Applying Tsallis Divergence to Proteins Organization Prediction Problems

Kirill R. Chernyshov

V.A. Trapeznikov Institute of Control Sciences
of the Russian Academy of Sciences
65 Profsoyuznaya, Moscow 117997, Russia
(e-mail: myau@ipu.ru)

Abstract. The paper considers the application of consistent measures of dependence of random values (in other words, those measures that vanish only under the conditions of stochastic independence of such values) to solve the problems of selecting representative indicators characterizing the predictability (stability) of the behavior of biological objects, such as protein units.

Keywords: biomedical systems, measures of dependence, mutual information, prediction, proteins, stochastic signal analysis, Tsallis entropy.

1. PRELIMINARIES

In the analysis of biomedical experimental data (Ciarletta et al., 2016, Garfinkel et al., 2017, Rangayyan, 2005, 2015) in the framework of stochastic studies, the necessity of applying measures of dependence of random variables inevitably arises. In most cases, the traditional linear correlation is used as such a measure. Its application is directly caused by the very formulation of the measurement processing problem, when such a task is based on the application of traditional approaches related to the least-squares method. The main advantage of such a measure is its ease of application, which implies the possibility of deriving explicit analytical relations to determine the required characteristics of the system, and building methods of evaluation using sample data, including those based on the use of dependent observations.

However, it is known that linear correlation as a measure of dependence is able to become zero even in the presence of a deterministic dependence between random values. In particular, this is true for the quadratic dependence, $Y = X^2$, when X is a Laplacian random value (Rajbman, 1981), as well as for an odd transformation in the form $Y = 5X^3 - 3X$, where the random value X has a uniform distribution within the interval $(-1,+1)$ (Rényi, 1959).

To eliminate this disadvantage, it is important and necessary to use more complex nonlinear measures of dependence in the process of stochastic signal processing. The main issue considered in this paper is the application of consistent measures of dependence. In accordance with the terminology of A.N. Kolmogorov, a measure of the dependence $\mathrm{M}(X,Y)$ of two random values X and Y is consistent if $\mathrm{M}(X,Y) = 0$ if and only when the random variables X and Y are stochastically independent. The paper is devoted to issues of consistency of measures of dependence of random values with application to solving problems of selecting representative indicators characterizing the predictability (stability) of the behavior of biological objects, such as protein units.

2. A PROBLEM OF SELECTING INFORMATIVE INDICATORS TO PREDICT THE STABILITY OF PROTEINS FRAGMENTS

The present paper approach based on involving consistent measures of dependence is aimed to be applied within the problem of selecting informative indicators to predict the stability of short fragments of proteins consisting of five amino-acid residues, the pentapeptides, that were earlier shown (Nekrasov et al., 2014) to be able to describe adequately the three-dimensional protein structure. Such proteins units are basic structure elements of protein molecules and play important role in forming their structure. Studying physical-chemical and functional protein properties that are defined by their amino acid sequence is a key problem of advanced biology. Predicting a structure that a protein will take within its folding process (Fabian and Naumann, 2012) is of importance, in particular, to elaborate medicinal agents influencing the performance of biological systems. Problems of such a kind are characterized by a large dimension of indicators spaces, For instance, under studying a cell or tissue sample by use of techniques of sequence analysis of the genome of new generation it is possible to obtain information about the expression of practically all protein-coding genes, as well as short and lengthy non-coding RNA. Typical size of such data amount is extremely small in the comparison with the indicators number. The indicators number is of the order of thousands, while the number of samples is, in the best case, of the order of hundreds (Petrov et al., 2019).

To increase the efficiency and reliability of data analysis results, methods of selecting indicators, revealing regularities inherent in data, clustering data in order to form homogeneous groups are applied (Petrov et al., 2019). Within such a stochastic framework, as a selection criterion, it is natural to apply a suitable measure of dependence of an indicator with the goal variable. Meanwhile, suitability of such a measure of dependence is just to be characterized by its ability of proper revealing such dependence between the indicator and goal variable. Selecting a corresponding measure of dependence meeting the requirements considered is not straightforward and assumes applying a specific apparatus of the probability theory and based on the A. Rényi axioms.

In 1959, A. Rényi formulated axioms that were found to be the most suitable for determining the measure of dependence $\mu(X,Y)$ between two random values X and Y, which is intended to exhaustively characterize such dependence. These axioms are presented below (Rényi, 1959):

A) $\mu(X,Y)$ is defined for any pair of random values X and Y, if none of them is constant with probability 1.

B) $\mu(X,Y)=\mu(Y,X)$.

C) $0 \le \mu(X,Y) \le 1$.

D) $\mu(X,Y)=0$ only if X and Y are independent.

E) $\mu(X,Y)=1$ if there exists a deterministic relationship between X and Y, so that either $Y=\varphi(X)$, or $X=\psi(Y)$, where φ and ψ are some Borel measurable functions.

F) If φ and ψ are some one-to-one Borel measurable functions, then $\mu(\varphi(X),\psi(Y))=\mu(X,Y)$.

G) If the joint probability distribution of X and Y is Gaussian, then $\mu(X,Y)=|r(X,Y)|$, where $r(X,Y)$ is the conventional correlation coefficient between X and Y.

Measures of dependency corresponding to Rényi's axioms, with the possible exception of axiom F, will be called hereinafter consistent in the sense of Rényi.

The conventional correlation coefficient $r(X,Y)$ is, of course, the best known among the various measures of dependence. A more subtle approach to the characterization of the dependence of random variables is considered when applying the correlation ratio

$$\theta(X,Y)=\frac{\mathbf{var}\left(\mathbf{E}\left(Y\big/X\right)\right)}{\mathbf{var}(Y)},\quad \mathbf{var}(Y)>0,$$

and the maximum correlation coefficient $S(X,Y)$, originally introduced by H. Gebelein (1941) and investigated in papers of O.V. Sarmanov (Sarmanov 1963a,b, Sarmanov and Zakharov, 1960), A. Rényi (1959), and others

$$S(X,Y)=\sup_{\{B\},\{C\}}\frac{\mathbf{cov}(B(Y),C(X))}{\sqrt{\mathbf{var}(B(Y))\mathbf{var}(C(X))}},$$

$$\mathbf{var}(B(Y))>0,\ \ \mathbf{var}(C(X))>0.$$

In the formula above, the upper bound is taken over the sets of Borel measurable functions, $\{B\}$ and $\{C\}$, and also, $B\in\{B\}$, $C\in\{C\}$, while $\mathbf{cov}(\cdot,\cdot)$ is the covariance symbol.

However, in the paper of Rényi (1959) it was shown that only the maximum correlation coefficient $S(X,Y)$ corresponds to the above axioms, whereas the conventional correlation coefficient $r(X,Y)$ and correlation ratio $\theta(X,Y)$ do not correspond. In particular, the axioms D, E, F have not been satisfied for the correlation coefficient, and the axioms D, F have not been satisfied for the correlation ratio.

Despite the existing shortcomings, the measure of dependence such as the conventional correlation coefficient is widely applicable for certain biomedical purposes, especially when solving protein research problems (Yanchun Tao et al., 2018, Yuqing Wu et al., 2018), but within the framework of solving another type of problems, for example, predicting the stability of short protein areas, the use of conventional correlation may be unacceptable in the view of the above disadvantages.

In accordance to preceding sections considerations, a measure of dependence to be applied within selecting informative indicators to predict the stability of short proteins fragments is to be consistent in the Kolmogorov sense at least, but at the same time, it is very advisable that the measure would be consistent in the Rényi sense as well due to the necessity to take its values in the unit interval only, since such a normalization can be a reliable characteristic of the selection. In turn, the reliability takes its values in the unit interval. From another hand side, such a consistent in the Rényi sense measure of dependence is to admit its suitable estimation by use of sample date.

3. BUILDING CONSISTENT IN THE RÉNYI SENSE MEASURES OF DEPENDENCE

In turn, it must be emphasized that Kolmogorov consistent measures of dependence will not necessarily be consistent by Rényi based measures. First of all, this refers to the correspondence between the axioms of C and G according to Rényi. This Section presents an approach to building measures of dependence in accordance with the above mentioned Rényi axioms. In particular, the approach includes the implementation of the following three steps.

1) for any measure of dependence M_{XY} *between random values X and Y, it is necessary to calculate this measure for two-dimensional Gaussian density depending on the correlation coefficient* $r(X,Y)$.

2) Represent the resulting expression as a function of the correlation coefficient module

$$\Theta_{\mathrm{M}_{XY}}\left(\left|r(X,Y)\right|\right), \qquad (1)$$

and invert this function.

3) The resulting expression

$$\Theta^{-1}_{\mathrm{M}_{XY}}\left(\mathrm{M}_{XY}\right), \qquad (2)$$

(as a function of the initial measure of dependence M_{XY}*) defines a measure of dependence between two random values X and Y, satisfying the Rényi's axioms C and N.*

In particular, for the maximum correlation coefficient $S(X,Y)$, the corresponding function

$$\Theta_{\mathrm{M}_{XY}}\left(\left|r(X,Y)\right|\right)=\Theta_{S(X,Y)}\left(\left|r(X,Y)\right|\right)$$

is the identical transformation. It should be noted that the calculation of the maximum correlation coefficient is associated with the need to apply a complex iterative method for determining the first eigenvalue and the pair of the first eigenfunctions (corresponding to this first eigenvalue) of the stochastic kernel

$$\frac{p_{xy}(x,y)}{\sqrt{p_y(y)p_x(x)}}.$$

Along with maximum correlation based on the comparison of the moment characteristics and the unconditional joint probability distributions of the pair of random values under consideration, a wide class of measures of dependence is formed by using the direct matching of unconditional marginal and joint probability distribution of random variables. Such a class is known as measures of divergence of probability distributions. Among such measures is the Kullback-Leibler divergence,

$$D^{KL}(f\|g)=-\int\limits_{R^n} f(\mathbf{z})\ln\left(\frac{g(\mathbf{z})}{f(\mathbf{z})}\right)d\mathbf{z},$$

perhaps the most widely known and used, while directly leading to mutual information as per Shannon.

4. NORMALIZED MUTUAL INFORMATION OF THE TSALLIS DIVERGENCE OF THE ORDER ½

In addition to the Kullback-Leibler divergence, which leads to the definition of mutual information according to Shannon, many more general approaches in defining the divergence measure of two probability distributions are known. In particular, the Tsallis divergence of the order α has the form (Tsallis, 2009)

$$D^T_\alpha(f\|g)=\frac{1}{\alpha-1}\left(1-\int\limits_{R^n} f(\mathbf{z})\left(\frac{g(\mathbf{z})}{f(\mathbf{z})}\right)^{\alpha-1}d\mathbf{z}\right). \qquad (3)$$

From a computational point of view, especially when performing calculations using sample data, Tsallis divergence is more preferable than the Kullbak-Leibler divergence, since the latter includes a "logarithm of integral", which is generally recognized to be much more complex in comparison with the Tsallis divergence, where there is no logarithm at all.

Since at $\alpha\to 1$ $D^T_\alpha(f\|g)$ tends to Kullback-Leibler divergence, the latter can be considered as a special case of the Tsallis divergence of the order 1. It is obvious that

$$D^T_\alpha(f\|g)=D^T_\alpha(g\|f)\Leftrightarrow\alpha={}^1\!/_2.$$

Then, for $n=2$ and $f=p_{xy}$, $g=p_x\cdot p_y$ from the expression (3) directly ensues mutual information of the order ½ of random values *X* and *Y*, having the form

$$I^T_{1/2}(X,Y)=2\left(1-\mathbf{E}_{p_{xy}}\left\{\sqrt{\frac{p_x(x)p_y(y)}{p_{xy}(x,y)}}\right\}\right), \qquad (4)$$

where the mathematical expectation is taken over $p_{xy}(x,y)$.

In addition, using the method mentioned in Section 3, including expressions (1), (2), we get

$$\iota^T_{1/2}(X,Y)=2\left(\frac{I^T_{1/2}(X,Y)}{2}-1\right)^{-2}\sqrt{\left(\frac{I^T_{1/2}(X,Y)}{2}-1\right)^4+\sqrt{1-3\left(\left(\frac{I^T_{1/2}(X,Y)}{2}-1\right)^4-1\right)}-2}. \qquad (5)$$

The behavior of the measure of dependence $\iota^T_{1/2}(X,Y)$, represented by expression (5), as a function of the Tsallis mutual information of the order ½ $I^T_{1/2}(X,Y)$, is shown in Fig. 1.

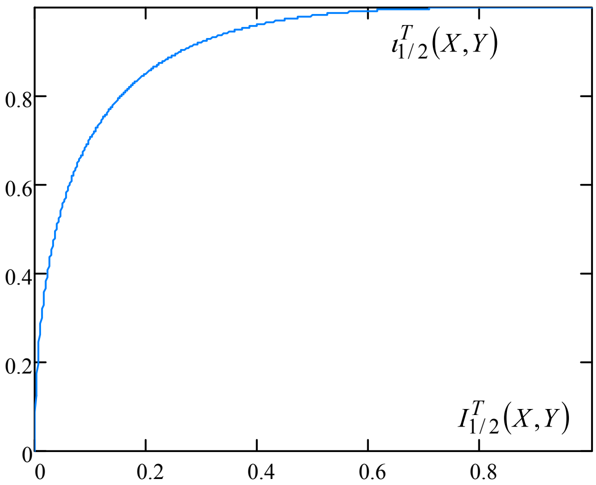


Figure 1. The behavior of the measure of dependence $\iota^T_{1/2}(X,Y)$, presented in (5) as a function of $I^T_{1/2}(X,Y)$.

In turn, estimation of $I^T_{1/2}(X,Y)$ (4) on the basis of sample data can be performed by direct application of Sklar's theorem (Sklar, 1959) to the decomposition of the joint density of the probability distribution using their copula function. In particular, for the probability distribution density $p_{xy}(x,y)$ of random values X and Y with the corresponding marginal probability distribution densities $p_x(x)$, $p_y(y)$ the following expansion is justified:

$$p_{xy}(x,y) = c(P_x(x),P_y(y))p_x(x)p_y(y)\,, \qquad (6)$$

where

$$P_x(x) = \int_{-\infty}^{x} p_x(z)dz\,,\ P_y(y) = \int_{-\infty}^{y} p_y(z)dz$$

are marginal probability distribution functions of random values X and Y, while $c(P_x(x),P_y(y))$ is the copula density function.

In accordance with the presentation of formula (6), expression (4) takes the form

$$I^T_{1/2}(X,Y) = 2\left(1-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sqrt{c(P_x(x),P_y(y))}p_x(x)p_y(y)dxdy\right) = 2\left(1-\int_0^1\int_0^1 c(P_x(x),P_y(y))dP_v(v)dP_u(u)\right). \qquad (7)$$

Representation (7) allows us to apply the method of estimating mutual information according to Shannon (Zeng and Durrani, 2011). At the same time, the case of the Tsallis mutual information of order ½ (4) in the frame of this context becomes much simpler, since in the case of mutual information according to Shannon, the copula density function based on the expansion formula (6), similar to the above, includes the logarithm of the copula density function:

$$c(P_x(x),P_y(y))\ln(c(P_x(x),P_y(y))).$$

At the same time, the application of the estimation method using the copula density function avoids the difficulties that always accompany the division operation.

Summarizing all these justifications, one should conclude that just normalized Tsallis mutual information of order ½ defined by (5) is the measure of dependence meeting the requirements of the described problem of selecting informative indicators to predict the stability of short proteins fragments.

5. EXAMPLE: ZERO CORRELATION UNDER STOCHASTIC DEPENDENCE

As mentioned in Section 1, there are many examples where the use of conventional correlation methods in building models has not yielded satisfactory results. Among such systems, we can distinguish those in which the dependence between the input and output variables is described by the probability distribution density

$$p_{yx;\lambda}(y,x) = p_y(y)p_x(x)\left(1+\lambda\phi_y(y)\phi_x(x)\right), \qquad (8a)$$

with marginal probability distribution densities $p_y(y)$ and $p_x(x)$, as well as functions $\phi_y(y)$ and $\phi_x(x)$, satisfying the conditions

$$\int p_y(y)\phi_y(y)dy = 0\,,\ \int p_x(x)\phi_x(x)dx = 0\,, \qquad (8b)$$

where the parameter λ ensures that the condition is fulfilled:

$$1+\lambda\phi_y(y)\phi_x(x) \geq 0\,. \qquad (8c)$$

Density (8) belongs to the class of probability distributions by O.V. Sarmanov (Balakrishnan and Lai, 2009, Sarmanov, 1967).

For probability distribution density (8), both the correlation coefficient $r(X,Y)$, and the correlation ratio $\theta(X,Y)$ are identically zero if the functions $p_y(y)\phi_y(y)$ or $p_x(x)\phi_x(x)$ are even.

In particular, to the class of distributions of O.V. Sarmanov (8) refers to the following density:

$$p_{yx;\lambda}(y,x)=$$

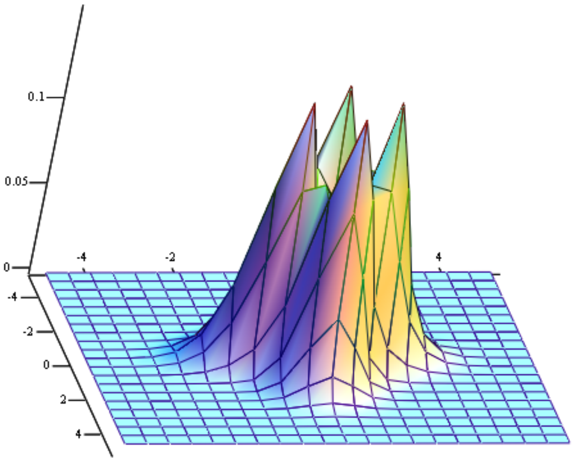
$$=\frac{e^{-\frac{x^2+y^2}{2}}}{2\pi}\left(1+\lambda\left(2e^{-\frac{3}{2}x^2}-1\right)\left(2e^{-\frac{3}{2}y^2}-1\right)\right), \qquad (9)$$

$$-1\le\lambda\le 1\,.$$

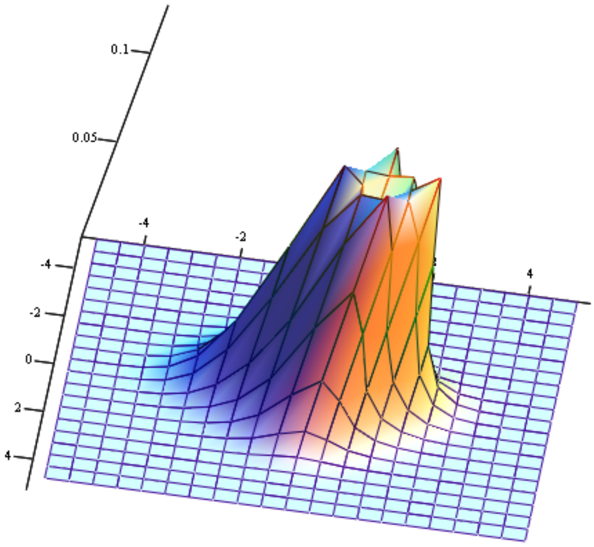
The maximum correlation coefficient for it is as follows:

$$S(X,Y)=\left(\frac{4}{\sqrt{7}}-1\right)|\lambda|\,.$$

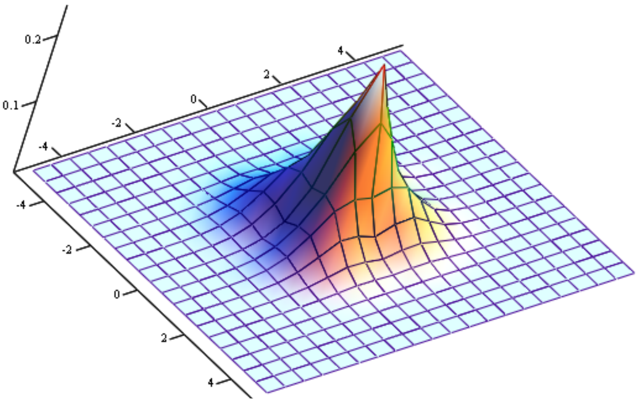
Despite its scalar nature, the magnitudes of the parameter λ significantly affect the shape of the probability distribution density (9). Fig. 2 shows the shape of the probability distribution density (9) for different values of the parameter λ.



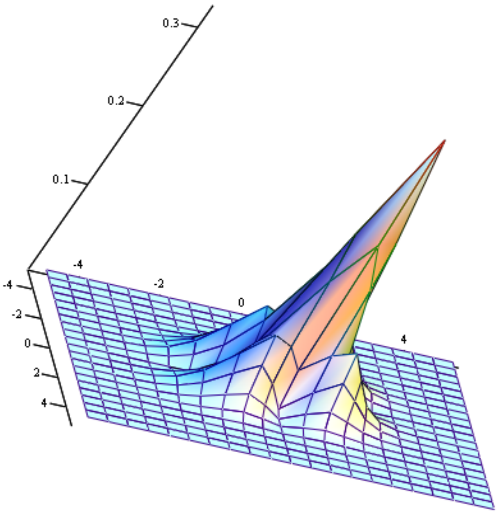
λ = −1



$\lambda = -1/2$



λ = 1/2



λ = 1

Figure 2. The density form of the probable distribution (9) for different values of the parameter λ .

For probability distribution density (9), both the correlation coefficient $r(X,Y)$, and the correlation ratio $\theta(X,Y)$ are identically zero.

In turn, in Fig. 3, the dependence of the values $\iota^T_{1/2}(X,Y)$ (5) on the parameter λ of the probability distribution density (9) is presented in comparison with the values of the maximum correlation $S(X,Y)$ coefficient (dotted line).

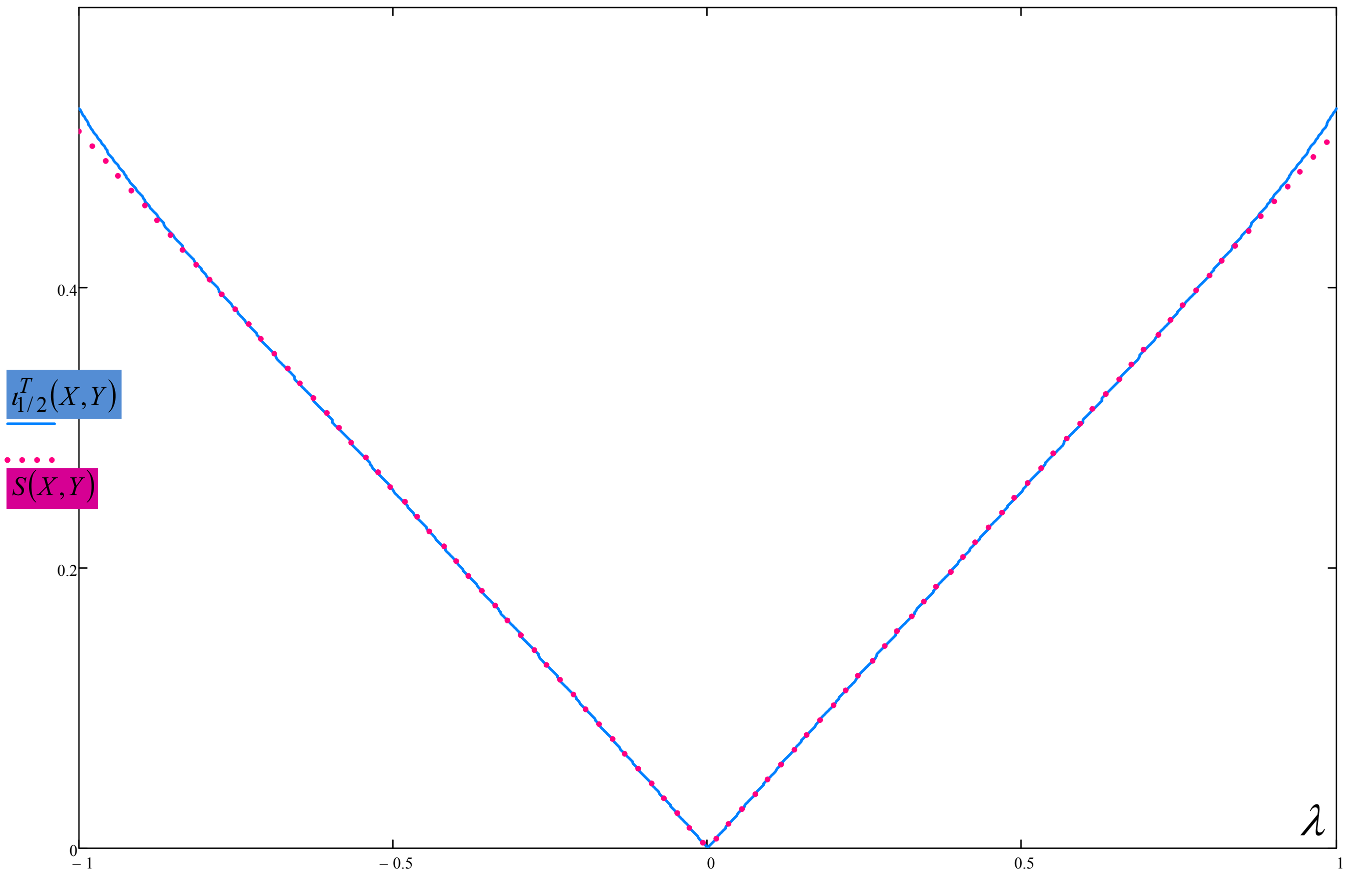


Figure 3. Comparison of quantities $\iota^T_{1/2}(X,Y)$ and $S(X,Y)$ for various values of the parameter λ in the probability distribution density (9)..

6. CONCLUSIONS

The paper has been concerned with issues of application of consistent measures of dependence, taking into account the ability of their application within stochastic signal processing, since it always assumes applying a measure of dependence, while conventional ones, based on the linear correlation, may lead to unacceptable results due to the possibility of vanishing even under the availability of deterministic dependence between random values. A procedure that enables one to construct corresponding consistent in the Rényi sense measure of dependence from a consistent in the Kolmogorov sense measure of dependence has been proposed.

In the paper, a consistent in the Kolmogorov sense measure of dependence was referred as consistent in the Rényi sense, if such a measure meets all Rényi axioms (Rényi, 1959). In particular, such a consistent in the Rényi sense measure of dependence has been constructed by us of Tsallis divergence of the order ½, which has been proposed to be applied within the problem of selecting informative indicators to predict the stability of short fragments of proteins.

REFERENCES

Balakrishnan, N., and Chi-Diw Lai (2009). *Continuous Bivariate Distributions* / Second Edition, Wiley, 2009, 714 p.

Ciarletta, P., Hillen, T., Othmer, H., Preziosi, L., and D. Trucu (2016). *Mathematical Models and Methods for Living Systems*, Springer, 332 p.

Fabian, H. and D. Naumann (Eds.) (2012). Protein Folding and Misfolding. Shining Light by Infrared Spectroscopy, Springer, 256 p.

Garfinkel, A., Shevtsov, J., and Yina Guo (2017). *Modeling Life. The Mathematics of Biological Systems*, Springer, 456 p.

Gebelein, H. (1941). “Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichungsrechnung”, *Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, no. 6, zz. 364-379.

Nekrasov, A.N., Anashkina, A.A., and A.A. Zinchenko (2014). “A new paradigm of protein structural organization”, *Theoretical Approaches to BioInformation Systems*. Institute of Physics, Belgrade, pp. 1-24.

Mikhalskii, A.I., Petrov, I.V., Tsurko, V.V., Anashkina, A.A., and A. N. Nekrasov (2020). “Application of mutual information estimation for predicting the structural stability of pentapeptides”, *Russian Journal of Numerical Analysis and Mathematical Modelling*, vol. 35, no. 5, pp. 263-271.

Rajbman, N.S. (1981). “Extensions to nonlinear and minimax approaches”, *Trends and Progress in System Identification*, ed. P. Eykhoff, Pergamon Press, Oxford, 1981, pp. 185-237.

Rangayyan, R.M. (2005). *Biomedical Image Analysis*, CRC Press, 1272 p.

Rangayyan, R.M. (2015). *Biomedical Signal Analysis*, 2nd Edition, Wiley-IEEE Press, 720 p.

Rényi, A. (1959). “On measures of dependence”, *Acta Math. Acad. Sci. Hung.*, vol. 10, no 3-4, pp. 441-451.

Sarmanov, O.V and E.K. Zakharov (1960). “Measures of dependence between random variables and spectra of stochastic kernels and matrices”, *Matematicheskiy Sbornik*, 1960, vol. 52(94), pp. 953-990. (in Russian)

Sarmanov, O.V. (1963a). “The maximum correlation coefficient (nonsymmetric case)”, *Sel. Trans. Math. Statist. Probability*, vol. 4, pp. 207-210.

Sarmanov, O.V. (1963b). “Investigation of stationary Markov processes by the method of eigenfunction expansion”, *Sel. Trans. Math. Statist. Probability*, vol. 4, pp. 245-269.

Sarmanov, O.V. (1967). “Remarks on uncorrelated Gaussian dependent random variables”, *Theory Probab. Appl.,* vol. 12, issue 1, pp. 124-126.

Tsallis, C. (2009). *Introduction to Nonextensive Statistical Mechanics. Approaching a Complex World*, Springer, 388 p.

Yanchun Tao, Yuqing Wu, and Liping Zhang (2018). “Advancements of two dimensional correlation spectroscopy in protein researches”, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 197, pp. 185-193.

Yuqing Wu, Liping Zhang, Young Mee Jung, and Y. Ozaki (2018). “Two-dimensional correlation spectroscopy in protein science, a summary for past 20 years”, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 189, pp. 291-299.

Zeng, X. and T.S. Durrani (2011). “Estimation of mutual information using copula density function”, *Electronics Letters*, vol. 47, no. 8, pp. 493-494.