Differential Entropy Estimation via One-Class SVM

Milena Marinho Arruda, Luciana Ribeiro Veloso and Francisco Marcos de Assis

***Abstract*— This paper introduces the use of Support Vector Machine for entropy estimation of continuous random variables with well-defined probability density function. The method is based on support estimation and can converge to Shannon entropy or zero-order Rényi entropy depending of effective support set delimited. Simulated results indicate that the method proposed for effective support characterization gives asymptotically good results to Shannon entropy estimation in comparative with other three estimators based on: histogram, kernel smoothing and neighbor distances.**

***Keywords*— Entropy, Estimation, Support Vector Machine.**

I. Introduction

The information theory consolidates a mathematical approach to understanding and development of research in different scientific fields. It has been recently used for example in neuroscience [1], [2], biomedical data [3], [4] and plant wide diagnosis in industrial process [5], [6].

Entropy is the most common and basic concept of information theory. Generally, entropy refers to disorder, uncertainty or amount of information. Something transmits information when it expresses messages that are not expected or unknown, in this way, the least likely messages are those that carry more information (when the transmitted message is known, then the amount of information received is null).

However, although the concepts in information theory are relatively simple, in practice, their estimation can be a complex process. Histogram-based estimation and kernel density estimation are widely used to estimate differential entropy, but, for other measures these estimators can produce bias [7], [8].

Furthermore, estimators that use nearest neighbors distances [9], [7] has been used to minimize such problems. In addition, although techniques using classifiers such as Support Vector Machine (SVM) were slightly exploited for estimation of measures in theory information, they have already shown good results in plant wide diagnosis [10], [5].

In such context, this paper attempts to investigate the use of these four estimators for differential entropy estimation and compare them with theoretical values of entropy to random variables with well defined distributions, i.e. exponential, gamma, normal and uniform. In addition, a method to characterize the effective support set is proposed for these estimates in the one-class SVM algorithm.

This paper is structured as follows. Section II organizes notations. Section III reviews definitions of entropy. Section IV introduces briefly four entropy estimators. In section V, a method is proposed to effective support characterization. Section VI presents the performed simulations. Finally, section VII concludes the paper.

Milena Marinho Arruda, Luciana Ribeiro Veloso and Francisco Marcos de Assis, Center of Electrical Engineering and Informatics, Federal University of Campina Grande (UFCG), Campina Grande-PB, Brazil, E-mails: milena.arruda@ee.ufcg.edu.br, luciana.veloso@dee.ufcg.edu.br, fmarcos@dee.ufcg.edu.br. This work was partially supported by CAPES.

II. Notation and Terminology

During our discussions we denote random variables by uppercase letters, their realizations by lowercase letters, stochastic processes by uppercase bold letters and realizations of d-dimensional random variables by lowercase bold letters. The nth output of the process is indicated by subscripts, X_n. The finite length sequence of a random variable is indicated by subscript and superscript, $X_{N-k}^{N} = \{X_{N-k}, \cdots, X_N\}$. Probability density function is denoted by $f(\cdot)$ and the set where $f(\cdot) > 0$ is called support set, it is expressed by $[\![\cdot]\!]$.

III. Differential Entropy

The most common concept in information theory is that defined by Claude Shannon [11]. For discrete random variables, entropy is a measure of the average uncertainty and the number of bits on average required to describe them.

When the random variable is continuous, there is a differential entropy. According to Shannon it is related to the shortest description volume of these variables [12]. Unlike discrete entropy, differential entropy can be negative and it occurs when this volume is less than one.

A generalization for differential entropy with parameter α was defined by Alfréd Rényi [13] as

$$h_\alpha(X) = \frac{1}{1-\alpha} \log \left[\int f^\alpha(x) d\mu \right], \tag{1}$$

for $0 < \alpha < \infty$ and $\alpha \neq 1$. When $\alpha \to 1$ we obtain the Shannon entropy function

$$h(X) = h_1(X) = -\int f(x) \log f(x) dx, \tag{2}$$

and when $\alpha \to 0$, we obtain zero-order Rényi entropy

$$h_0(X) = \log \phi[\![X]\!], \tag{3}$$

where $\phi(\cdot)$ denotes the Lebesgue measure [14]. When the random variables is discrete, the case of $\alpha = 0$ is associated with Hatley definition [15].

The asymptotic equipartition property (AEP) is a direct consequence of the weak law of large numbers. Considering continuous random variables, the AEP provides an interpretation about differential entropy in which it is the logarithm of the smallest volume that contains most of the probability [12].

Differential Shannon entropy and zero-order Rényi entropy can be related through AEP. Shannon entropy is related with the logarithm of the size of the effective support set and Rényi

entropy gives the logarithm of the Lebesgue measure of the support set. In general, entropy will refer to Shannon entropy [12].

IV. Estimators of Entropy

In this section, four estimators to entropy are presented for continuous random variables. Here we will consider two estimators that may depend on the estimate of the probability density function and two that are direct entropy estimation.

A. Histogram Approach

The simplest and most common approach to estimate the differential entropy of a continuous random variable uses histogram-based estimation technique.

This technique divides the range of random variable into bins of length Δ, and furthermore, it assume that the probability density function is continuous within the bins.

For each box size, the number of observations is counted and the estimate is defined by

$$\widehat{f}(x) = \frac{\text{number of observations in the same bin as } x}{N}, \quad (4)$$

where N is the sample size. The estimation of differential entropy is related by discrete entropy in the sense [12]

$$\widehat{H}(X) + \log \Delta \to h(X), \quad (5)$$

where $\widehat{H}(X)$ is the discrete entropy of quantized variable into bins and it is interpreted as $\widehat{H}(X) = -\sum \widehat{f}(x) \log \widehat{f}(x)$.

B. Kernel Density Estimator

A kernel density estimator (KDE) is a nonparametric estimation of the probability density function of a random variable. This distribution is defined by a smoothing function (defines the shape of the curve used to estimate), and a bandwidth value (controls the smoothness of the resulting density curve).

Considering $x_1, x_2, \cdots, x_N$ realizations of a continuous random variable with unknown distribution. For any real values of x, the kernel density estimator is given by

$$\widehat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} \Theta\left(\frac{x - x_i}{h}\right), \quad (6)$$

where N is the sample size, h is the bandwidth and $\Theta(\cdot)$ is the kernel smoothing function.

A kernel is a non-negative real-valued integrable function that satisfy two additional requirements: 1) Normalization: $\int_{-\infty}^{\infty} \Theta(u)du = 1$; and 2) Symmetry: $\Theta(u) = \Theta(-u)$.

Several types of kernel functions are commonly used, such as: uniform, triangle, Epanechnikov and Gaussian. In this paper the gaussian kernel was used, so, $\Theta(u) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}u^2}$.

After determine density function, the entropy can be estimated using (5).

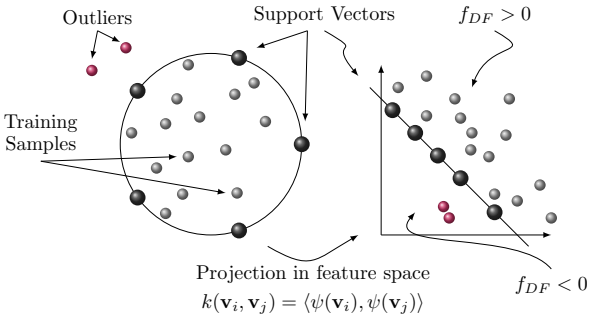

Fig. 1. Classification based on one-class SVM: the training samples are separated from the origin into the feature space by support vectors.

C. Kozachenko-Leonenko Estimator

Unlike a histogram and KDE methods, the Kozachenko-Leonenko (KL) estimator was defined to direct entropy estimation [9].

The KL estimator considers that the probability distribution of the distances between x_i and its k-th nearest neighbor is a trinomial distribution and estimates the entropy by

$$\widehat{h}(X) = -\psi(k) + \psi(N) + \log c_d + \frac{d}{N} \sum_{i=1}^{N} \log \epsilon(i), \quad (7)$$

where $\psi(\cdot)$ is digamma function, N is the sample size, c_d is the volume of the d-dimensional unit ball (for the Euclidean norm $c_d = \pi^{d/2}/\Gamma(1 + d/2)$) and $\epsilon(i)$ is twice the distance from x_i to its k-th neighbor.

D. One-class SVM

One-class SVM was proposed by Schölkopf *et al.* [10] for estimating the support of a high-dimensional distribution. In addition, it was originally proposed for zero-order Rényi entropy estimation by [5].

This algorithm map the training data into the feature space using an inner product space (computed by a kernel) and then a decision function is used to identify whether any samples is within the support set, see Fig. 1.

Let training vectors (d-dimensional vectors of realizations of continuous random variables) denoted by $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N \in \mathcal{X}$, where N is the sample size and $\mathcal{X}$ is a compact subset of $\mathbb{R}^N$. In addition, $\Psi(\cdot)$ is a feature map $\mathcal{X} \to F$, that is, a map into an inner product space F such that the inner product in the image of $\Psi(\cdot)$ can be computed by evaluating some simple kernel,

$$k(\mathbf{v}_i, \mathbf{v}_j) = \langle \Psi(\mathbf{v_i}), \Psi(\mathbf{v_j}) \rangle, \quad (8)$$

where, $\langle \cdot \rangle$ denotes inner product and in this paper it is the radial basis function (RBF) kernel.

The basic idea of the one-class SVM is to map the training vector on a characteristic space in order to separate the vectors of the origin with maximum margin. Therefore, Schölkopf *et al.* [10] proposes the following quadratic problem

$$\min_{\omega \in F, \xi \in \mathbb{R}^N, \rho \in \mathbb{R}} \frac{1}{2}||\omega||^2 + \frac{1}{\nu N} \sum_i \xi_i - \rho$$

$$\text{subject to} \quad (\omega \cdot \psi(\mathbf{X_i})) \leq \rho - \xi_i, \quad \xi_i \leq 0. \tag{9}$$

And reduces it to the dual problem:

$$\min_{\alpha} \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu N}, \quad \sum_i \alpha_i = 1, \tag{10}$$

where α denotes a vector with the coefficients of support vectors, $\nu \in (0, 1]$ denotes an upper bound on the fraction of training points outside the estimated region (outliers).

The decision function check whether each data point of quantized maximum range is within the support set of the model of random variables and it is defined as follows

$$f_{DF}(\mathbf{x}) = \text{sgn}\left(\sum_i \alpha_i(\mathbf{x}_i, \mathbf{x}_j) - \rho\right), \tag{11}$$

where $\rho = \sum_j \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$ is the distance from the separate hyperplane to the origin and sgn is the sign function.

Note that we need to delimit the maximum region in which the support set is defined. For this, a lower bound ($\mathbf{x}_{lb}$) and upper bound ($\mathbf{x}_{ub}$) are defined as minimum and maximum values of training vectors added by the margin called δ according to following inequality

$$\sum_{i=1}^{N} \alpha_i e^{-\gamma||\mathbf{x}_{lb}-\mathbf{x}_i||^2} < \sum_{i=1}^{N} \alpha_i e^{-\gamma||\mathbf{x}_{lb}-\mathbf{x}_{min}||^2} = e^{-\gamma\delta^2}, \tag{12}$$

where $e^{-\gamma\delta^2} = \rho$.

Then, the maximum support is uniformly quantized into $nbins$ intervals with bin size Δ. If the random variable is one-dimensional, the entropy estimation of variable is

$$\widehat{h}(X) = \log(n_x \times \Delta), \tag{13}$$

where n_x is the number of points within the range. Otherwise, for each dimension the distribution model is specified in order to determine the maximum support (Cartesian product) to then estimate the entropy, see Algorithm 1.

Algorithm 1 presents the pseudocode of one-class SVM entropy estimator to multi-dimensional random variables. To attend zero-order Rényi entropy, this method consider the fraction of outliers of the data is quite small ($\nu = 0.01$) [5]. While to attend Shannon entropy, effective support needs a characterization.

V. Effective Support Set Characterization

The applications that uses one-class SVM algorithm to estimation of entropy have been used to attend zero-order Rényi entropy [5]. However, both estimations of zero-order Rényi entropy and Shannon entropy differ only on characterization of support set.

Shannon entropy estimations are made from the logarithm of the size of the effective support set. Therefore, it is necessary establish a characterization of effective support set in one-class SVM algorithm. The effective support is one that contains most of data sample.

Algorithm 1: One-class SVM for entropy estimation

Input: X, $nbins$ `// X is` $N \times d$ `matrix`
Output: $\widehat{h}$

$\gamma = 0.1$
$\nu = 0.01$
$vbins = 1$
for *i = 1 to d* **do**
  Determine the model and δ for $X(\cdot, i)$
  $lb(i) = \min_{1:n} X(\cdot, i) - \delta$
  $ub(i) = \max_{1:n} X(\cdot, i) + \delta$
  $\Delta(i) = \frac{ub(i)-lb(i)}{nbins}$
  $vbins = vbins \times \Delta(i)$
$\widehat{\text{supp}}(X) = [lb(1), \Delta(1), ub(1)] \times \cdots \times [lb(3), \Delta(3), ub(3)]$
Determine the model for X
$n_x = |f_{DF}(\widehat{\text{supp}}(X)) == 1|$
$\widehat{h} = \log(n_x \times vbins)$
return $\widehat{h}$

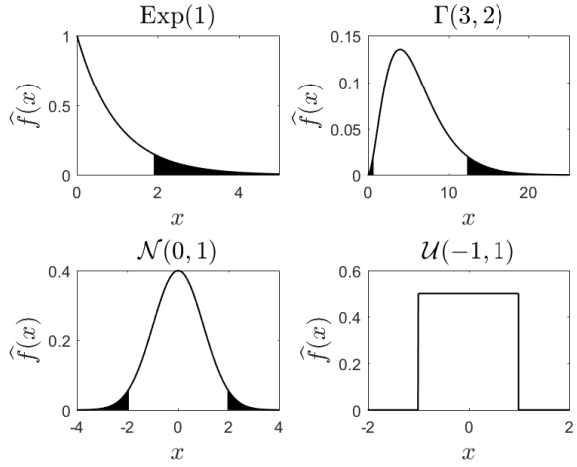

Fig. 2. Probability density function for a random variable X in four scenarios: gamma, exponential, normal and uniform distributed. The filled area represents the region with probability less than 15% of the maximum probability.

In this section we propose a way to determine the effective support on the distributions for use of one-class SVM algorithm to the Shannon entropy estimation by adjusting the fraction of outliers in the algorithm.

Considering the probability density function of some well-defined distribution, we propose that the fraction of outliers is equivalent to fractions of samples that occur with probability less than 15% of the maximum probability. When this area is null we consider $\nu = 0.01$.

Example 1: Let a random variable X gamma-distributed with shape k and scale θ, their probability density function is

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}, \tag{14}$$

with $x \in (0, \infty)$. Suppose that $k = 3$ and $\theta = 2$, the maximum value of $f(x)$ is 0.1353 and occur when $x = 4$. The area with probability less than $0.15 \times 0.1353 = 0.0203$ represents approximately 6%, then, $\nu = 0.06$, see Fig. 2.

VI. Results

With the purpose of evaluating the four presented estimation methods, we generated 20000 samples independent and identically distributed of the continuous random variables with a well-defined distribution, among which: exponential, gamma, normal and uniform. Theoretical values for differential entropy of these distributions will be our referential during subsequent simulations.

In the simulations, parameters were defined for each method. Histogram-based technique has always used 100 bins. KDE estimator used $h = 0.01$. Kozachenko-Leonenko estimator used $k = 1$. One-class SVM was implemented according to Algorithm 1 including the characterization of the effective support set (see Fig. 2). In the Table I are arranged the value of the parameter ν using the methodology presented for two set of parameter for each one of four different distributions.

Fig. 3, 4, 5 and 6 show the results for entropy estimation as function of the number of data samples, considering exponential, gamma, normal and uniform distributions respectively. In all figures, black lines indicate the analytical entropy value, continuous magenta lines indicate histogram technique, continuous green lines indicate KDE, continuous red lines indicate KL estimation and continuous blue line indicate one-class SVM estimation.

In most simulations, KDE was the estimator that took longer to converge (see Fig. 3, 4 and 5). Algorithms that have no dependency on the estimation of probability density function (KL and one-class SVM estimators) presented faster convergence. The KL estimator has a correction term that is crucial for debiasing it for large numbers of samples [16]. As well as histogram-based technique both gives asymptotically good results.

The first interesting features of the one-class SVM method is that using our proposal of effective support characterization, it is possible choose to estimate both zero-order Rényi entropy and Shannon entropy. For Shannon entropy estimation, this method gives good results in our simulations considering different parameters of each distribution. In some situations the algorithm was the first to converge to the analytical values in function of number of samples analyzed.

However, if the probability density function takes an infinite value for some sample realization, the effective support characterization is not efficient. For example for $\mathcal{X} \sim \Gamma(0.5, 1)$ (see Fig. 4), the estimation converge to zero-order Rényi entropy. This is because the outliers are those samples that happen with probability less than a fraction of the maximum value of density probability. In this case, histogram technique and KDE also presented most significant errors in estimation.

Concerning the time of the entropy estimation all methods consume little time. Nevertheless, the one-class SVM estimator is more time consuming and its time is related with parameter ν. When ν assume the minimum value (0.01) each trial takes approximately $0.15s$ while histogram, KDE and KL takes $0.05s$, $0.006s$ and $0.015s$, respectively. In our simulations the maximum value of ν was 0.15 and each trial takes approximately $2.7s$. We ran these estimates on a computer with a 2.70 GHz processor.

TABLE I

Parameter ν for distributions exponential, gamma, normal and uniform.

Distribution	Exp(1)	$\Gamma(3,2)$	$\mathcal{N}(2,9)$	$\mathcal{U}(-1,1)$
ν	0.15	0.0606	0.0514	0.01
Distribution	Exp(2)	$\Gamma(2,0.5)$	$\mathcal{N}(0,1)$	$\mathcal{U}(0,0.5)$
ν	0.15	0.0694	0.0478	0.01

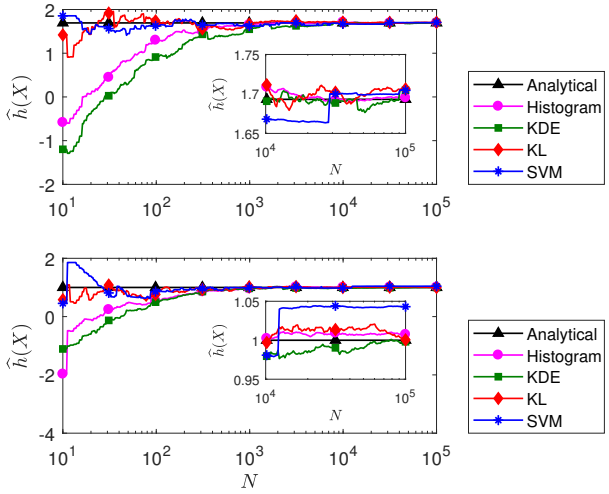

Fig. 3. Entropy estimation for 20000 realizations for $X \sim \text{Exp}(2)$ (upper) and $X \sim \text{Exp}(1)$ (lower).

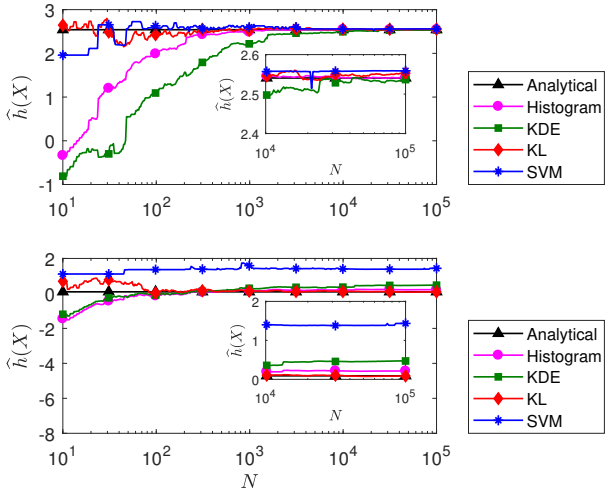

Fig. 4. Entropy estimation for 20000 realizations for $X \sim \Gamma(3,2)$ (upper) and $X \sim \Gamma(0.5,1)$ (lower).

VII. Conclusions

The one-class SVM was firstly used to attend zero-order Rényi entropy. However, zero-order Rényi entropy and Shannon entropy, both for continuous random variables, can be related through asymptotic equipartition property. Based on this relationship we proposed the effective support characterization that extend this method to Shannon entropy estimation.

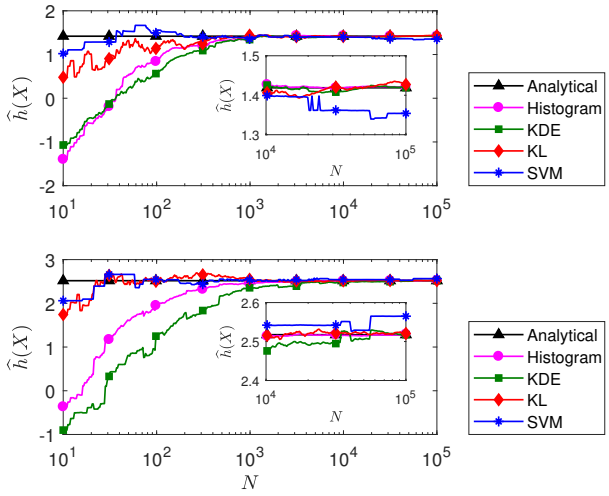

Fig. 5. Entropy estimation for 20000 realizations for $X \sim \mathcal{N}(0,1)$ (upper) and $X \sim \mathcal{N}(2,9)$(lower).

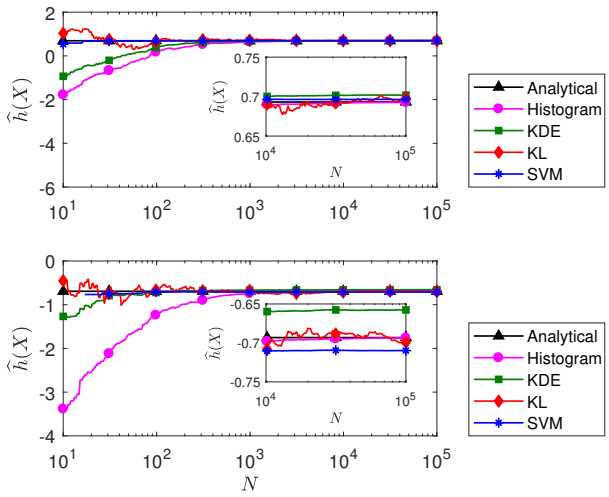

Fig. 6. Entropy estimation for 20000 realizations for $X \sim \mathcal{U}(-1,1)$ (upper) and $X \sim \mathcal{U}(0,0.5)$ (lower).

The method proposed gives asymptotically good results for Shannon entropy estimation in seven of eight cases analyzed. The special case occur when when probability density function assume infinity value and the estimation always converge to zero-order Rényi entropy.

Results of performed simulations to evaluate and compare the entropy estimation using the proposed SVM and three other estimators showed there is not significant difference between the methods. In some situations the proposed algorithm is the first to converge according to the number of sample data presented.

Although one-class SVM is more time consuming, it can be useful in applications for causality analysis with non-stationary data and can be applied to estimation of others measures like as transfer 0-entropy [5].

REFERENCES

[1] J. M. de Assis and F. M. de Assis, "An application of directed information to infer synaptic connectivity," in *Anais do XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, Santarém, Brazil*, p. 528–532, 2016.

[2] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy—a model-free measure of effective connectivity for the neurosciences," *Journal of Computational Neuroscience*, vol. 30, pp. 45–67, Feb 2011.

[3] M. M. Arruda, L. R. Veloso, and F. M. de Assis, "Transfer entropy characterization of tbi cases," in *Anais do XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, São Pedro, Brazil*, pp. 625–628, 2017.

[4] F. Marzbanrad, Y. Kimura, M. Endo, M. Palaniswami, and A. H. Khandoker, "Transfer entropy analysis of maternal and fetal heart rate coupling," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7865–7868, Aug 2015.

[5] P. Duan, F. Yang, S. L. Shah, and T. Chen, "Transfer zero-entropy and its application for capturing cause and effect relationship between variables," *IEEE Transactions on Control Systems Technology*, vol. 23, pp. 855–867, may 2015.

[6] M. Bauer, J. W. Cox, M. H. Caveness, J. J. Downs, and N. F. Thornhill, "Finding the direction of disturbance propagation in a chemical process using transfer entropy," *IEEE Transactions on Control Systems Technology*, vol. 15, pp. 12–21, Jan 2007.

[7] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004.

[8] J. M. de Assis, M. O. Santos, and F. M. de Assis, "Auditory stimuli coding by postsynaptic potential and local field potential features," *PLOS ONE*, vol. 11, p. e0160089, aug 2016.

[9] L. 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.

[10] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443–1471, jul 2001.

[11] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, july, october 1948.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[13] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, (Berkeley, Calif.), pp. 547–561, University of California Press.

[14] W. Rudin *et al.*, *Principles of mathematical analysis*, vol. 3. McGraw-hill New York, 1964.

[15] R. V. Hartley, "Transmission of information 1," *Bell System technical journal*, vol. 7, no. 3, pp. 535–563, 1928.

[16] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," in *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, jun 2017.