *Kozachenko–Leonenko* estimator.

## Estimating Mutual Information

1. The data is already discrete or binned.
2. The data is continuous and needs to be binned.

Calculating mutual information is tricky and needs to be validated case by case. For continuous distributions, we first need to bin the data to determine probability distributions. Good estimate for number of bins is square root of N, the number of data points.

Other approach is to use bins of various sizes. For each marginal, the amount of bins should be the ceil(square root of (N/5))

### The KSG (Kraskov, Stögbauer, and Grassberger) Algorithm

An effective estimator for MI for continuous distributions. The key innovation of the KSG algorithm is getting the numerical errors to partially cancel in the marginal and joint entropy estimates.

### KSG for conditional MI

### Non-stationary case

window for which the probabilities are calculated, the series is stationary. The estimate of mutual information is not meaningful in this case.

# Transfer Entropy