2005 International Symposium on Nonlinear
Theory and its Applications (NOLTA2005)
Bruges, Belgium, October 18-21, 2005

A new approach to mutual information between pairs of time series

Andreas Galka[†‡], Tohru Ozaki[‡] and Okito Yamashita[§]

† Institute of Experimental and Applied Physics, University of Kiel, 24098 Kiel, Germany
‡ Institute of Statistical Mathematics (ISM), Minami-Azabu 4-6-7, Tokyo 106-8569, Japan
§ ATR Computational Neuroscience Laboratories, Hikaridai 2-2-2, Kyoto 619-0288, Japan
Email: galka@physik.uni-kiel.de

Abstract—We study the estimation of mutual information between pairs of simultaneous time series, taking into account temporal correlations within the time series. It is shown that an intimate relationship exists between parametric model fitting by Maximum-Likelihood and estimation of mutual information. As a result it becomes possible to detect weak correlations within short spatiotemporal data sets, such as provided by the fMRI technique in neuroscience.

1. Introduction

In many fields of science multivariate time series are obtained from spatially extended dynamical systems, e.g. in hydrodynamics, meteorology, geophysics, biology and medicine. As a particular example we mention neuroscience where spatiotemporal data sets are recorded routinely through well-established modalities like electroencephalography (EEG), magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI).

In order to investigate interactions between subsystems located at different spatial positions, measures for pairwise dependence are needed; linear correlation is a well-known measure for linear dependence, while mutual information (MI) has been introduced as a measure for general dependence in an information-theoretic framework [1]. For gaussian distributions both measures are equivalent, i.e. MI is a function of linear correlation.

In this paper the relationship between MI estimation and parametric modelling of time series is investigated, and a new parametric estimator of MI is derived.

2. The definition of Mutual Information revisited

The definition of MI is based on the probability distributions of two random variables x and y which can assume values out of a set of states; assume that the number of states is finite, say S. Let the index i, $i = 1, \ldots S$, label these states, denote the corresponding values by x_i and y_i and assume that joint and marginal probability distributions $p(x_i, y_j), p(x_i)$ and $p(y_i)$ for the occurrence of these states exist. Then the mutual information $I(x, y)$ between x and y is defined by

$$\begin{aligned} I(x,y) &= \sum_{i=1}^{S}\sum_{j=1}^{S} p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)} \\ &= \Big\langle \log p(x_i,y_j) - \log\big(p(x_i)p(y_j)\big)\Big\rangle_{p(x_i,y_j)} \quad , \end{aligned} \tag{1}$$

where $\langle . \rangle_{p(x_i,y_j)}$ denotes the average over (i, j) with respect to $p(x_i, y_j)$. Note that MI is estimated from distributions, without any reference to time and dynamics; the probability distributions $p(x_i, y_j), p(x_i)$ and $p(y_i)$ need to be estimated from the available data, assuming that this represents independently drawn samples.

If the data is given as a pair of time series x_t and y_t, $t = 1, \ldots, N$, the assumption of independent sampling will typically be invalid since most time series display serial correlations. Consequently we have to regard x_t and y_t as different random variables for each value of t, and Eq. (1) is replaced by

$$\begin{aligned} I(x,y) &= \log p\big((x_1,y_1),\ldots,(x_N,y_N)\big) \\ &\quad - \log\big(p(x_1,\ldots,x_N)p(y_1,\ldots,y_N)\big) \quad . \end{aligned} \tag{2}$$

Now all serial correlations are captured by the corresponding joint distributions. Reinterpreting Eq. (2) from the viewpoint of time series analysis, it can be seen that mutual information can be regarded as a difference between two terms representing log-likelihoods, the first referring to the bivariate time series (x_t, y_t), the second being the sum of the log-likelihoods of the two univariate time series x_t and y_t. Let log-likelihood be denoted by $\mathcal{L}$, then Eq. (2) corresponds to

$$I(x,y) = \mathcal{L}(x,y) - \big(\mathcal{L}(x) + \mathcal{L}(y)\big) \quad . \tag{3}$$

3. Expressing likelihood using predictive modelling

The high-dimensional distributions in Eq. (2) will be very difficult to estimate, therefore we propose to whiten the data by predictive modelling, i.e. using the

expected values as predictors and forming the *residuals* (also called *innovations*):

$$\epsilon_t(x|x) = x_t - \mathcal{E}(x_t|x_{t-1}, x_{t-2}, \ldots) \ , \tag{4}$$

$$\epsilon_t(y|y) = y_t - \mathcal{E}(y_t|y_{t-1}, y_{t-2}, \ldots) \ , \tag{5}$$

$$\big(\epsilon_t(x|x,y), \epsilon_t(y|x,y)\big)^\dagger = \big(x_t, y_t\big)^\dagger - \mathcal{E}\big((x_t,y_t)^\dagger|(x_{t-1},y_{t-1})^\dagger, (x_{t-2},y_{t-2})^\dagger, \ldots\big) \ . \tag{6}$$

According to a theorem from the theory of stochastic dynamical processes (see Theorem 41 in [2]), any continuous-time Markov process with continuous dynamics can be modelled such that the corresponding innovations are I.I.D. gaussian noise. For the predictive models of Eqs. (4)– (6) it can easily be shown that

$$p(x_1, \ldots, x_N) = p\big(\epsilon_1(x|x), \ldots, \epsilon_N(x|x)\big) \ , \tag{7}$$

$$p(y_1, \ldots, y_N) = p\big(\epsilon_1(y|y), \ldots, \epsilon_N(y|y)\big) \ , \tag{8}$$

$$p\big((x_1,y_1), \ldots, (x_N,y_N)\big) = p\Big(\big(\epsilon_1(x|x,y), \epsilon_1(y|x,y)\big), \ldots, \big(\epsilon_N(x|x,y), \epsilon_N(y|x,y)\big)\Big), \tag{9}$$

This fact renders it possible to replace the unknown and intractable joint distributions $p\big((x_1,y_1), \ldots, (x_N,y_N)\big)$, $p(x_1, \ldots, x_N)$ and $p(y_1, \ldots, y_N)$ in Eq. (2) by products of gaussian distributions.

Let the corresponding log-likelihoods for x and y be given by

$$\begin{aligned}\mathcal{L}(x) &= \log p(x_1, \ldots, x_N) \\ &= \log p\big(\epsilon_1(x|x), \ldots, \epsilon_N(x|x)\big) \\ &= -\frac{1}{2}\Big(N \log \sigma^2_{\epsilon(x|x)} + \sum_{t=1}^{N} \frac{\epsilon_t^2(x|x)}{\sigma^2_{\epsilon(x|x)}} + N\log(2\pi)\Big),\end{aligned} \tag{10}$$

by a corresponding expression for y, and for (x,y) by

$$\begin{aligned}&\mathcal{L}(x,y) = \log p\big((x_1,y_1), \ldots, (x_N,y_N)\big) = \\ &\log p\Big(\big(\epsilon_1(x|x,y), \epsilon_1(y|x,y)\big), \ldots, \big(\epsilon_N(x|x,y), \epsilon_N(y|x,y)\big)\Big) \\ &= -\frac{1}{2}\Big(N \log |\mathsf{S}_{\epsilon(x,y|x,y)}| + \sum_{t=1}^{N} \big(\epsilon_t(x|x,y), \epsilon_t(y|x,y)\big) \\ &\quad \times \mathsf{S}^{-1}_{\epsilon(x,y|x,y)} \big(\epsilon_t(x|x,y), \epsilon_t(y|x,y)\big)^\dagger + 2N\log(2\pi)\Big) \ .\end{aligned} \tag{11}$$

Here $\sigma^2_{\epsilon(x|x)}$ denotes the variance of the innovations for the case of a predictive model for x only; and $\mathsf{S}_{\epsilon(x,y|x,y)}$ denotes the covariance matrix of the bivariate innovations for the case of a predictive model for (x,y). The following structure is chosen for $\mathsf{S}_{\epsilon(x,y|x,y)}$:

$$\mathsf{S}_{\epsilon(x,y|x,y)} = \begin{pmatrix} \sigma^2_{\epsilon(x|x,y)} & \mathsf{S}_{12} \\ \mathsf{S}_{21} & \sigma^2_{\epsilon(y|x,y)} \end{pmatrix} \ , \tag{12}$$

where the off-diagonal element is given by

$$\mathsf{S}_{12} = \mathsf{S}_{21} = \sigma_{\epsilon(x|x,y)}\, \sigma_{\epsilon(y|x,y)}\, r\big(\epsilon(x), \epsilon(y)\big) \ ; \tag{13}$$

here $r\big(\epsilon(x), \epsilon(y)\big)$ denotes the normalised coefficient of linear correlation.

4. Estimation of Mutual Information through whitening

If we replace in Eqs. (10), (11) and Eq. (12) the parameters $\sigma_{\epsilon(x|x)}$, $\sigma_{\epsilon(y|y)}$, $\sigma_{\epsilon(x|x,y)}$, $\sigma_{\epsilon(y|x,y)}$ and $r\big(\epsilon(x), \epsilon(y)\big)$ by their appropriate Maximum-Likelihood estimators and insert the results into Eq. (3), we obtain after some transformations

$$\begin{aligned}I(x,y) = -\frac{1}{2}N\Big(&\log\big(1 - r^2(\epsilon(x), \epsilon(y))\big) \\ &+ \big(\log \sigma^2_{\epsilon(x|x,y)} - \log \sigma^2_{\epsilon(x|x)}\big) \\ &+ \big(\log \sigma^2_{\epsilon(y|x,y)} - \log \sigma^2_{\epsilon(y|y)}\big)\Big) \ .\end{aligned} \tag{14}$$

Note that Eq. (2) can be regarded as the likelihood ratio test (LRT) statistic of the null hypothesis of independence of the time series x_t and y_t [3]. By Eq. (14) possible deviations from independence are decomposed into three components, the first describing instantaneous correlations between the innovations of x and y, quantified by $r\big(\epsilon(x), \epsilon(y)\big)$, while the second and the third describe dependence of x on the past of y and vice versa. If knowing the past of y does not improve predictions of x, and vice versa, the mutual information can still be non-zero, as given by the first term on the rhs of Eq. (14); however, since any estimates of the mutual information obtained from actual finite samples will follow a χ^2-distribution (as it is usually the case for any LRT statistics in the null case), it is to be expected that even in this case there will be a small positive bias resulting from the second and third terms on the rhs of Eq. (14). In contrast to this, in the non-null case the estimate of mutual information can be expected to follow a Gaussian distribution [4].

For any particular application a suitable predictive model needs to be formulated and fitted to the data, and the MI estimator, Eq. (14), needs to be reformulated accordingly; we will now show an example.

5. The case of linear autoregressive modelling

Assume that for a given pair of time series x_t and y_t all mutual dependences can be modelled as *instantaneous*, i.e. not involving any time lags, then modelling can be performed by two linear autoregressive (AR)

models

$$\begin{aligned}\epsilon_t(x) &= x_t - \Big(\mu_x + \sum_{\tau=1}^{p} a_\tau\, x_{t-\tau}\Big) \\ \epsilon_t(y) &= y_t - \Big(\mu_y + \sum_{\tau=1}^{p} b_\tau\, y_{t-\tau}\Big) \ ,\end{aligned} \tag{15}$$

where the covariance matrix of the estimated innovations $\big(\epsilon_t(x), \epsilon_t(y)\big)$ is given by Eq. (12); in the standard Maximum-Likelihood model estimation approach the elements of $\mathsf{S}_{\epsilon(x,y|x,y)}$ need to be estimated by numerical optimisation. Alternatively, the pair of time series may be modelled by

$$\begin{aligned}\epsilon_t(x) &= x_t - \Big(\mu_x + \sum_{\tau=1}^{p} a_\tau\, x_{t-\tau}\Big) \\ \epsilon_t(y|x) &= y_t - \Big(\mu_y + \sum_{\tau=1}^{p} b_\tau\, y_{t-\tau} + c_{xy} x_t\Big) \ ,\end{aligned} \tag{16}$$

such that the instantaneous dependence is captured by an additional *coupling term* $c_{xy}x_t$ [5]; then the covariance matrix of $\big(\epsilon_t(x), \epsilon_t(y|x)\big)$ will be diagonal. The log-likelihood of model (16) is given by

$$\begin{aligned}\mathcal{L}(x,y|x) = -\frac{1}{2}\sum_{t=p+1}^{N} \Big(&\log\big|\mathsf{S}_{\epsilon(x,y|x)}\big| \\ &+ \big(\epsilon_t(x), \epsilon_t(y|x)\big)\mathsf{S}^{-1}_{\epsilon(x,y|x)}\big(\epsilon_t(x), \epsilon_t(y|x)\big)^\dagger \\ &+ 2\log(2\pi)\Big) \ ,\end{aligned} \tag{17}$$

By transforming Eq. (16) back into proper autoregressive form (such that no terms depending on time t remain on the right-hand side), the corresponding covariance matrix $\mathsf{S}_{\epsilon(x,y|x)}$ can be obtained:

$$\mathsf{S}_{\epsilon(x,y|x)} = \begin{pmatrix} \sigma^2_{\epsilon(x)} & c_{xy}\sigma^2_{\epsilon(x)} \\ c_{xy}\sigma^2_{\epsilon(x)} & c^2_{xy}\sigma^2_{\epsilon(x)} + \sigma^2_{\epsilon(y|x)} \end{pmatrix} \ . \tag{18}$$

After some further transformations we arrive at

$$\begin{aligned}\mathcal{L}(x,y|x) = &-\frac{1}{2}(N-p)(\log\sigma^2_{\epsilon(x)} + \log\sigma^2_{\epsilon(y|x)}) \\ &- (N-p)\big(1+\log(2\pi)\big) \\ &+ (N-p)\frac{2c_{xy}\,\sigma^2_{\epsilon(x),\epsilon(y|x)} \ - \ c^2_{xy}\,\sigma^2_{\epsilon(x)}}{2\sigma^2_{\epsilon(y|x)}} \ ,\end{aligned} \tag{19}$$

where we have defined $\sigma^2_{\epsilon(x),\epsilon(y|x)} = \mathcal{E}\big(\epsilon_t(x)\epsilon_t(y|x)\big)$.

Note that the log-likelihood of the uncoupled model (as in Eq. (15), but without any dependence between x_t and y_t, i.e. with diagonal $\mathsf{S}_{\epsilon(x,y|x,y)}$) is given by

$$\begin{aligned}\mathcal{L}(x,y) = &-\frac{1}{2}(N-p)(\log\sigma^2_{\epsilon(x)} + \log\sigma^2_{\epsilon(y)}) \\ &- (N-p)\big(1+\log(2\pi)\big) \ .\end{aligned} \tag{20}$$

Now from Eqs. (2) it follows that MI can also be expressed as

$$\begin{aligned}I(x,y) = &\log\big(p(x_1,\ldots,x_N|y_1,\ldots,y_N)\,p(y_1,\ldots,y_N)\big) \\ &- \log\big(p(x_1,\ldots,x_N)\,p(y_1,\ldots,y_N)\big) \ ,\end{aligned} \tag{21}$$

Then from Eqs. (19, (20) and (21) the following estimator of MI is obtained:

$$I(x,y) = (N-p)\frac{2c_{xy}\,\sigma^2_{\epsilon(x),\epsilon(y|x)} \ - \ c^2_{xy}\,\sigma^2_{\epsilon(x)}}{2\sigma^2_{\epsilon(y|x)}} \ . \tag{22}$$

6. A numerical example

As an example for the application of the approach outlined so far we consider a chain of 64 coupled stochastic nonlinear oscillators with nearest-neighbours coupling and periodic boundary conditions, each of them being driven by independent white Gaussian noise; this system is an example of a one-dimensional coupled map lattice [6].

The state of the vth oscillator ($v = 1 \ldots 64$) is given by

$$y^{(v)}_t = \tanh\Big(\sum_{\tau=1}^{p} a^{(v)}_\tau\, y^{(v)}_{t-\tau} + \sum_{v\pm1} b^{(v,v\pm1)}_1\, y^{(v\pm1)}_{t-1}\Big) + \eta^{(v)}_t \ , \tag{23}$$

where the first sum describes a local autoregressive (AR) dynamics and the second sum extends over the nearest neighbours of each oscillator; in a chain with periodic boundary conditions there will be two nearest neighbours. The hyperbolic tangens provides a nonlinear element and prevents the dynamics from diverging. Each oscillator is driven by an individual noise term $\eta^{(v)}_t$; however, two non-neighbouring oscillators (with numbers 14 and 41) share a common driving noise term, thereby representing a situation where two different parts of a system are intrinsically connected. Further details on the definition of this simulation will be published in [7].

For an AR model order of $p = 2$ a time series of length $N = 1024$ points is generated, and measurements of the state of each oscillator are simulated by recording noisy data according to

$$x^{(v)}_t = y^{(v)}_t + n^{(v)}_t \ , \tag{24}$$

where $n^{(v)}_t$ denotes a small Gaussian noise component.

The mutual information matrix for the original data (estimated by a nonparametric estimator, based on histograms [4], left figure) does not reveal the intrinsic connection between these two oscillators, but shows a diffuse pattern of pair dependences, mostly reflecting correlations between neighbours. The same matrix for the innovations instead of the original data (middle figure) shows clearly the connection, since all

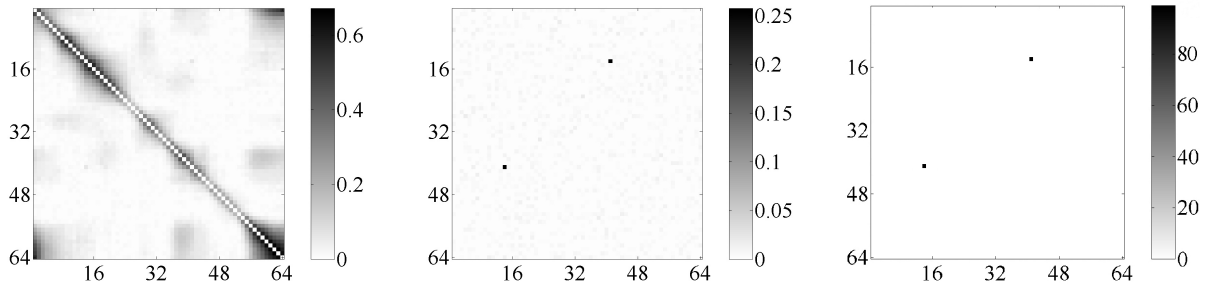

Figure: Histogram estimate of mutual information matrix for simulated data (left) and innovations (middle); parametric estimate for innovations (right). Horizontal and vertical axes represent the chain of oscillators (periodic boundary conditions); entries on the diagonals have been omitted.

other correlations have been removed by the whitening transformation (which in this simulation contains also explicit terms for the instantaneous and delayed interactions between neighbours, details can be found in [7]). The parametric estimate of the mutual information matrix (right figure) according to Eq. (22) reproduces this result with even less spurious dependences between other pairs of oscillators.

We are currently applying this methodology to fMRI data sets recorded during cognition experiments, with the aim of investigating, within a dynamical framework, the connectivity structure of human brain.

Acknowledgments

T.Ozaki gratefully acknowledges support by the Japanese Society for the Promotion of Science (JSPS) through grants KIBAN 13654075 and KIBAN 15500193. A.Galka gratefully acknowledges support by the Deutsche Forschungsgemeinschaft (DFG) through project GA 673/1-1 and by JSPS through fellowship ID No. P 03059.

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, 1948.

[2] P. Protter. *Stochastic Integration and Differential Equations*. Springer-Verlag, Berlin, 1990.

[3] D. R. Brillinger. Second-order moments and mutual information in the analysis of time series. In Y. P. Chaubey, ed., *Recent Advances in Statistical Methods*, pages 64–76. Imperial College Press, London, 2002.

[4] R. Moddemeijer. A statistic to estimate the variance of the histogram based mutual information estimator based on dependent pairs of observations. *Signal Proc.*, **75**:51–63, 1999.

[5] J. Geweke. Inference and causality in economic time series models. In Z. Grilliches and M. Intriligator, eds., *Handbook of Econometrics*, pages 1101–1144. North-Holland, Amsterdam, 1984.

[6] K. Kaneko. Spatiotemporal chaos in one- and two-dimensional coupled map lattices. *Physica D*, **37**:60–82, 1989.

[7] A. Galka, T. Ozaki, J. Bosch Bayard and O. Yamashita. Whitening as a tool for estimating mutual information in spatiotemporal data sets. *J. Stat. Phys.* (submitted), 2005.