INT J COMPUT COMMUN, ISSN 1841-9836
8(5):689-698, October, 2013.

Asymptotically Unbiased Estimator of the Informational Energy with kNN

A. Caţaron, R. Andonie, Y. Chueh

Angel Caţaron
Electronics and Computers Department
Transylvania University of Braşov, Romania
cataron@unitbv.ro

Răzvan Andonie*
1. Computer Science Department
Central Washington University, Ellensburg, USA
andonie@cwu.edu
2. Electronics and Computers Department
Transylvania University of Braşov, Romania
*Corresponding author

Yvonne Chueh
Department of Mathematics
Central Washington University, Ellensburg, USA
chueh@cwu.edu

Abstract: Motivated by machine learning applications (e.g., classification, function approximation, feature extraction), in previous work, we have introduced a nonparametric estimator of Onicescu's informational energy. Our method was based on the k-th nearest neighbor distances between the n sample points, where k is a fixed positive integer. In the present contribution, we discuss mathematical properties of this estimator. We show that our estimator is asymptotically unbiased and consistent. We provide further experimental results which illustrate the convergence of the estimator for standard distributions.
Keywords: machine learning, statistical inference, asymptotically unbiased estimator, k-th nearest neighbor, informational energy.

1 Introduction

Inference is based on a strong assumption: using a *representative* training set of samples to infer a model. In this case, we select a random sample of the population, perform a statistical analysis on this sample, and use these results as an estimation to the desired statistical characteristics of the population as a whole. The more representative the sample is, the higher our confidence level reaches so that the statistical results obtained by using this sample are indeed a good estimation to the desired population parameters. We gauge the representativeness of a sample by how well its statistical characteristics reflect the probabilistic characteristics of the entire population. Many standard techniques may be used to select a representative sample set [15]. However, if we do not use expert knowledge, selecting the most representative training set from a given dataset was proved to be computationally difficult (NP-hard) [10]. The problem is actually more difficult, since in most applications the complete dataset is unknown, or too large to be analyzed. Therefore, we have to rely on a more or less representative training set.

A critical aspect of many machine learning approaches is how well an information theory measure is estimated from the available training set. This relates to a fundamental concept in statistics: probability density estimation. *Density estimation* is the construction of an estimate of the density function from the observed data [20]. We will refer here only to *nonparametric*

estimation, where less rigid assumptions will be made about the distribution of the observed data. Although it will be assumed that the distribution has the probability density f, the data will be allowed to speak for themselves in determining the estimate of f more than would be the case if f were constrained to fall in a given parametric family. A common measure used in machine leaning is mutual information (MI). Several methods were proposed for MI estimation [18], [22], [13]: histogram based estimators, kernel density estimators, B-spline estimators, k-th nearest neighbor (kNN) estimators, and wavelet density estimators.

Estimating entropy and MI is known to be a non-trivial task [4]. Naïve estimations (which attempt to construct a histogram where every point is the center of a sampling interval) are plagued with both systematic (bias) and statistical errors. An ideal estimator does not exist, instead the choice of the estimator depends on the structure of data to be analyzed. It is not possible to design an estimator that minimizes both the bias and the variance to arbitrarily small values. The existing studies have shown that there is always a delicate trade off between the two types of errors [4].

MI is generally based on the classical Shannon type MI. However, it is computationally attractive to use one of its generalized forms: the Rényi divergence measure, which uses Rényi's quadratic entropy. The reason is that, as proved by Principe *et al.*, Rényi's quadratic entropy (and Rényi's divergence measure) can be estimated from a set of samples using Parzen's windows approach [19]. The MI and Rényi's divergence measure are equivalent, but only in the limit $\alpha = 1$, where α is the order of Rényi's divergence measure [19].

A unilateral dependency measure can be derived from Onicescu's informational energy (IE). This measure proved to be an efficient alternative to MI, and we have estimated it from sample datasets using the Parzen windows approach. We used this approach in classification and feature weighting [2], [5], [6], [3], [7]. An important drawback of this approach is the fact that Parzen windows estimate cannot be applied on continuous spaces. This is also true for Shannon's type MI. Therefore, This means an important machine learning domain - continuous function approximation (or prediction), is left out.

In previous work [8], we introduced a kNN IE estimator which may be used to approximate the unilateral dependency measure both in the discrete and the continuous case. An important theoretical aspect was not discussed yet: the asymptotic behavior of this estimator in terms of unbiasedness and consistency. Generally, any statistic whose mathematical expectation is equal to a parameter is called *unbiased* estimator of that parameter. Otherwise, the statistic is said to be *biased*. Any statistic that converges asymptotically to a parameter is called *consistent* estimator of that parameter [12].

Consistent and unbiased are not equivalent. A simple example of a biased consistent estimator is if the mean of samples $x_1, x_2, \ldots, x_n$ is estimated by $1/n\sum x_i + 1/n$. This estimate is biased but consistent, since it approaches asymptotically the correct value. An *asymptotically unbiased* estimator is an estimator that is unbiased as the sample size tends to infinity. Some biased estimators are asymptotically unbiased but all unbiased estimators are asymptotically unbiased. The previous estimator is biased but asymptotically unbiased. One way to prove that an estimator is consistent is to prove that it is asymptotically unbiased and the variance goes to zero.

This gives the motivation for the present work. We show that our IE estimator is asymptotically unbiased and consistent. This will imply that the estimator is "good".

First, we summarize (Section 2) the IE and the kNN method. Section 3 describes our IE approximation method, including the novel theoretical results. After the experimental results, exposed in Section 4, we conclude with final remarks and a description of future work (Section 5).

2 Background

2.1 Onicescu's Informational Energy

Generally, information measures refer to uncertainty. Since Shannon defined his probabilistic information measure in 1948, many other authors, with Rényi, Daroczy, Bongard, Arimoto, and Guiaşu among them, have introduced new measures of information. However, information measures can also refer to certainty, and probability can be considered as a measure of certainty. More general, any monotonically growing and continuous probability function can be considered as a measure of certainty. For instance, Onicescu's IE was interpreted by several authors as a measure of expected commonness, a measure of average certainty, or as a measure of concentration.

For a continuous random variable X with probability density function $f(x)$, the IE is [11,17]:

$$IE(X) = \int_{-\infty}^{+\infty} f^2(x)dx \tag{1}$$

2.2 The nearest neighbor method

Although classification remains the primary application of kNN, we can use it to do density estimation also. Since kNN is non parametric, it can do estimation for arbitrary distributions. The idea is very similar to use of Parzen window. Instead of using hypercube and kernel functions, here we do the estimation as follows.

The kNN estimators represent an attempt to adapt the amount of smoothing to the "local" density of data. The degree of smoothing is controlled by an integer k, chosen to be considerably smaller than the sample size; typically $k \approx n^{1/2}$. Define the distance $d(x, y)$ between two points on the line to be $|x \times y|$ in the usual way, and for each t define $d_1(t) \leq d_2(t) \leq \ldots \leq d_n(t)$ to be the distances, arranged in ascending order, from t to the points of the sample.

The kNN density estimate $f(t)$ is defined by [20]:

$$\hat{f}(t) = \frac{k}{2nd_k(t)} \tag{2}$$

The kNN was used for non-parametrical estimate of the entropy based on the k-th nearest neighbor distance between n points in a sample, where k is a fixed parameter and $k \leq n - 1$. Based on the first nearest neighbor distances, Leonenko *et al.* [14] introduced an asymptotic unbiased and consistent estimator H_n of the entropy $H(f)$ in a multidimensional space. When the sample points are very close one to each other, small fluctuations in their distances produce high fluctuations of H_n. In order to overcome this problem, Singh *et al.* [21] defined an entropy estimator based on the k-th nearest neighbor distances. A kNN estimate of the Kullback-Leibler divergence was obtained by Wang *et al.* in [23]. A mean of several kNN estimators corresponding to different values of k was used by Faivishevsky *et al.* in [9] for developing a smooth estimator of differential entropy, mutual information, and divergence.

According to [22], kNN MI estimation outperforms histogram methods. kNN works well if the value of k is optimally chosen. However, there is no model selection method for determining the number of neighrest neighbors k. This is a limitation of the kNN estimation.

3 Estimation of the Informational Energy

We are ready now to introduce our kNN method for IE approximation, using results from our previous work [8]. The described theoretical properties are however novel. Mathematical proofs are omitted, since they would not fit into the page limit of this paper.

The IE can be easily computed if the data sample is extracted from known distributions. When the underlying distribution of data sample is unknown, the IE has to be estimated. More formally, our goal is to estimate (1) from a random sample X_1, X_2, ..., X_n of n d-dimensional observations from a distribution with the unknown probability density $f(x)$. This problem is even more difficult if the number of available points is small.

The $IE_{empirical}$ is not a good estimate especially when the relative frequencies are far from the true probabilities. This is generally the case for small datasets and, in accordance to the central limit theorem, for an increasing number of samples, $IE_{empirical}$ converges probabilistically to IE.

The IE is the average of $f(x)$, therefore we have to estimate $f(x)$. The n observations from our samples have the same probability $\frac{1}{n}$. A convenient estimator of the IE is:

$$\hat{IE}_k^{(n)}(f) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(X_i). \tag{3}$$

We will determine first the probability density $P_{ik}(\epsilon)$ of the random distance $R_{i,k,n}$ between a fixed point X_i and its k-th nearest neighbor from the remaining $n-1$ points. Probability $P_{ik}(\epsilon)d\epsilon$ of the k-th nearest neighbor to be within distance $R_{i,k,n} \in [\epsilon, \epsilon + d\epsilon]$ from X_i, $k-1$ points at a smaller distance and $n-k-1$ at a larger distance can be expressed in terms of the trinomial formula [9]:

$$P_{ik}(\epsilon)d\epsilon = \frac{(n-1)!}{1!(k-1)!(n-k-1)!}dp_i(\epsilon)p_i^{k-1}(1-p_i)^{n-k-1},$$

where $p_i(\epsilon) = \int_{\|x-X_i\|<\epsilon} f(x)dx$ is the mass of the ϵ-ball centered at X_i and $\int P_{ik}(\epsilon)d\epsilon = 1$.

We can express the expected value of the $p_i(\epsilon)$ using the probability mass function of the trinomial distribution:

$$E_{P_{ik}(\epsilon)}(p_i(\epsilon)) = \int_0^\infty P_{ik}(\epsilon)p_i(\epsilon)d\epsilon =$$

$$= k\binom{n-1}{k}\int_0^1 p^{k-1}(1-p)^{n-k-1}pdp =$$

$$= k\binom{n-1}{k}\int_0^1 p^{(k+1)-1}(1-p)^{(n-k)-1}dp.$$

This equality can be reformulated using the *Beta* function:

$$B(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1} = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}.$$

We obtain:

$$E_{P_{ik}(\epsilon)}(p_i(\epsilon)) = k\binom{n-1}{k}\frac{\Gamma(k+1)\Gamma(n-k)}{\Gamma(n+1)} =$$

$$= k\frac{(n-1)!}{(n-k-1)!k!}\frac{k!(n-k-1)!}{n!},$$

which can be rewritten as:

$$E_{P_{ik}(\epsilon)}(p_i(\epsilon)) = \frac{k}{n}. \tag{4}$$

On the other hand, assuming that $f(x)$ is almost constant in the entire ϵ-ball around X_i [9], we have:

$$p_i(\epsilon) \approx V_1 R_{i,k,n}^d f(X_i),$$

where we denote the volume of the ball of radius $\rho_{r,n}$ in a d-dimensional space by:

$$V_{\rho_{r,n}} = V_1 \rho_{r,n}^d = \frac{\pi^{\frac{p}{2}} \rho_{r,n}^d}{\Gamma(\frac{p}{2}+1)}.$$

V_1 is the volume of the unit ball and $R_{i,k,n}$ is the Euclidean distance between the reference point X_i and its k-th nearest neighbor. This means that $V_1 R_{i,k,n}^d$ is the volume of the d-dimensional ball of radius $R_{i,k,n}$.

We obtain the expected value of $p_i(\epsilon)$:

$$E(p_i(\epsilon)) = E(V_1 R_{i,k,n}^d f(X_i)) = V_1 R_{i,k,n}^d \hat{f}(X_i). \tag{5}$$

Equations (4) and (5) both estimate $E(p_i(\epsilon))$. Their results are approximatively equal:

$$V_1 R_{i,k,n}^d \hat{f}(X_i) = \frac{k}{n},$$

That is:

$$\hat{f}(X_i) = \frac{k}{n V_1 R_{i,k,n}^d}, i = 1...n. \tag{6}$$

This is the estimate of the probability density function. By substituting $\hat{f}(X_i)$ in (3), we finally obtain the following IE approximation:

$$\hat{IE}_k^{(n)}(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{k}{n V_1 R_{i,k,n}^d}. \tag{7}$$

We have introduced this result in [8]. Our main question now is to analyze its asymptotic behavior.

Consistency of an estimator means that as the sample size gets large, the estimate gets closer and closer to the true value of the parameter. Unbiasedness is a statement about the expected value of the sampling distribution of the estimator. The ideal situation, of course, is to have an unbiased consistent estimator. This may be very difficult to achieve.

Yet unbiasedness is not essential, and just a little bias is permitted as long as the estimator is consistent. Therefore, an asymptotically unbiased consistent estimator may be acceptable. In the following, we will use the following mathematical property (from [16]): An asymptotically unbiased estimator with asymptotic zero variance is consistent.

We are ready now to state our theoretical results:

1. The informational energy estimator $\hat{IE}_k^{(n)}(f)$ is asymptotically unbiased.
2. $\lim_{n\to\infty} Var\left[\hat{H}_k^{(n)}(f)\right] = 0.$

Therefore, we can conclude that the $\hat{IE}_k^{(n)}(f)$ estimator is consistent.

4 Experiments

When the distribution of a sample is unknown, the statistical measures cannot be calculated directly, and we have to use an estimate. The quality of an estimator can be determined by studying its asymptotic behavior. We proved that the informational energy estimator $\widehat{IE}$ is

asymptotically unbiased and consistent. It provides an approximation of the informational energy regardless of the distribution where the sample was drawn from. It is interesting to compare the estimated value of the IE with its real value. For an unidimensional distribution, we can achieve this goal by generating a random sample from a known distribution $f(x)$, with x_{min} and x_{max} being the minimum / maximum values. Then, the informational energy of the distribution $f(x)$ on the subdomain $\mathcal{D}_{sample} = \{x | x \in [x_{min}, x_{max}]\}$ is

$$IE_{\mathcal{D}_{sample}} = \int_{x_{min}}^{x_{max}} f^2(x)dx = \int_{\mathcal{D}_{sample}} f^2(x)dx, \tag{8}$$

while the estimated informational energy $\widehat{IE}$ is given by the formula (7). The information energy of the same distribution has a fixed value when it is computed on its entire definition domain $\mathcal{D}$:

$$IE_{\mathcal{D}} = \int_{\mathcal{D}} f^2(x)dx. \tag{9}$$

Our experiments focus on the following distributions: Exponential, unidimensional Gaussian, Beta, Cauchy, Gamma, and Weibull. We use the R programming language and environment functions to generate the random samples from each distribution, with the parameters listed in Tables 1–6. The first line in each table contains: the probability density function of the distribution, the support of this function (which is the domain $\mathcal{D}$), the values of the parameters, and the IE computed with formula (9).

Sample size is the number of values from the random sample, and *Range* is the interval limited by the minimum / maximum values from the sample used to compute the IE with formula (8). In order to study the asymptotically unbiasedness and consistency of the estimator, we determine the value of $\hat{IE}$ for samples with 10, 100, 1000 values, and with increasing values of k, from 1 to the squared root of the sample size [20].

Table 1: Exponential distribution

$f(x) = \theta e^{-\theta x}$, $x \geq 0$, $\theta = 3$, $IE_{\mathcal{D}} = 1.5$						
Sample size: 10; Range: [0.022, 0.777]; $IE_{\mathcal{D}_{sample}} = 1.3$						
k	1	2	3			
$\hat{IE}$	8.133	2.464	1.023			
Sample size: 100; Range: [0.0007, 2.599]; $IE_{\mathcal{D}_{sample}} = 1.493$						
k	1	2	3	5	7	9
$\hat{IE}$	4.953	2.312	2.289	2.167	1.854	1.743
Sample size: 1000; Range: [0.0001, 2.237]; $IE_{\mathcal{D}_{sample}} = 1.499$						
k	1	2	3	10	20	30
$\hat{IE}$	6.829	3.093	2.269	1.605	1.527	1.507

In general, the expected behavior was confirmed by experiments: the larger the sample size n, the more accurate estimation of the informational energy.

Table 2: Unidimensional Gaussian distribution

$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $\mu = 0$, $\sigma = 1$, $IE_{\mathcal{D}} = 0.282$						
Sample size: 10, Range: $[-1.018, 1.395]$, $IE_{\mathcal{D}_{sample}} = 0.254$						
k	1	2	3			
$\hat{IE}$	2.559	0.444	0.402			
Sample size: 100, Range: $[-2.263, 2.484]$, $IE_{\mathcal{D}_{sample}} = 0.281$						
k	1	2	3	5	7	9
$\hat{IE}$	0.998	0.462	0.333	0.286	0.275	0.271
Sample size: 1000, Range: $[-3.596, 2.781]$, $IE_{\mathcal{D}_{sample}} = 0.282$						
k	1	2	3	10	20	30
$\hat{IE}$	1.419	0.526	0.421	0.315	0.296	0.293

Table 3: Beta distribution

$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1-x)^{\beta-1}x^{\alpha-1}$, $0 \leq x \leq 1$, $\alpha = 2, \beta = 3$, $IE_{\mathcal{D}} = 1.371$						
Sample size: 10, Range: $[0.194, 0.773]$, $IE_{\mathcal{D}_{sample}} = 1.170$						
k	1	2	3			
$\hat{IE}$	4.873	2.452	2.083			
Sample size: 100, Range: $[0.010, 0.842]$, $IE_{\mathcal{D}_{sample}} = 1.369$						
k	1	2	3	5	7	9
$\hat{IE}$	5.084	2.588	2.167	1.899	1.522	1.523
Sample size: 1000, Range: $[0.010, 0.930]$, $IE_{\mathcal{D}_{sample}} = 1.371$						
k	1	2	3	10	20	30
$\hat{IE}$	101.969	2.617	2.103	1.516	1.450	1.430

Table 4: Cauchy distribution

$f(x) = \frac{b}{\pi[(x-m)^2+b^2]}$, $x \in R$, $m = 0$, $b = 1$, $IE_{\mathcal{D}} = 0.159$						
Sample size: 10, Range: $[-29.068, 61.499]$, $IE_{\mathcal{D}_{sample}} = 0.159$						
k	1	2	3			
$\hat{IE}$	1.342	0.253	0.193			
Sample size: 100, Range: $[-19.543, 17.052]$, $IE_{\mathcal{D}_{sample}} = 0.159$						
k	1	2	3	5	7	9
$\hat{IE}$	37.599	0.204	0.188	0.154	0.158	0.158
Sample size: 1000, Range: $[-232.181, 165.562]$, $IE_{\mathcal{D}_{sample}} = 0.159$						
k	1	2	3	10	20	30
$\hat{IE}$	0.859	0.343	0.284	0.170	0.164	0.161

Table 5: Gamma distribution

$f(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\Gamma(\alpha)\theta^{\alpha}}$, $x \geq 0$, $\theta = 1$, $\alpha = 3$, $IE_{\mathcal{D}} = 0.187$						
Sample size: 10, Range: [1.381, 5.340], $IE_{\mathcal{D}_{sample}} = 0.156$						
k	1	2	3			
$\hat{IE}$	0.204	0.232	0.240			
Sample size: 100, Range: [0.556, 9.053], $IE_{\mathcal{D}_{sample}} = 0.186$						
k	1	2	3	5	7	9
$\hat{IE}$	0.624	0.318	0.285	0.264	0.233	0.232
Sample size: 1000, Range: [0.092, 11.866], $IE_{\mathcal{D}_{sample}} = 0.187$						
k	1	2	3	10	20	30
$\hat{IE}$	1.768	0.344	0.280	0.210	0.201	0.196

Table 6: Weibull distribution

$f(x) = \frac{\alpha x^{\alpha-1}}{\beta^{\alpha} e^{\left(\frac{x}{\beta}\right)^{\alpha}}}$, $x \geq 0$, $\alpha = 3$, $\beta = 4$, $IE_{\mathcal{D}} = 0.213$						
Sample size: 10, Range: [2.040, 5.202], $IE_{\mathcal{D}_{sample}} = 0.191$						
k	1	2	3			
$\hat{IE}$	1.315	0.583	0.523			
Sample size: 100, Range: [0.899, 7.295], $IE_{\mathcal{D}_{sample}} = 0.212$						
k	1	2	3	5	7	9
$\hat{IE}$	0.551	0.368	0.296	0.227	0.209	0.215
Sample size: 1000, Range: [0.393, 7.438], $IE_{\mathcal{D}_{sample}} = 0.213$						
k	1	2	3	10	20	30
$\hat{IE}$	1.416	0.379	0.287	0.234	0.226	0.223

5 Conclusions and Future Work

We have introduced a novel non-parametric kNN approximation method for computing the IE from data samples. In accordance to our results, the $\hat{IE}_k^{(n)}(f)$ estimator is consistent.

In order to study the interaction between two random variables X and Y, the following measure of unilateral dependency was defined by Andonie *et al.* [1]:

$$o(Y, X) = IE(Y|X) - IE(Y)$$

This measure quantifies the unilateral dependence characterizing Y with respect to X and corresponds to the amount of information detained by X about Y. There is an obvious analogy between $o(Y, X)$ and the MI, since both measure the same phenomenon. However, the MI is a symmetric, not a unilateral measure.

Rather than approximating $o(Y, X)$ as we did in our previous studies, in our future work we will approximate directly the IE from the available dataset, using the $\hat{IE}_k^{(n)}(f)$ estimator. We also plan to apply our IE estimator to machine learning techniques.

Bibliography

[1] Andonie, R; Petrescu, F.; Interacting systems and informational energy, *Foundation of Control Engineering*, 11:53-59, 1986.

[2] Andonie, R.; Caţaron, A.; An informational energy LVQ approach for feature ranking, *Proc. of the European Symposium on Artificial Neural Networks ESANN 2004, Bruges, Belgium, April 28-30, 2004*, D-side Publications, 471-476, 2004.

[3] Andonie, R.; How to learn from small training sets, *Dalle Molle Institute for Artificial Intelligence (IDSIA)*, Manno-Lugano, Switzerland, September, invited talk, 2009.

[4] Bonachela, J.A.; Hinrichsen, H.; Munoz, M.A.; Entropy estimates of small data sets, *J. Phys. A: Math. Theor.*, 41:202001, 2008.

[5] Caţaron, A.; Andonie, R.; Energy generalized LVQ with relevance factors, *Proc. of the IEEE International Joint Conference on Neural Networks IJCNN 2004*, Budapest, Hungary, July 26-29, 2004, ISSN 1098-7576, 1421-1426, 2004.

[6] Caţaron, A.; Andonie, R.; Informational energy kernel for LVQ, *Proc. of the 15th Int. Conf. on Artificial Neural Networks ICANN 2005, Warsaw, Poland, September 12-14, 2005*, W. Duch et al. (Eds.): Lecture Notes in Computer Science 3697, Springer-Verlag Berlin Heidelberg, 601-606, 2005.

[7] Caţaron, A.; Andonie, R.; Energy supervised relevance neural gas for feature ranking, *Neural Processing Letters*, 1(32):59-73, 2010.

[8] Caţaron, A.; Andonie, R.; How to infer the informational energy from small datasets, *Proc. of the Optimization of 13th International Conference on Electrical and Electronic Equipment (OPTIM2012)*, Brasov, Romania, May 24-26, 1065-1070, 2012.

[9] Faivishevsky, L.; Goldberger, J.; ICA based on a smooth estimation of the differential entropy, *Proc. of the Neural Information Processing Systems*, NIPS 2008.

[10] Gamez, J.E.; Modave, F.; Kosheleva, O.; Selecting the most representative sample is NP-hard: Need for expert (fuzzy) knowledge, *Proc. of the IEEE World Congress on Computational Intelligence WCCI 2008*, Hong Kong, China, June 1-6, 1069-1074, 2008.

[11] Guiasu, S.; *Information theory with applications*, McGraw Hill, New York, 1977.

[12] Hogg, R.V.; *Introduction to mathematical statistics, 6/E*, Pearson Education, ISBN 9788177589306, 2006.

[13] Kraskov, A.; Stögbauer, H.; Grassberger, P.; Estimating mutual information, *Phys. Rev. E*, American Physical Society, 6(69):1-16, 2004.

[14] Kozachenko, L. F.; Leonenko, N. N.; Sample estimate of the entropy of a random vector, *Probl. Peredachi Inf.*, 2(23):9-16, 1987.

[15] Lohr, H.; *Sampling: Design and analysis*, Duxbury Press, 1999.

[16] Miller, M.; Miller M.; *John E. Freund's mathematical statistics with applications*, Pearson-/Prentice Hall, Upper Saddle River, New Jersey, 2004.

[17] Onicescu, O.; Theorie de l'information. Energie informationelle, *C. R. Acad. Sci. Paris, Ser. A–B*, 263:841-842, 1966.

[18] Paninski, L.; Estimation of entropy and mutual information, *Neural Comput.*, MIT Press, Cambridge, MA, USA, ISSN 0899-7667, 6(15):1191-1253, 2003.

[19] Principe, J. C.;Xu, D.;Fisher, J. W. III.; Information-theoretic learning, *Unsupervised adaptive filtering*, ed. Simon Haykin, Wiley, New York, 2000.

[20] Silverman, B.W.; *Density Estimation for statistics and data analysis (Chapman & Hall/CRC Monographs on statistics & Applied Probability)*, Chapman and Hall/CRC, 1986.

[21] Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; Demchuk, E.; Nearest neighbor estimates of entropy, *American Journal of Mathematical and Management Sciences*, 23:301-321, 2003.

[22] Walters-Williams, J.; Li, Y.; Estimation of mutual information: A survey, *Proc. of the 4th International Conference on Rough Sets and Knowledge Technology*, RSKT 2009, Gold Coast, Australia, July 14-16, 2009, Springer-Verlag, Berlin, Heidelberg, 389-396, 2009.

[23] Wang, Q.; Kulkarni, S. R.; Verdu, S. (2006); A nearest-neighbor approach to estimating divergence between continuous random vectors, *Proc. of the IEEE International Symposium on Information Theory*, ISIT 2006, Seattle, WA, USA, July 9-14, 2006, 242-246, 2006.