Mutual Information: estimation and applications to neuroscience

Roberta Sirovich

Department of Mathematics G. Peano
University of Torino

Statistical Challenges in Neuroscience, 3rd September, 2014

joint work with M. T. Giraudo and L. Sacerdote (UniTo)

Outline

- The setting of the work:
  - ▷ Dependences between variables
  - ▷ Information–theoretic approaches
  - ▷ Mutual Information
- The result:
  - ▷ The estimation of MI
  - ▷ The MI as the entropy of the linkage function
  - ▷ The method

Dependences between variables

- ▷ The question: qiven two (or more) time sequences, are we able to characterize the statistical relationship between the two (or more) vectors?

$$\Downarrow$$

Dependences detection

- ▷ Moreover: what do we mean by “dependence”? Do we have an operational definition? And a measure?

Dependence?

Let us just review briefly the main points of the discussion

- ▷ Interdependence
  - general statistical dependence which only reflects the statistical covariation of signals
  - coherency measures
  - cross–correlograms
  - mutual information
- ▷ Causal dependence
  - distinguish driving and responding elements
  - detect asymmetry in the coupling of subsystems
  - polish up information due to common history and input signals

Mutual Information

- ▷ Benefits
  - sensitive to general (not only linear) dependences, i.e. it is zero only if the two random variables are strictly independent
- ▷ Drawbacks
  - neither contains dynamical nor directional information, i.e. it is symmetric
  - so... causal relationships can be detected only if associated to time delays
  - ...but still does not distinguish information actually exchanged from that due to the response of a common input
  - It is difficult to estimate

A good measure of dependence but still not the best one.

The statistical issue: d=2 to understand

Definition

The mutual information (MI) of a 2-dimensional random vector $X = (X_1, X_2)$ is given by

$$MI(X_1, X_2) = \int_{\mathbf{R}^2} f_{1,2}(x_1, x_2) \log_2 \left[\frac{f_{1,2}(x_1, x_2)}{f_1(x_1) f_2(x_2)} \right] dx_1 dx_2. \qquad (1)$$

Remark:

- If X_1 and X_2 are independent $MI(X_1, X_2) = 0$.
- MI and Entropy are related through the well known equation

$$MI(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2).$$

- MI is the Kullback–Leibler distance between $\mathbb{P}_{1,2}$ and $\mathbb{P}_1 \times \mathbb{P}_2$.

Copulas and Mutual Information

Theorem

Let $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$. The MI (1) of the 2-dimensional random vector $X = (X_1, X_2)$ can be obtained as

$$MI(X_1, X_2) = -H\left(U_1, U_2\right) \tag{2}$$

where F_1 is the distribution function of X_1 and H is the differential entropy.

Eq. (2) reads: MI is minus the entropy of the "copula", i.e. the entropy of the random vector U whose joint distribution is the copula function associated to the original random vector X.

Defs

Estimation of MI

Idea: Use the relationship between MI and Entropy:

- ▷ transform the original sample in a new sample with uniform marginals through $U_1 = F(X_1), U_2 = F(X_2)$;
- ▷ estimate the entropy of the obtained sample.

$$\Downarrow$$

Extension to the $d-$ dimensional case!

Mutual Information: general d

The definition is not unique as it depends on the grouping chosen for the components of the random vector $X = (X_1, \ldots, X_d)$.

Definition

For any n multi-indices $(\alpha^1, \ldots, \alpha^n)$ of dimensions $h_1, \ldots, h_n$ respectively, such that $h_1 + \cdots + h_n = d$ and partitioning the set of indices $\{1, 2, \ldots, d\}$ the following quantities

$$
\begin{aligned}
MI(X_{\alpha^1}, \ldots, X_{\alpha^n}) &= \int_{\mathbf{R}^d} f_{\alpha^1,\ldots,\alpha^n} \log_2 \frac{f_{\alpha^1,\ldots,\alpha^n}}{f_{\alpha^1} \cdots f_{\alpha^n}} \\
&= \int_{\mathbf{R}^d} f_{1,\ldots,d}(x_1, \ldots, x_d) \times \\
&\quad \log_2 \left[\frac{f_{1,\ldots,d}(x_1, \ldots, x_d)}{f_{\alpha^1_1,\ldots,\alpha^1_{h_1}}(x_{\alpha^1_1}, \ldots, x_{\alpha^1_{h_1}}) \cdots f_{\alpha^n_1,\ldots,\alpha^n_{h_n}}(x_{\alpha^n_1}, \ldots, x_{\alpha^n_{h_n}})} \right] dx_1 \ldots dx_d,
\end{aligned}
$$

are all d–dimensional extensions of the bidimensional MI.

Mutual Information and Entropy

- The $d-$dimensional MI can be expressed as a sum of Entropies

$$MI(X_{\alpha^1}, \ldots, X_{\alpha^n}) = H(X_{\alpha^1}) + \cdots + H(X_{\alpha^n}) - H(X_1, \ldots, X_d). \quad (3)$$

- MI is the Kullback–Leibler distance between $\mathbb{P}_{1,\ldots,d}$ and $\mathbb{P}_{\alpha^1} \times \cdots \times \mathbb{P}_{\alpha^n}$.

??

Is it possible again to transform the sample and get the MI as the entropy of the transformed sample?

Copulas and MI: dimension d

- It is not possible to use copula functions to handle multivariate distribution with given marginal distributions of general dimensions.

!!

The only copula compatible with any assigned multidimensional marginal distributions is the independent one.

Defs

Linkage and MI

Theorem

Let $X = (X_1, \ldots, X_d)$ be a d–dimensional random vector. For any n multi-indices $(\alpha^1, \ldots, \alpha^n)$ of dimensions $(h_1, \ldots, h_n)$ respectively, such that $h_1 + \cdots + h_n = d$ and partitioning the set of indices $\{1, 2, \ldots, d\}$, it holds

$$MI(X_{\alpha^1}, \ldots, X_{\alpha^n}) = -H(U_{\alpha^1}, \ldots, U_{\alpha^n}), \tag{4}$$

where $(U_{\alpha^1}, \ldots, U_{\alpha^n}) = (\Psi_{\alpha^1}(X_{\alpha^1}), \ldots, \Psi_{\alpha^n}(X_{\alpha^n}))$.

Proof

Ψ_{α^i}: Linkage function

Definition

The linkage corresponding to the d-dimensional random vector $(X_{\alpha^1}, \ldots, X_{\alpha^n})$ is defined as the joint distribution L of the vector $(U_{\alpha^1} \ldots, U_{\alpha^n})$

$$(U_{\alpha^1_1}, \ldots, U_{\alpha^1_{h_1}}, \ldots, U_{\alpha^n_1}, \ldots, U_{\alpha^n_{h_n}}) = (\Psi_{\alpha^1}(X_{\alpha^1}), \ldots, \Psi_{\alpha^n}(X_{\alpha^n})) \,. \quad (5)$$

where

- $\Psi_{\alpha^i} : \mathbf{R}^{h_i} \to [0, 1]^{h_i}$, $i = 1, \ldots, n$ with
$\Psi_{\alpha^i}(x_{\alpha^i_1}, \ldots, x_{\alpha^i_{h_i}}) =$
$(F_{\alpha^i_1}(x_{\alpha^i_1}), F_{\alpha^i_2|\alpha^i_1}(x_{\alpha^i_2}|x_{\alpha^i_1}), \ldots, F_{\alpha^i_{h_i}|\alpha^i_1,\ldots,\alpha^i_{h_i-1}}(x_{\alpha^i_{h_i}}|x_{\alpha^i_1}, \ldots, x_{\alpha^i_{h_i-1}}))$;
- $(\alpha^1, \ldots, \alpha^n)$ multi-indices of dimensions $(h_1, \ldots, h_n)$ respectively, such that $h_1 + \cdots + h_n = d$ partitioning the set $\{1, 2, \ldots, d\}$;
- $F_{\alpha^i}, i = 1, \ldots, n$: h_i-dimensional c.d.f. of $X_{\alpha^i} = (X_{\alpha^i_1}, \ldots, X_{\alpha^i_{h_i}})$
- $F_{\alpha^1,\ldots,\alpha^n}$: $d-$dimensional joint c.d.f. of $X_{\alpha^1}, \ldots, X_{\alpha^n}$.

The estimation algorithm

- Estimate the conditional c.d.f.'s in eq. (9). Denote these functions as $\tilde{\Psi}_{\alpha^i} = (\tilde{F}_{\alpha^i_1}, \tilde{F}_{\alpha^i_2|\alpha^i_2}, \ldots, \tilde{F}_{\alpha^i_{h_i}|\alpha^i_{h_i-1}})$, for $i = 1, \ldots, n$;
- For $k = 1, \ldots, N$ calculate $U^k = (U^k_{\alpha^1}, \ldots, U^k_{\alpha^n})$, where $U^k_{\alpha^i} = (\tilde{\Psi}_{\alpha^1}(X^k_{\alpha^1}), \ldots, \tilde{\Psi}_{\alpha^n}(X^k_{\alpha^n}))$, for $i = 1, \ldots, n$;
- Estimate the $MI(X_{\alpha^1}, \ldots, X_{\alpha^n})$ as the differential entropy in eq. (4) of the transformed sample $(U^1, \ldots, U^N)$.

For the particular case when $d = 2$ the procedure becomes the following:

- estimate the c.d.f.'s $U_1 = F_1(X_1), U_2 = F_2(X_2)$. Denote the estimated functions as $(\tilde{F}_1, \tilde{F}_2)$;
- calculate $U^k = (\tilde{F}_1(X^k_1), \tilde{F}_2(X^k_2))$, for $k = 1, \ldots, N$;
- estimate $MI(X_1, X_2)$ as the differential entropy in eq. (2) of the transformed sample $(U^1, \ldots, U^N)$.

Algorithm details

- Use the kernel method to estimate the the linkage functions
- Use the nearest-neighbor method to estimate the differential entropy:

$$\widehat{H} = \frac{d}{N}\sum_{j=1}^{N}\log_2\left(\lambda_j\right) + \log_2\left[\frac{S_d(N-1)}{d}\right] + \frac{\gamma}{\ln(2)} \qquad (6)$$

where $\gamma = -\int_0^\infty e^{-v}\ln v dv \cong 0.5772156649$ is the Euler-Mascheroni constant, λ_j is the Euclidean distance of each sample point to its nearest neighbor and $S_d = \frac{d\pi^{r/2}}{\Gamma(\frac{d}{2}+1)}$ with Γ the gamma function is the area of a unit d-dimensional spherical surface (for example $S_1 = 2,\ S_2 = 2\pi,\ S_3 = 4\pi, \dots$). K-L

Results: Gaussian bivariate vector

Comparison between the proposed, the KSG and plain entropy methods. KSG plain

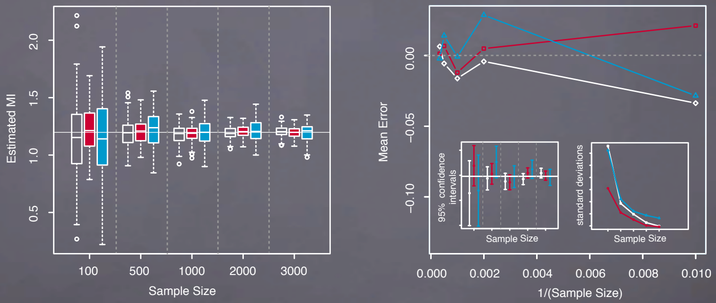

Figure : Standard Gaussian vector with $\rho = 0.9$. Here $MI(X_1, X_2) = 1.1980$ *bit*. Color map: black and white for the estimator we propose, red for KSG and blue for plain entropy.

Results: assigned bivariate distribution

X_1,X_2 have joint c.d.f.

$$F_{1,2}(x_1, x_2) = \begin{cases} \frac{(x_1+1)(e_2^x-1)}{x_1+2e_2^x-1} & (x_1, x_2) \in [-1, 1] \times [0, \infty] \\ 1 - e^{-x_2} & (x_1, x_2) \in (1, \infty] \times [0, \infty] \end{cases}$$

and marginal Uniform on $[-1, 1]$ and Exponential with $\mathbb{E}(X_2) = 1$.

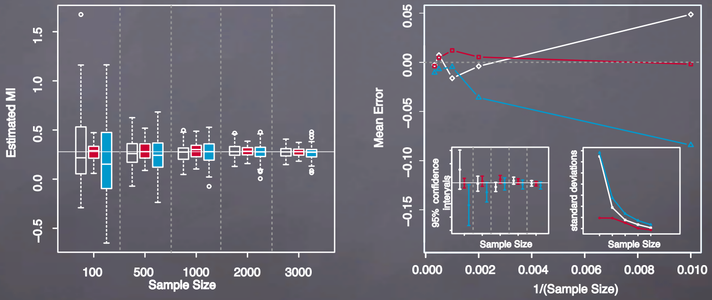

Figure : Color map: black and white for the proposed estimator, red for KSG and blue for plain entropy.

Results: Three dimensional vectors

$X = (X_1, X_2, X_3)$ Gaussian random vector with standard normal components and covariance matrix $\rho_{X_1,X_2} = \rho_{X_2,X_3} = \rho_{X_1,X_3} = 0.9$.

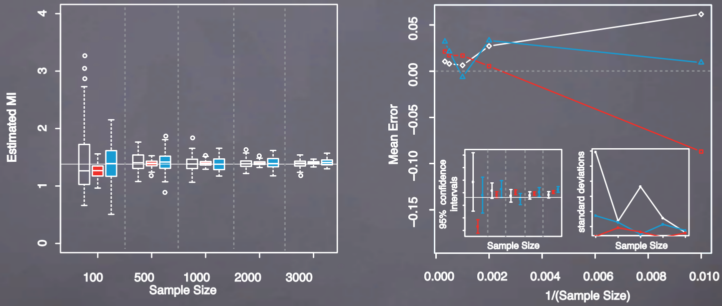

Figure : Color map: black and white for the proposed estimator, blue for plain entropy.

Results: Four dimensional vectors

Multivariate Gaussian random vector, with multi–indices to group the components $\alpha^1 = (1, 2)$ and $\alpha^2 = (3, 4)$.

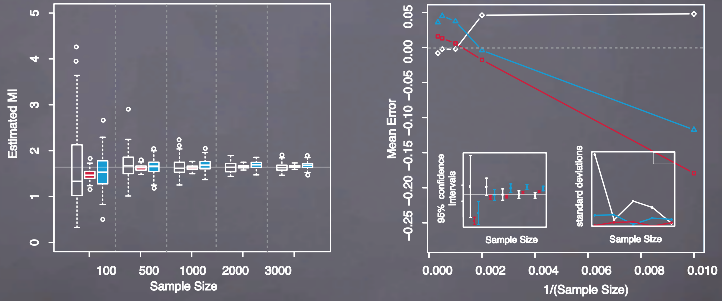

Figure : Color map: black and white for the proposed estimator, blue for plain entropy.

References

[1] Giraudo, M.T., Sacerdote, L. and Sirovich, R. (2013) Non–parametric Estimation of Mutual Information through the Entropy of the Linkage. Entropy 15(12), 5154-5177.

[2] Li, H., Scarsini, M. and Shaked, M. (1996) Linkages: a tool for the construction of multivariate distributions with given nonoverlapping multivariate marginals. *J. Multivariate Anal.*, *56*, 20–41.

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

[4] Kozachenko, L.F. and Leonenko, N.N. (1987) A statistical estimate for the entropy of a random vector. *Problemy Peredachi Informatsii*, *23*, 9–16.

Algo R

Copula function

Definition

A two-dimensional copula is a function $C : [0,1]^2 \to [0,1]$ with the following properties:

1. $C(u;0) = C(0;v) = 0$ and $C(u;1) = u, C(1;v) = v$ for every $u, v \in [0;1]$;
2. C is 2-increasing, i.e. for every $u_1, u_2, v_1, v_2 \in [0;1]$ such that $u_1 \leq u_2, v_1 \leq v_2$,

$$C(u_1, v_1) + C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) \geq 0$$

Remark:

- ▷ A copula function is a 2-dimensional joint distribution.

(Sklar's) Theorem

Theorem

Let F_1 and F_2 be two univariate distributions. It comes that $C(F_1(x_1), F_2(x_2))$ defines a bivariate probability distribution with margins F_1 and F_2.

Theorem

Let $F_{1,2}$ be a two-dimensional distribution function with margins F_1 and F_2. Then $F_{1,2}$ has a copula representation:

$$F_{1,2}(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

The copula C is unique if the margins are continuous.

back

Copulas: general d

Definition

A $d-$dimensional copula (or $d-$copula) is a function $C : [0,1]^d \to [0,1]$ with the following properties:

1. for every $u = (u_1, \dots, u_d) \in [0,1]^d$, $C(u) = 0$ if at least one coordinate is null and $C(u) = u_k$ if all coordinates are 1 except u_k;
2. for every $a = (a_1, \dots, a_d)$ and $b = (b_1, \dots, b_d) \in [0,1]^d$ such that $a \leq b$, $V_C([a,b]) \geq 0$.

Here V_C is the so called $C-$volume of $[a,b]$, i. e. the $n-$th order difference of C on $[a,b]$.

$$V_C([a,b]) = \Delta_{a_d}^{b_d} \Delta_{a_{d-1}}^{b_{d-1}} \dots \Delta_{a_1}^{b_1} C(u), \tag{7}$$

where

$\Delta_{a_k}^{b_k} C(u) = C(u_1, \dots, u_{k-1}, b_k, u_{k+1}, \dots, h_d) - C(u_1, \dots, u_{k-1}, a_k, u_{k+1}, \dots, h_d)$.

Sklar's Theorem: dimension d

Theorem

For any $d-$dimensional c.d.f. $F_{1,\dots,d}$ of the random vector $X = (X_1, \dots, X_d)$ there exists a $d-$copula C such that for all $x = (x_1, \dots, x_d) \in \mathbf{R}^d$

$$F_{1,\dots,d}(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)), \tag{8}$$

where F_i are the univariate margins. If the margins are continuous, then the copula C is uniquely determined. Otherwise, C is uniquely determined over $RanF_1 \times \cdots \times RanF_d$, where $RanF_i$ is the range of the function F_i.

Conversely, if C is a copula and F_i, $i = 1, \dots, d$ are one-dimensional distribution functions, then the function $F_{1,\dots,d}(x_1, \dots, x_d)$ defined in (8) is a d-dimensional distribution function with margins F_i, $i = 1, \dots, d$.

back

Linkage and MI (II)

Proof: Consider the following change of variables:

$$\left\{ \begin{array}{lcl} U_{\alpha_1^1} & = & F_{\alpha_1^1}(X_{\alpha_1^1}) \\ U_{\alpha_2^1} & = & F_{\alpha_2^1|\alpha_1^1}(X_{\alpha_2^1}|X_{\alpha_1^1}) \\ & \vdots & \\ U_{\alpha_{h_1}^1} & = & F_{\alpha_{h_1}^1|\alpha_1^1,\alpha_2^1,\ldots,\alpha_{h_1-1}^1}(X_{\alpha_{h_1}^1}|X_{\alpha_1^1},X_{\alpha_2^1},\ldots,X_{\alpha_{h_1-1}^1}) \\ U_{\alpha_1^2} & = & F_{\alpha_1^2}(X_{\alpha_1^2}) \\ & \vdots & \\ U_{\alpha_1^n} & = & F_{\alpha_1^n}(X_{\alpha_1^n}) \\ & \vdots & \\ U_{\alpha_{h_n}^n} & = & F_{\alpha_{h_n}^n|\alpha_1^n,\ldots,\alpha_{h_n-1}^n}(X_{\alpha_{h_n}^n}|X_{\alpha_{h_1}^n},\ldots,X_{\alpha_{h_n-1}^n}). \end{array} \right. \tag{9}$$

back