INVESTIGATING BIAS IN NON-PARAMETRIC MUTUAL INFORMATION ESTIMATION

Jie Zhu[⋆†§], *Jean-Jacques Bellanger*[⋆†], *Huazhong Shu*[§‡], *Régine Le Bouquin Jeannès*[⋆†§]

[⋆] INSERM, U 1099, Rennes, F-35000, France
[†] Université de Rennes 1, LTSI, F-35000, France
[§] Centre de Recherche en Information Biomédicale sino-français (CRIBs), Rennes, France
[‡] LIST, School of Computer Science and Engineering, Southeast University, Nanjing, China

ABSTRACT

In this paper, our aim is to investigate the control of bias accumulation when estimating mutual information from nearest neighbors non-parametric approach with continuously distributed random data. Using a multidimensional Taylor series expansion, a general relationship between the estimation bias and neighborhood size for plug-in entropy estimator is established without any assumption on the data for two different norms. When applied with the maximum norm, our theoretical analysis explains experimental simulation tests drawn in existing literature. In the experiments, two different strategies are tested and compared to estimate mutual information on independent and dependent simulated signals.

Index Terms— Entropy estimation, bias reduction, mutual information, independence test

1. INTRODUCTION

Mutual Information (MI) is a widely used independence measurement, which has received particular attention during the past few decades. Compared to dependence characterization based on linear or non-linear correlation, MI is a more general dependence measure and its estimation is of great importance when testing the independence between distinct data sources [1, 2, 3, 4]. However, it remains a tough task while carried out on finite sample length signals, particularly in the field of neuroscience, where getting large amounts of stationary data is an issue. More precisely, let (X, Y) be a pair of multidimensional random variables with a continuous distribution specified by a joint probability density $p_{X,Y}$ with marginal densities p_X and p_Y. The joint and marginal entropies, namely $\mathcal{H}(X, Y)$, $\mathcal{H}(X)$ and $\mathcal{H}(Y)$, respectively linked to (X, Y), X and Y, are defined as $\mathcal{H}(X, Y) = -\mathrm{E}\left[\log p_{X,Y}(X, Y)\right]$, $\mathcal{H}(X) = -\mathrm{E}\left[\log p_X(X)\right]$ and $\mathcal{H}(Y) = -\mathrm{E}\left[\log p_Y(Y)\right]$. Mutual information between X and Y is then defined as [5]

$$\mathcal{I}(X, Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X, Y). \tag{1}$$

According to (1), MI estimation from n independent occurrences $z_i = (x_i, y_i)$ of $Z = (X, Y)$ could be simply obtained by estimating three individual entropies separately and then summing them. The individual entropies can be estimated by the following Kozachenko-Leonenko entropy estimator

$$\widehat{\mathcal{H}(U)} = \psi(n) - \psi(k) + \frac{1}{n}\sum_{i=1}^{n} \log v_i, \tag{2}$$

where $U \in \{X, Y, (X, Y)\}$, $\psi(\cdot)$ stands for the digamma function, n denotes the signal length, k the number of neighbors and v_i is the volume of the ball $\{u : \|u - u_i\| < \mathcal{R}_U(u_i)\}$ where the radius $\mathcal{R}_U(u_i)$ is equal to the distance value $d_U(u_i)$ between the ith data sample u_i and its kth nearest neighbor for a given norm, for instance Euclidean or maximum norm. It is a natural choice [6] to impose the same number k of neighbors for the estimation of $\mathcal{H}(X)$, $\mathcal{H}(Y)$ and $\mathcal{H}(X, Y)$. However, a recurrent question arises: is it possible to adapt the values of k to cancel out the bias errors in individual estimations to avoid adverse accumulations of errors when using algebraic summation of the 3 entropy estimations in (1)? To deal with this question, Kraskov *et al.* [6] proposed to use a common neighborhood size $\mathcal{R}_U(u_i)$ for both joint and marginal spaces, when selecting nearest neighbors. This strategy consisted in fixing the number of neighbors in the joint space $\mathcal{S}_Z$ $[Z = (X, Y)]$, then using the resulting distance $\mathcal{R}_Z(z_i)$ as the neighborhood radius value for both $\mathcal{S}_X$ and $\mathcal{S}_Y$. In [6], through numerical simulations, the effectiveness of this strategy is claimed compared to the case where the same number of neighbors is imposed when estimating the 3 individual entropies. This strategy has been widely used since and also extended to the calculation of other information theory functionals, such as divergence [7] or partial mutual information [8]. The interesting conjecture proposed from numerical results in [6] leads to write:

$$\mathrm{E}\left[\widehat{\mathcal{I}(X, Y)}_K\right] = 0, \quad \text{iif } \mathcal{I}(X, Y) = 0, \tag{3}$$

where $\widehat{\mathcal{I}(X, Y)}_K$ is the MI estimated with the strategy proposed in [6].

In the present work, we propose theoretical arguments to justify bias cancellation observed experimentally in [6] when using the strategy mentioned above with the maximum norm and extend these theoretical developments to the Euclidean norm.

$$\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} \approx p_X(x) + \left[\frac{\partial p_X(x)}{\partial x}\right]^T \frac{1}{v(x)} \int_{\mathcal{L}(x)} (y-x)\mathrm{d}y + \frac{1}{2v(x)} \int_{\mathcal{L}(x)} (y-x)^T \left[\frac{\partial^2 p_X(x)}{\partial x^2}\right] (y-x)\mathrm{d}y \tag{10}$$

$$\begin{aligned} \frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} &\approx p_X(x) + \frac{1}{2v(x)} \int_{\mathcal{L}(x)} (y-x)^T \left[\frac{\partial^2 p_X(x)}{\partial x^2}\right] (y-x)\mathrm{d}y \\ &= p_X(x) + \frac{1}{2v(x)} \mathrm{tr}\left\{\left[\int_{\mathcal{L}(x)} (y-x)(y-x)^T \mathrm{d}y\right]\left[\frac{\partial^2 p_X(x)}{\partial x^2}\right]\right\} \end{aligned} \tag{11}$$

2. METHODS AND MATERIALS

2.1. New bias expression for the plug-in entropy estimator

Let us consider a random variable X which takes its values in $\mathbb{R}^{d_X}$. If for any x in $\mathbb{R}^{d_X}$, $\mathcal{L}(x)$ stands for a small region around x, we introduce the Lebesgue measure (volume) $v(x) = \int_{\mathcal{L}(x)} \mathrm{d}z$ of $\mathcal{L}(x)$ and the probability density function $p_X(x)$ attached to the distribution probability of X. In most existing non-parametric density estimation algorithms, including either KDE (Kernel Density Estimation) or kNN (k-Nearest Neighbor), $p_X(x)$ is estimated as

$$\widehat{p_X(x)} = \frac{\widehat{P\left[X \in \mathcal{L}(x)\right]}}{v(x)} = \frac{\widehat{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}}{v(x)}, \tag{4}$$

where $\widehat{P\left[X \in \mathcal{L}(x)\right]}$ corresponds to an estimation of the probability that X belongs to the volume $v(x)$.

In (4), the estimation $\widehat{\log p_X(x)}$ of $\log p_X(x)$ is currently built as follows

$$\begin{aligned} \widehat{\log p_X(x)} &= \log \widehat{p_X(x)} \\ &= \log \frac{\widehat{P\left[X \in \mathcal{L}(x)\right]}}{v(x)} \\ &= \log\left[\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} + \varepsilon\right], \end{aligned} \tag{5}$$

where the random estimation error ε given by

$$\varepsilon = \frac{\widehat{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}}{v(x)} - \frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)}. \tag{6}$$

Note that, using Jensen's inequality, $\mathrm{E}\left(\widehat{\log p_X(x)}\right) = \mathrm{E}\left(\log \widehat{p_X(x)}\right) \leq \log \mathrm{E}\left(\widehat{p_X(x)}\right)$. So, if we assume that $\widehat{p_X}(x)$ is unbiased, it leads to $\mathrm{E}\left(\log \widehat{p_X(x)}\right) \leq \log\left(p_X(x)\right)$. This last inequality implies that $\widehat{\log p_X(x)}$ in (5) is a biased estimation of $\log p_X(x)$. In this study, we focus on this source of bias, and so we assume that $\widehat{P\left[X \in \mathcal{L}(x)\right]}$ is unbiased and ϵ is zero mean (these last assumptions are realistic, at least approximately).

From observations X_i (random variables issued from P_X), the corresponding differential entropy $\mathcal{H}(X)$ can be estimated as

$$\widehat{\mathcal{H}(X)} = -\frac{1}{n}\sum_{i=1}^{n} \widehat{\log p_X(X_i)}, \tag{7}$$

where n is the number of observed occurrences of X. Then, when $\|y-x\|$ is small, a Taylor approximation around x leads to approximate the probability density $p_X(y)$ by

$$\begin{aligned} p_X(y) \approx\ & p_X(x) + \left[\frac{\partial p_X(x)}{\partial x}\right]^T (y-x) \\ & + \frac{1}{2}(y-x)^T \left[\frac{\partial^2 p_X(x)}{\partial x^2}\right] (y-x), \end{aligned} \tag{8}$$

where the superscript T stands for matrix transposition. We analyze the bias of $\widehat{\mathcal{H}(X)}$ writing

$$\begin{aligned} \widehat{\mathcal{H}(X)} &= -\frac{1}{n}\sum_{i=1}^{n} \log \widehat{p_X(X_i)} \\ &= -\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{\int_{\mathcal{L}(X_i)} p_X(y)\mathrm{d}y}{v(X_i)} + \varepsilon_i\right]. \end{aligned} \tag{9}$$

Integrating (8) on both sides and dividing by $v(x)$, we get (10).

As $\mathcal{L}(x)$ admits x as a center of symmetry for any chosen norm, then $\int_{\mathcal{L}(x)}(y-x)\mathrm{d}y = 0$, and the first order term on the right hand side of (10) is zero. According to the fact that $\mathrm{tr}\left(ABC\right) = \mathrm{tr}\left(CAB\right)$, where $\mathrm{tr}\left(\cdot\right)$ stands for the trace operator, (10) is transformed into (11) (note that $\int_{\mathcal{L}(x)} (y-x)\,(y-x)^T \mathrm{d}y$ is a diagonal matrix).

Finally, the estimator $\widehat{\log p_X(x)}$ of $\log p_X(x)$ can be approximated by (12), where the term $\left[\frac{1}{p_X(x)} \cdot \varepsilon\right]$ is zero mean.

According to the Taylor expansion of $\log(\cdot)$ function, the bias $\mathcal{B}_X$ in $\widehat{\mathcal{H}(X)}$ is approximated by the second term in the right hand side of (12) and could be used as a correcting term if it was possible to evaluate it. To build the ball $\mathcal{L}(x) = \{y : \|y-x\| \leq \mathcal{R}(x)\}$, we retain two norms, the Euclidean norm and the maximum norm resulting respectively in a standard ball and in a d_X dimensional cube. The value $\mathcal{R}(x)$ fixes respectively the radius of the ball or the half of the edge length of the cube.

$$\log\left[\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)}+\varepsilon\right] \approx \log\left(p_X(x)+\frac{1}{2v(x)}\mathrm{tr}\left\{\left[\int_{\mathcal{L}(x)}(y-x)(y-x)^T\mathrm{d}y\right]\left[\frac{\partial^2 p_X(x)}{\partial x^2}\right]\right\}+\varepsilon\right)$$
$$\approx \log p_X(x)+\underbrace{\frac{1}{p_X(x)}\frac{1}{2v(x)}\mathrm{tr}\left\{\left[\int_{\mathcal{L}(x)}(y-x)(y-x)^T\mathrm{d}y\right]\left[\frac{\partial^2 p_X(x)}{\partial x^2}\right]\right\}}_{\approx \mathcal{B}_X}+\frac{1}{p_X(x)}\varepsilon \tag{12}$$

$$\frac{1}{p_Z(z_i)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_Z(z_i)}{\partial z^2}\right]=\frac{1}{p_X(x_i)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_X(x_i)}{\partial x^2}\right]+\frac{1}{p_Y(y_i)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_Y(y_i)}{\partial y^2}\right] \tag{15}$$

After calculation, using the Euclidean norm, we get

$$\mathcal{B}_X(x)\approx\frac{\mathcal{R}^2(x)}{2(d_X+2)}\cdot\frac{1}{p_X(x)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_X(x)}{\partial x^2}\right]. \tag{13}$$

Similarly, using the maximum norm distance, we get

$$\mathcal{B}_X(x)\approx\frac{\mathcal{R}^2(x)}{6}\cdot\frac{1}{p_X(x)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_X(x)}{\partial x^2}\right]. \tag{14}$$

2.2. Bias reduction of MI estimator based on the new bias expression

Considering the ith data point, if the signals X and Y are independent, i.e., $p_Z(z_i)=p_X(x_i)p_Y(y_i)$, with $Z=(X,Y)$, we obtain (15).

Now, if we focus on testing an independence hypothesis between X and Y, in order to cancel out the bias, we solve

$$\mathcal{B}_X(x_i)+\mathcal{B}_Y(y_i)-\mathcal{B}_Z(z_i)=0, \tag{16}$$

with respect to $\mathcal{R}(x_i)$ and $\mathcal{R}(y_i)$. With the Euclidean norm, it yields to

$$\mathcal{R}(x_i)=\sqrt{\frac{d_X+2}{d_Z+2}}\cdot\mathcal{R}(z_i)\ \text{ and }\ \mathcal{R}(y_i)=\sqrt{\frac{d_Y+2}{d_Z+2}}\cdot\mathcal{R}(z_i), \tag{17}$$

where $\mathcal{R}(x_i)$, $\mathcal{R}(y_i)$ and $\mathcal{R}(z_i)$ are the distances used for the estimation of $\widehat{p_X(x_i)}$, $\widehat{p_Y(y_i)}$ and $\widehat{p_Z(z_i)}$ at the ith point, d_X, d_Y and d_Z are the dimensions of the signals X, Y and Z respectively. Similarly, using the maximum norm, we obtain

$$\mathcal{R}(x_i)=\mathcal{R}(z_i)\ \text{ and }\ \mathcal{R}(y_i)=\mathcal{R}(z_i). \tag{18}$$

Until now, no particular form of density estimator was specified in our bias analysis. In [6], the Kozachenko-Leonenko estimator [9] is used without calculating the densities for each sample point. However, since $\psi(n)\approx\log(n)$ for large n, (2) is equivalent to

$$\widehat{\mathcal{H}(U)}=-\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{\psi(k)}}{nv_i}\right). \tag{19}$$

Therefore, using (5), (7) and (19), we consider the following density estimator

$$\widehat{p_U(u_i)}=\frac{e^{\psi(k)}}{nv_i}. \tag{20}$$

So, (2) can still be considered as an estimator with the same structure as in (9) and explained under the framework of our bias analysis.

Finally, (18) and (20) formally confirm (as suggested but not proved in [6]) that, if X and Y are independent, using the maximum norm and constraining the values $\mathcal{R}(x_i)$ and $\mathcal{R}(y_i)$ to be equal to $\mathcal{R}(z_i)$ allow to decrease the bias $\widehat{\mathcal{I}(X,Y)}-\mathcal{I}(X,Y)$. (17) extends this result when the Euclidean norm is used for the 3 individual spaces. Let us mention that (15) no longer holds if signals X and Y are not independent. In this case only a part of the bias is a priori expected to be cancelled out.

To conclude, in the case of independence between X and Y, MI is estimated by

$$\widehat{\mathcal{I}(X,Y)}=-\frac{1}{n}\sum_{i=1}^{n}\left[\log\widehat{p_X(x_i)}+\log\widehat{p_Y(y_i)}-\log\widehat{p_Z(z_i)}\right] \tag{21}$$

with an approximately zero bias by fixing $\mathcal{R}(z_i)$ and properly defining $\mathcal{R}(x_i)$ and $\mathcal{R}(y_i)$ using (17) or (18), where $\widehat{p_X(x_i)}$, $\widehat{p_Y(y_i)}$ and $\widehat{p_Z(z_i)}$ are estimated using (20).

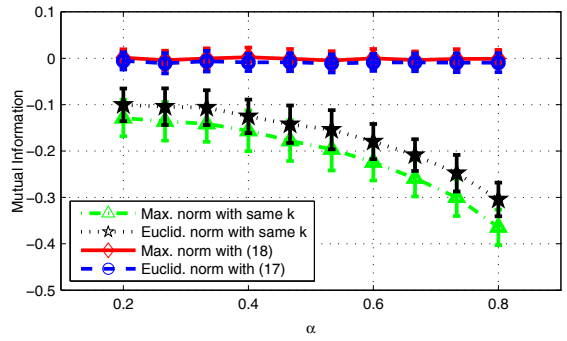
3. SIMULATION RESULTS

To validate our analysis, we generated two independent d-dimensional signals X and Y, both of them following a zero mean Gaussian distribution $\mathcal{N}(0,\mathcal{C})$, where $\mathcal{C}$ is a Toeplitz matrix with first line $[1,\alpha,\ldots,\alpha^{d-1}]$. Clearly, whatever the value of $\alpha\in[0,1[$, the theoretical value of the mutual information $\mathcal{I}(X,Y)$ is zero. For our simulations, we used sequences of n independent samples $(X_i,Y_i)\,,i=1,\ldots,n$ from the distribution of (X,Y).

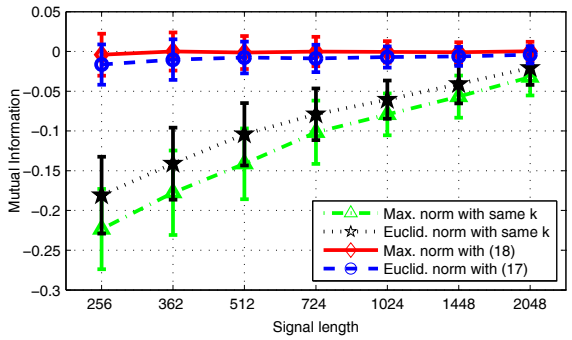
Additionally, in order to briefly investigate the effect of non-independence on the bias of MI estimation when applying the different strategies, we replaced the independent pair (X,Y) by a dependent one, (X,Y_1), where the first coordinate X is the same as previously and Y is replaced by

$$Y_1=\cos\theta\cdot X+\sin\theta\cdot Y. \tag{22}$$

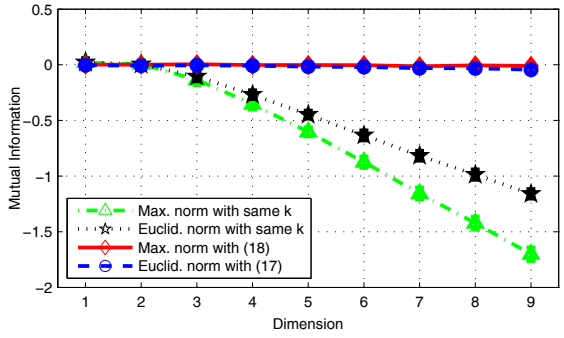
The parameter θ, $\theta\in\left[0,\frac{\pi}{2}\right]$, allows to tune the dependence between X and Y_1, since it modifies the cross covariance ma-



(a) Mutual information $\widehat{\mathcal{I}(X,Y)}$ (in nats) estimated with varying α, $d = 3, n = 512$.



(b) Mutual information $\widehat{\mathcal{I}(X,Y)}$ (in nats) estimated with different signals lengths, $\alpha = 0.3, d = 3$.



(c) Mutual information $\widehat{\mathcal{I}(X,Y)}$ (in nats) estimated with varying dimensions, $\alpha = 0.3, n = 512$.

Fig. 1. Mutual information $\widehat{\mathcal{I}(X,Y)}$ estimation for independent signals using different strategies with 100 trials.

trix $\mathcal{C}_{X,Y_1}$ ($\mathcal{C}_{X,Y_1} = \cos\theta \cdot \mathcal{C}$) whereas the marginal covariance matrices $\mathcal{C}_X$ and $\mathcal{C}_{Y_1}$ remain unchanged ($\mathcal{C}_X = \mathcal{C}_{Y_1} = \mathcal{C}$). Note that, for $\theta = \frac{\pi}{2}$, X and Y_1 are independent. The theoretical value of $\mathcal{I}(X,Y_1)$ is equal to $-d\log(\sin\theta)$.

We tested the 2 different strategies to estimate mutual information either with the maximum norm or the Euclidean

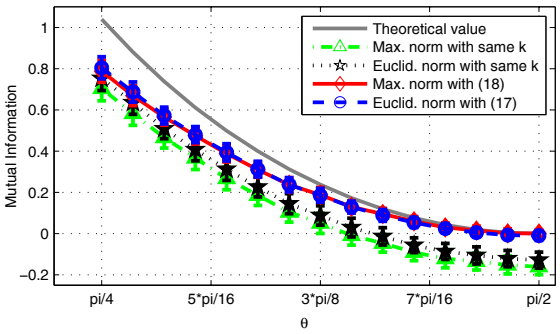


Fig. 2. Mutual information $\widehat{\mathcal{I}(X,Y_1)}$ (in nats) estimation using different strategies with varying θ, $\alpha = 0.4$, $d = 3$, $n = 512$, 100 trials.

norm: (i) we imposed the same number of neighbors k for the 3 individual entropies, (ii) we determined $\mathcal{R}(z_i)$ from k, and then derived $\mathcal{R}(x_i)$ and $\mathcal{R}(y_i)$ using (17) or (18). Throughout the experimentation, k was fixed to 6 and the statistical mean and variance of the different estimators were estimated by an averaging on 100 trials.

Fig. 1 displays the performance of both approaches in the independence case with the two norms (maximum and Euclidean norms). For the estimators using the same k, the performance drastically falls with larger α (Fig. 1(a)), shorter signal length (Fig. 1(b)) and higher dimension (Fig. 1(c)). The estimators with chosen neighborhood size clearly outperform the former significantly whatever the norm, in terms of estimation bias and standard deviation. In other words, the new justified strategy provides reliable mutual information values for independence test, even with short signal lengths or high dimensional signals. As for dependent signals, (model (22)), results are displayed in Fig. 2. Here again, the second strategy is preferred: whatever the norm, mutual information is properly estimated, the experimental values being very close to the theoretical one. Of course, this approach provides the best performance when the dependence between the signals decreases (θ close to $\frac{\pi}{2}$).

4. CONCLUSION

In this paper, we investigated the difficult issue of bias reduction on mutual information estimation. To this end, we established a relation between the systematic bias and the distance parameter for plug-in entropy estimator. Experimental results allowed us to assess the performance of the novel strategy using either Euclidean or maximum norm to get a more accurate estimation of mutual information for independent signals. A preliminary study also reveals its interest in the dependence case and will be further investigated.

5. REFERENCES

[1] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[2] C. J. Cellucci, A. M. Albano, and P. E. Rapp, "Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms," *Physical Review E*, vol. 71, no. 6, pp. 066208, 2005.

[3] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, no. 2, pp. 026209, 2007.

[4] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, no. 1, pp. 1–46, 2007.

[5] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[6] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, pp. 066138, 2004.

[7] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multidimensional densities via-nearest-neighbor distances," *Information Theory, IEEE Transactions on*, vol. 55, no. 5, pp. 2392–2405, 2009.

[8] S. Frenzel and B. Pompe, "Partial mutual information for coupling analysis of multivariate time series," *Physical review letters*, vol. 99, no. 20, pp. 204101, 2007.

[9] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.