1 Mutual Information

The sample correlation coefficient measures the strength of the linear relationship between two signals. What happens when this relationship is nonlinear? Consider the following example

$$\begin{aligned} x_i &= \sin 2\pi\gamma_i \\ y_i &= \cos 2\pi\gamma_i, \end{aligned} \tag{1}$$

where $\gamma_i \sim U(0,1)$ i.i.d. The correlation between these two signals is displayed in Figure 1 and can be seen to very small, even though knowing x_i gives us a good deal of information about y as can be seen in the right panel. In order to detect nonlinear relationships, as is the case in equation(1), we need a nonlinear statistic. The *mutual information* is a nonlinear statistic that measures the information between two systems. Given simultaneous measurements of systems X and Y, the mutual information measures how much information, on average, we possess about Y given that X has been observed. This is in contrast to the correlation coefficient which tells us how much variance in X can be explained by observing just Y. In order to apply such a measure we must discuss our joint system in the language of information theory.

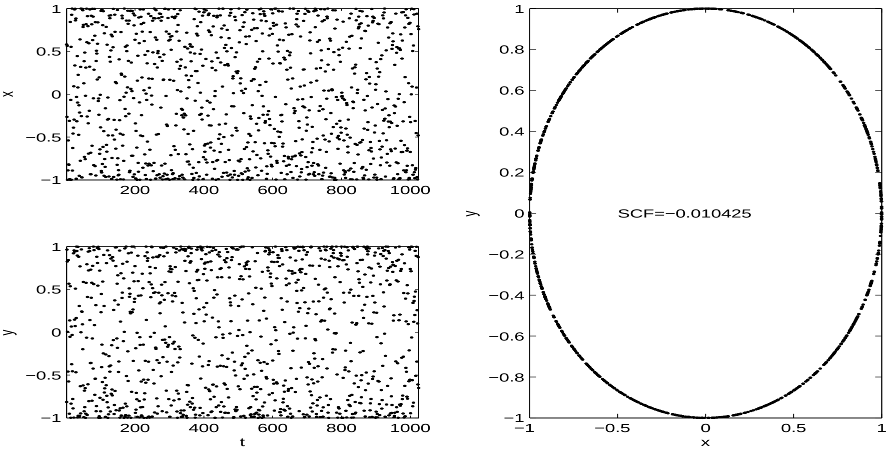

Figure 1: Time series and sample correlation for equation (1).

Let X denote the whole system consisting of a set of possible signals $\{x_i : i = 1, \ldots, B\}$ with associated probabilities $\{P_X(x_i) : i = 1, \ldots, B\}$. For our purposes, the set of possible "signals" will correspond to intervals of the real line, i.e. the bins of a histogram. To transform our time-series into a set of discrete signals it is common to partition the observation space $\mathcal{O}$ into bins.

The average amount of information gained from a measurement that specifies X is defined to be the entropy $H(X)$ of a system where

$$H(X) \quad = \quad -\sum_i P_X(x_i) \log P_X(x_i). \tag{2}$$

If the log is taken to base two, then the unit of H is the bit (binary digit).

Entropy measures the disorder of the system and can be interpreted as the surprise one should feel upon taking a measurement of the system. For a system which is completely determined there is only one outcome and it occurs with probability 1. The entropy is therefore zero. Given n events, maximum entropy is achieved with a uniform distribution, where $p_i = 1/n$ and $H(X) = \log n$.

Given a time series $\{x_i\}_{i=1}^N$ we can estimate the probabilities of events and hence the entropy by binning the data. An example is given in Figure 2. We can see that estimates of the entropy are dependent on the partitioning (binning) of X. Different binnings give different results. In isolation, the entropy lacks rigour for it to be a useful characterising statistic. The mutual information, however, quantifies the relative information between two systems and is less sensitive to the partitioning. In a general coupled system (X, Y) we ask, "Given that X has been measured and found to be x_i, what uncertainty is there in a measurement of Y?". By considering the conditional

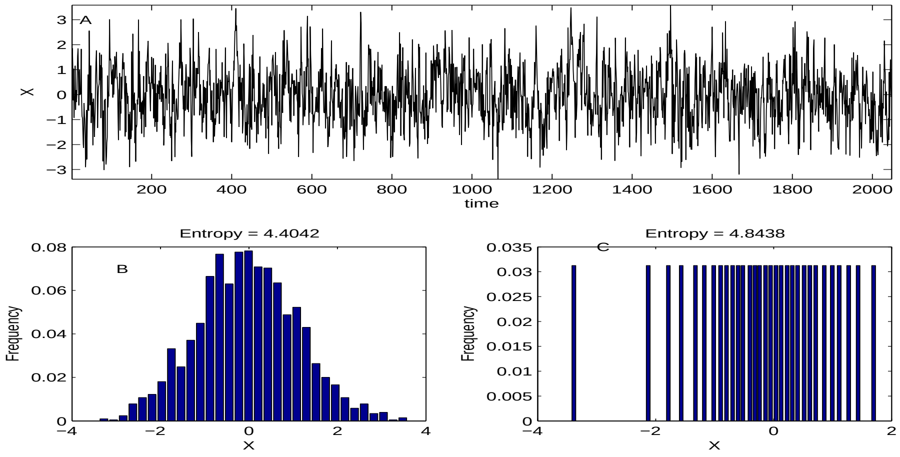

Figure 2: Example of estimating probabilities from data. The `gluttony.dat` time series (Panel A) is put into bins (Panel B and C). Panel B shows a binning with uniform bin sizes, Panel C shows a binning with an equal number of observations.

probabilities $P(A|B) = P(A \cap B)/P(B)$, we have

$$H(Y|x_i) = -\sum_j \frac{P_{XY}(x_i, y_j)}{P_x(x_i)} \log \frac{P_{XY}(x_i, y_j)}{P_X(x_i)}. \tag{3}$$

The average uncertainty in a measurement of Y given x_i is provided by averaging $H(Y|x_i)$ over x_i, giving

$$\begin{aligned} H(Y|X) &= \sum_i P_X(x_i) H(Y|x_i) \\ &= -\sum_i \sum_j P_{XY}(x_i, y_j) \log(\frac{P_{XY}(x_i, y_j)}{P_X(x_i)}) \\ &= H(X,Y) - H(X). \end{aligned} \tag{4}$$

We can interpret $H(Y)$ as the uncertainty of Y in isolation and $H(Y|X)$ is the uncertainty of Y given a measurement of X. So, the average amount by which a measurement of X reduces the uncertainty of Y is the mutual information

$$\begin{aligned} I(Y;X) &= H(Y) - H(Y|X) \\ &= H(Y) + H(X) - H(X,Y) \\ &= I(X;Y), \end{aligned} \tag{5}$$

where $I(X;Y)$ is the mutual information. We can see by the definition that the mutual information is symmetric, i.e. $I(X;Y) = I(Y;X)$.

One often wants to quantify the relationship between two experimental time series, or between the time lagged values of the same time series. The relationship between non-linear systems and their time lagged values may not however be detectable using linear correlation.

The principle difficulty in calculating the mutual information from experimental data is in estimating the joint probability distribution $P(X,Y)$. A straightforward approach is to use a histogram with uniform grid size. For observed data the noise level imposes a lower limit on the grid size, since we do not want fluctuations due to noise to be interpreted as small scale structure. Choosing any single box size has advantages and disadvantages. For a given number of data points, larger boxes have more points, and hence, estimates of the average probability are more accurate, while the estimates of P_{XY} are too flat, underestimating $I(X;Y)$. Smaller boxes account for the changes in P_{XY} over short length scales yet they also allow fluctuations that are due to small sample size to be included in P_{XY}, thus overestimating $I(X;Y)$.

Consider the following example using the data set `gluttony.dat`. We want to compute the mutual information as a function of lag. The X component is the original data set and Y is the lagged values of X. We can partition both data sets into 32 separate bins. We are then interested in the joint probability distribution $P(X, Y)$. We can estimate this by counting the number of points in each bin $B_{i,j}$. Figure 3 shows intensity plots for the joint distribution for various values of the lag. The mutual information, for the time series `gluttony.dat`, as a function of lag is displayed in Figure 4.

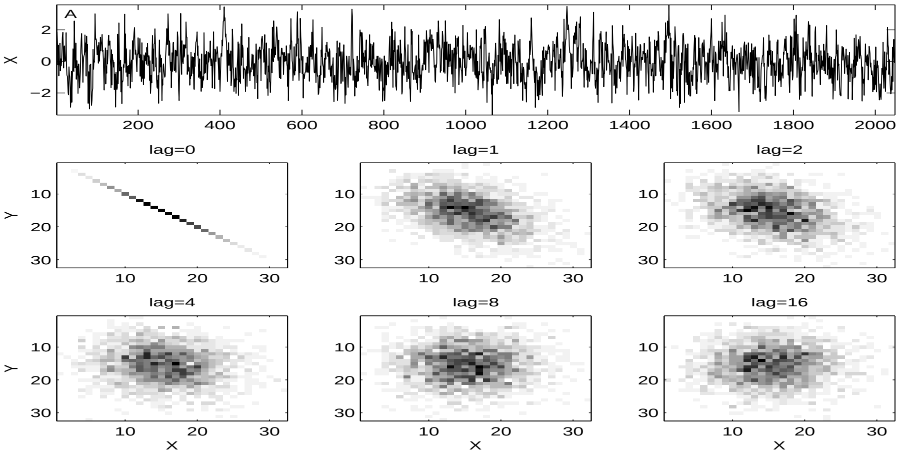

Figure 3: Estimates of the joint probability distribution $P(x_i, x_{i-l})$ for $l = 0, 1, 2, 4, 8, 16$. Darker regions indicate higher frequency.

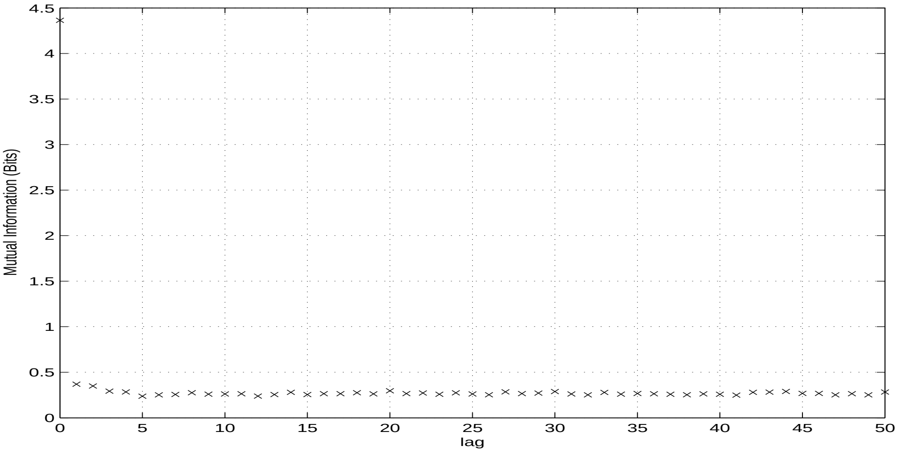

Figure 4: Mutual information as a function of lag for the time series `gluttony.dat`.