


Tsallis Divergence of Order $\frac{1}{2}$ in System Identification Related Problems

Kirill Chernyshov ^a

V. A. Trapeznikov Institute of Control Sciences, 65 Profsoyuznaya Street, Moscow, Russia

Keywords: Tsallis Divergence, System Identification, Mutual Information.

Abstract: The measure of divergence and the corresponding Hellinger-Tsallis mutual information have been introduced within the information-theoretic approach to system identification based on Tsallis divergence and Hellinger distance properties for a pair of probability distributions to be used in statistical linearization problems. The introduced measure in this case is used ambivalently: as mutual information, a measure of random vector dependence, — as a criterion of statistical linearization of multidimensional stochastic systems, and as a measure of divergence of probability distributions — as an anisotropic norm of input process used to quantify the correspondence between the observable data and the assumptions of the original problem statement as such.

1 PRELIMINARIES

The measure of inequality between continuous multiple probability distributions, such as $p_1(\mathbf{z})$ and $p_2(\mathbf{z})$ of a κ -dimensional random vector \mathbf{Z} , are well known as measures of divergence, including Kullback-Leibler divergence,

$$D^{KL}(p_1 \| p_2) = \mathbf{E}_{p_1} \left\{ \ln \left(\frac{p_1(\mathbf{z})}{p_2(\mathbf{z})} \right) \right\}, \quad (1)$$

is probably the most widely known and applied tool in various problems related to stochastic system analysis. In (1) and below, $\mathbf{E}_{\varphi}\{\cdot\}$ is a mathematical expectation with respect to probability distribution φ .

Meanwhile, there are broader approaches that enable the characterization of inequalities between two probability distributions, including Kullback-Leibler divergence. In particular, this includes α order Tsallis (2009) divergence of continuous multidimensional probability distributions, e. g. $p_1(\mathbf{z})$ and $p_2(\mathbf{z})$, which is defined as follows:

$$\begin{aligned} D_{\alpha}^T(p_1 \| p_2) &= \\ &= \frac{1}{1-\alpha} \left(1 - \mathbf{E}_{p_1} \left\{ \left(\frac{p_1(\mathbf{z})}{p_2(\mathbf{z})} \right)^{\alpha-1} \right\} \right), \quad \alpha > 0, \alpha \neq 1. \end{aligned} \quad (2)$$

As is known, if parameter α tends to 1 then value $D_{\alpha}^T(p_1 \| p_2)$ in (2) tends to $D^{KL}(p_1 \| p_2)$ in (1), and, therefore, Kullback-Leibler divergence may be treated as Tsallis divergence of the order 1.


The expedience of considering Tsallis divergence in construction of an anisotropic norm is due to the fact that Kullback-Leibler divergence (1) is a marginal case of Tsallis divergence of the order α with α tending to 1. Specifically, Tsallis divergence $D_{\alpha}^T(p_1 \| p_2)$ $\alpha > 0, \alpha \neq 1$ in (2) is a straightforward generalization of Kullback-Leibler divergence in replacing the logarithm with an appropriate exponential:

$$\ell_{\alpha}(z) = \frac{z^{1-\alpha} - 1}{1-\alpha}, \quad z > 0, \alpha > 0, \alpha \neq 1.$$

At the same time,

$$\lim_{\alpha \rightarrow 1} \ell_{\alpha}(z) = \ln(z), \quad z > 0.$$

In turn, Figure 1 depicts the function $\ell_{\alpha}(z)$ behavior with certain α values compared with the conventional logarithmic function.

^a  <http://orcid.org/0000-0003-4637-6161>

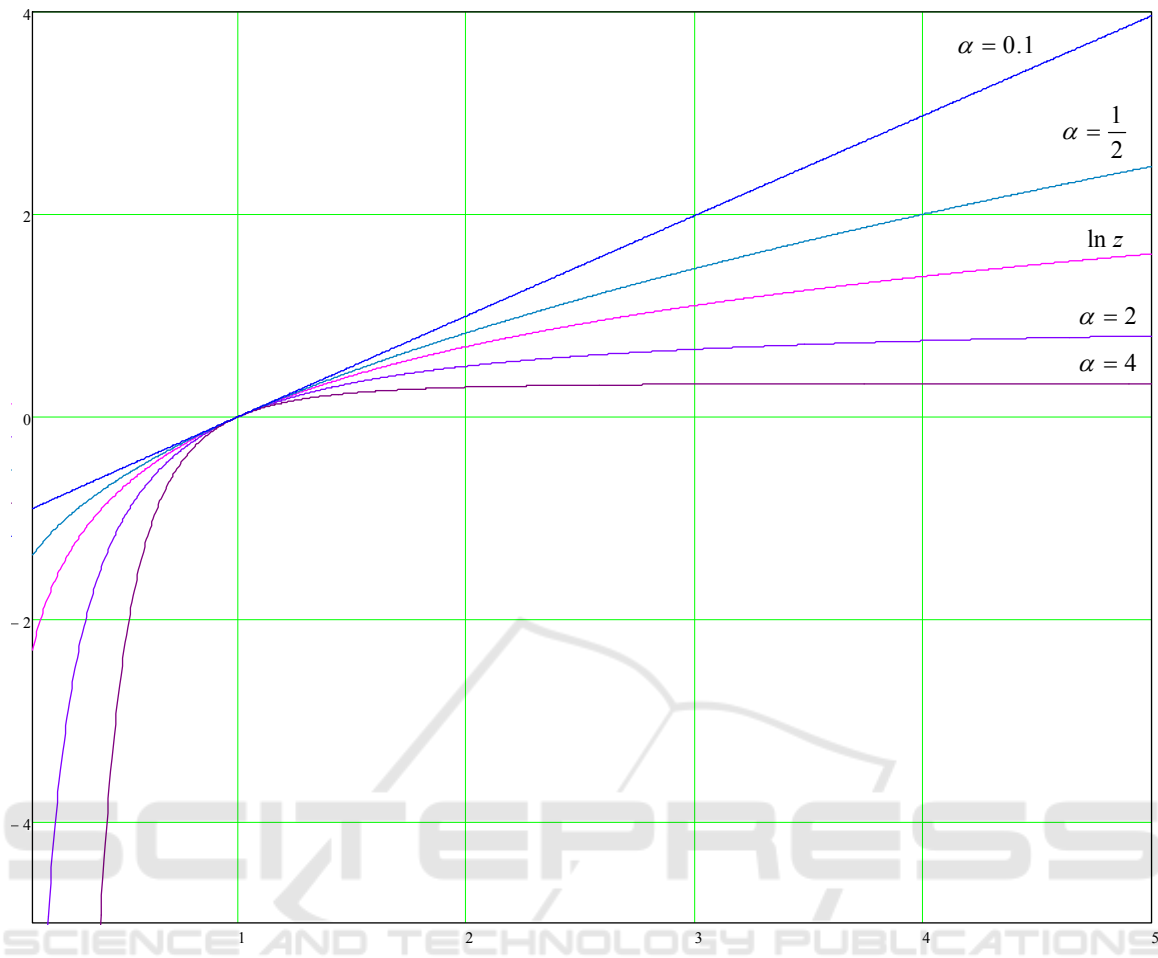


Figure 1: Plots of the logarithm and function $\ell_\alpha(z)$ under certain values of parameter α .

From the computational point of view, especially when calculations are based on sampled data, Tsallis divergence is more useful compared to Kullback-Leibler divergence as the latter includes an “integral of logarithm”, which is generally recognized as more complex for computational purposes compared to Tsallis divergence, which includes no logarithm at all. Meanwhile, the choice of a specific α order value is important as the larger it is, the more complex the computational process. On the other hand, there is only one α value that renders Tsallis divergence symmetrical relative to the probability distribution densities being compared. This value is, obviously, $\alpha = 1/2$. Consequently, its analytic expression is as follows:

$$D_{1/2}^T(p_1||p_2) = 2 \left(1 - \mathbf{E}_{p_1} \left\{ \sqrt{\frac{p_2(\mathbf{z})}{p_1(\mathbf{z})}} \right\} \right). \quad (3)$$

On the other hand, it can be noted that $D_{1/2}^T(p_1||p_2)$ is neither more nor less than double Hellinger (1907) distance between probability distributions defined as

$$H^2(p_1||p_2) = \frac{1}{2} \int_{R^\kappa} (\sqrt{p_1(\mathbf{z})} - \sqrt{p_2(\mathbf{z})})^2 d\mathbf{z}, \quad (4)$$

where $\kappa = \dim \mathbf{z}$. Based on expressions (3) and (4), Hellinger-Tsallis divergence is naturally defined as follows:

$$D^{HT}(p_1||p_2) = 1 - \mathbf{E}_{p_1} \left\{ \sqrt{\frac{p_2(\mathbf{z})}{p_1(\mathbf{z})}} \right\} = \frac{1}{2} D_{1/2}^T(p_1||p_2). \quad (5)$$

Apart from the symmetry, $D^{HT}(p_1||p_2)$ is characterized by the fact that its values fall within the unit interval. This particular case of Tsallis divergence will form the basis for further

constructions within the framework of the approaches applied in this article.

Measures of divergence may be treated as a quality criterion in the context of various theoretical and practical problems.

In particular, Hellinger-Tsallis divergence that is defined by (5) results in an expression, which can be referred as Hellinger-Tsallis $I^{HT}\{\mathbf{Z}_1, \mathbf{Z}_2\}$ mutual information for random vectors \mathbf{Z}_1 and \mathbf{Z}_2 of dimensions κ_1 and κ_2 , respectively, where one probability distribution density in $D^{HT}(p_1||p_2)$, i. e. $p_1(\mathbf{z}) = p_{\mathbf{z}_1\mathbf{z}_2}(\mathbf{z}_1, \mathbf{z}_2)$, is a joint probability distribution density of these random vectors, while the other probability distribution density, $p_2(\mathbf{z}) = p_{\mathbf{z}_1}(\mathbf{z}_1)p_{\mathbf{z}_2}(\mathbf{z}_2)$, is the product of marginal probability distribution densities of \mathbf{Z}_1 and \mathbf{Z}_2 . Similarly, the corresponding Hellinger-Tsallis divergence $D^{HT}(p_{\mathbf{z}_1\mathbf{z}_2}||p_{\mathbf{z}_1}p_{\mathbf{z}_2})$ provides an information theoretic quality criterion that may be treated as a basis for creating an identification criterion, which in turn defines the information theoretic approach to system identification:

$$\begin{aligned} D^{HT}(p_{\mathbf{z}_1\mathbf{z}_2}||p_{\mathbf{z}_1}p_{\mathbf{z}_2}) &= I^{HT}\{\mathbf{Z}_1, \mathbf{Z}_2\} = 1 - \\ &- \int_{R^{\kappa_1}} \int_{R^{\kappa_2}} \sqrt{p_{\mathbf{z}_1\mathbf{z}_2}(\mathbf{z}_1, \mathbf{z}_2)p_{\mathbf{z}_1}(\mathbf{z}_1)p_{\mathbf{z}_2}(\mathbf{z}_2)} d\mathbf{z}_1 d\mathbf{z}_2 = \\ &= 1 - \mathbf{E}_{p_{\mathbf{z}_1\mathbf{z}_2}} \left\{ \sqrt{\frac{p_{\mathbf{z}_1}(\mathbf{z}_1)p_{\mathbf{z}_2}(\mathbf{z}_2)}{p_{\mathbf{z}_1\mathbf{z}_2}(\mathbf{z}_1, \mathbf{z}_2)}} \right\}. \end{aligned} \quad (6)$$

Handling the system identification problems is always based on the use of various measures of dependence of random values as is the case with the input-output representation of the system in question or with the approach to state-space description. In the majority of cases, conventional linear correlation/covariance measures of dependence are used whose direct application results from the identification problem statement itself if the case is based on a conventional root mean square criterion. The main advantage of these measures of dependence is their usability: the feasibility of finding explicit analytic expressions to define the necessary system features and the relative simplicity of formulation of their respective estimates, including those based on the need to apply dependent observations. Nonetheless, the chief deficiency of the linear correlation-based measures of dependence is their possible vanishing even in the presence of deterministic dependence between random values (Rajbman, 1981, Rényi, 1959).

More complex, nonlinear measures of dependence are engaged to eliminate this deficiency and to solve the stochastic system identification problems. Among these measures, consistent measures of dependence represent the main priority.

In accordance with A. N. Kolmogorov's terminology, a measure of dependence for two random values is treated as consistent where it vanishes if and only if the above random values are stochastically independent.

Statistical linearization of the system input-output representation is actually related to nonlinear identification problems whose solution is largely defined by the characteristics of the input and output processes dependence within the system in question. On the other hand, the existing statistical linearization approaches are based on the application of conventional linear correlation, which, for reasons stated above, may result in construction of models with an output variable that is identically zero. In particular, the likelihood of such result is shown using an example in Section 5 in this article, which suggests an approach focused on eliminating the identified deficiencies and on applying consistent measures of dependence within the framework of system identification using linear representations of their respective input/output models. The information theoretic approach involves statement of the statistical linearization problem for multidimensional discrete-time systems.

2 PROBLEM STATEMENT

Let us assume that

$$V(t) = (v_1(t), \dots, v_n(t))^T$$

is the n -dimensional output random system process and

$$U(s) = (u_1(s), \dots, u_m(s))^T$$

is the m -dimensional input random system process within a multidimensional nonlinear dynamical stochastic system. Within the framework of the above process description, the processes $V(t)$ and $U(s)$ are treated as stationary processes or mutually stationary processes in a narrow sense. Then, the process $U(s)$ is white Gaussian noise with a covariance matrix C_U , while the dependence of input and output system processes is characterized (of which the researcher is, naturally, not aware) by probability distribution densities

$$\begin{aligned}
 & p_{v_i, u_j}(v, u; \tau), \\
 & i = 1, \dots, n, \\
 & j = 1, \dots, m \\
 & \tau = 1, 2, \dots
 \end{aligned} \tag{7}$$

For the sake of simplicity, but without loss of generality, the vector-valued process components $V(t)$ and $U(s)$ are treated with a zero mean and unit variance,

$$\begin{aligned}
 & \mathbf{E}\{v_i(t)\} = \mathbf{E}\{u_i(s)\} = 0; \\
 & \mathbf{var}\{v_i(t)\} = \mathbf{var}\{u_i(s)\} = 1, \\
 & i = 1, \dots, n, \quad j = 1, \dots, m,
 \end{aligned} \tag{8}$$

where $\mathbf{var}\{\cdot\}$ is variance. Under the above terms,

$$C_U = \begin{pmatrix} 1 & c_{12} & \dots & c_{1m} \\ c_{12} & \ddots & & \vdots \\ \vdots & & \ddots & c_{(m-1)m} \\ c_{1m} & \dots & c_{(m-1)m} & 1 \end{pmatrix}. \tag{9}$$

A corresponding linear *input/output* system model characterized by probability distribution densities (7) is sought in the following form:

$$\hat{V}(t; \mathbf{W}) = \sum_{\tau=1}^{\infty} W(\tau)U(t-\tau), \quad t = 1, 2, \dots \tag{10}$$

where

$$\hat{V}(t; \mathbf{W}) = (\hat{v}_1(t; \mathbf{W}), \dots, \hat{v}_n(t; \mathbf{W}))^T$$

is the output process of the model, $\mathbf{W} = \{W(\tau), \tau \in [1, \infty)\}$, $W(\tau), \tau = 1, 2, \dots$ are matrix-valued (dimensions $n \times m$) weight function coefficients of a linearized model to be identified as per the information theoretic criterion of statistical linearization.

This criterion is a condition for coincidence of the Hellinger-Tsallis mutual information (6) for the i^{th} component, $v_i(t)$, the system output process $V(t)$, and the j^{th} component, $u_j(s)$, the system input process $U(s)$, which are characterized by the probability distribution densities (7) and the Hellinger-Tsallis mutual information (6) the for i^{th} component, $\hat{v}_i(t; \mathbf{W})$, the output process $\hat{V}(t; \mathbf{W})$, and the j^{th} component, $u_j(s)$, the input process $U(s)$ of model (10) for all $i = 1, \dots, n, j = 1, \dots, m$. From the analytical point of view, this information-theoretic criterion is expressed as follows:

$$I^{HT}(v_i(t), u_j(s); \tau) = I^{HT}(\hat{v}_i(t; \mathbf{W}), u_j(s); \tau), \tag{11}$$

$$i = 1, \dots, n, \quad j = 1, \dots, m, \quad t - s = \tau = 1, 2, \dots$$

As regards designations in criterion (11), it should be noted that in the case of stationary and mutually stationary, in the strict sense, random processes, e. g. $z_1(t)$ and $z_2(s)$, $t - s = \tau$ Hellinger-Tsallis mutual information is the corresponding function of time $\tau = t - s$:

$$\begin{aligned}
 & I^{HT}(z_1(t), z_2(s); \tau) = \\
 & = \left(1 - \mathbf{E} p_{z_1 z_2} \left\{ \sqrt{\frac{p_{z_1}(z_1) p_{z_2}(z_2)}{p_{z_1 z_2}(z_1, z_2; \tau)}}} \right\} \right) \\
 & \tau = t - s,
 \end{aligned}$$

where $p_{z_1 z_2}(z_1, z_2; \tau)$, $p_{z_1}(z_1)$, $p_{z_2}(z_2)$ are mutual and marginal probability distribution densities of $z_1(t)$ and $z_2(s)$, $t - s = \tau$, respectively.

By all means, from the viewpoint of a statistical linearization problem, condition (11) must be supplemented with a condition of coincidence of mathematical expectations regarding the system and model output processes,

$$\mathbf{E}\{v_i(t)\} = \mathbf{E}\{\hat{v}_i(t; \mathbf{W})\} = 0, \quad i = 1, \dots, n. \tag{12}$$

It is evident that, within this problem description, condition (12) is met automatically.

Moreover, in accordance with normalization condition (8), model (10) output process components are imposed upon the unit variance condition,

$$\mathbf{var}\{\hat{v}_i(t; \mathbf{W})\} = 1, \quad i = 1, \dots, n, \tag{13}$$

and, consequently, the sequences of matrix-valued weight function coefficients of model (10) must meet the following condition:

$$\sum_{\tau=1}^{\infty} \bar{w}_i(\tau) C_U \bar{w}_i^T(\tau) = 1, \quad i = 1, \dots, n, \tag{14}$$

where

$$\bar{w}_i(\tau) = (w_{i1}(\tau), \dots, w_{im}(\tau))$$

is the i^{th} sequence of matrices $W(\tau), \tau = 1, 2, \dots$ in (10).

Relationship (14) is evidently defined by the sequence as follows:

$$\begin{aligned}
 1 &= \mathbf{var}\{\hat{v}_i(t; \mathbf{W})\} = \mathbf{var}\left\{\sum_{\tau=1}^{\infty} \bar{w}_i(\tau)U(t-\tau)\right\} = \\
 &= \sum_{\tau=1}^{\infty} \bar{w}_i(\tau)\mathbf{E}\left\{U(t-\tau)U^{\mathbf{T}}(t-\tau)\right\}\bar{w}_i^{\mathbf{T}}(\tau) + \\
 &+ \sum_{p \neq q} \bar{w}_i(p)\mathbf{E}\left\{U(t-p)U^{\mathbf{T}}(t-q)\right\}\bar{w}_i(q) = \\
 &= \sum_{\tau=1}^{\infty} \bar{w}_i(\tau)C_U\bar{w}_i^{\mathbf{T}}(\tau)
 \end{aligned}$$

based on the description of model (10) and normalization conditions (8), (9), (13).

Therefore, expressions (11) and (12) are information-theoretic criteria of statistical linearization of the system that is represented by probability distribution densities (7).

3 SOLUTION TECHNIQUE

As the first step, one can consider a sequence of random values

$$\begin{aligned}
 \zeta_i^{-\tau,t} &= \sum_{j=1}^{\tau-1} \bar{w}_i(j)U(t-j) + \sum_{j=\tau+1}^{\infty} \bar{w}_i(j)U(t-j), \\
 \tau &= 1, 2, \dots
 \end{aligned}$$

being evidently Gaussian ones with the zero mean and variance, due to (10), expressed as follows:

$$\begin{aligned}
 \mathbf{var}\left\{\zeta_i^{-\tau,t}\right\} &= \sum_{j=1}^{\tau-1} \bar{w}_i(j)C_U\bar{w}_i^{\mathbf{T}}(j) + \\
 &+ \sum_{j=\tau+1}^{\infty} \bar{w}_i(j)C_U\bar{w}_i^{\mathbf{T}}(j) = \\
 &= 1 - \bar{w}_i(\tau)C_U\bar{w}_i^{\mathbf{T}}(\tau), \quad \tau = 1, 2, \dots
 \end{aligned}$$

Then, within this set of notations, $(m + 1)$ -dimensional vector

$$Z_i(t, \tau) = \left(\zeta_i^{-\tau,t}, u_1(t-\tau), \dots, u_m(t-\tau)\right)^{\mathbf{T}}$$

is Gaussian with corresponding covariance matrix

$$C_{Z_i(t, \tau)} = \begin{pmatrix} \omega_{ij}(\tau) & 0 & \dots & \dots & 0 \\ 0 & 1 & c_{12} & \dots & c_{1m} \\ \vdots & c_{12} & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & c_{(m-1)m} \\ 0 & c_{1m} & \dots & c_{(m-1)m} & 1 \end{pmatrix},$$

$$\omega_{ij}(\tau) = 1 - \bar{w}_i(\tau)C_U\bar{w}_i^{\mathbf{T}}(\tau),$$

and the correlation for a bivariate Gaussian random vector $(\hat{v}_i(t; \mathbf{W}) \ u_j(t-\tau))^{\mathbf{T}}$ can be written as follows:

$$\begin{pmatrix} \hat{v}_i(t; \mathbf{W}) \\ u_j(t-\tau) \end{pmatrix} = \Psi_{ij}(\tau)Z_i(t, \tau),$$

where $(2 \times (m + 1))$ -dimensional matrix $\Psi_{ij}(\tau)$ is expressed as follows:

$$\begin{aligned}
 \Psi_{ij}(\tau) &= \\
 &= \begin{pmatrix} 1 & w_{i1}(\tau) & \dots & w_{ij}(\tau) & w_{i(j+1)}(\tau) & \dots & w_{im}(\tau) \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{pmatrix},
 \end{aligned}$$

and, as stated above, $w_{i1}(\tau), \dots, w_{im}(\tau)$ are the elements of the i^{th} row of matrices $W(\tau)$ from (10).

Hence, random vector $(\hat{v}_i(t; \mathbf{W}) \ u_j(t-\tau))^{\mathbf{T}}$ is Gaussian with corresponding covariance matrix $C_{(\hat{v}_i, u_j)}(\tau)$ expressed as follows:

$$C_{(\hat{v}_i, u_j)}(\tau) = \Psi_{ij}(\tau)C_{Z_i(t, \tau)}\Psi_{ij}^{\mathbf{T}}(\tau). \quad (15)$$

Calculation of the right side of (15) results in

$$C_{(\hat{v}_i, u_j)}(\tau) = \begin{pmatrix} 1 & \omega_{ij}(\tau) \\ \omega_{ij}(\tau) & 1 \end{pmatrix}, \quad (16)$$

where $\omega_{ij}(\tau)$ is the j^{th} component of the column vector $C_U\bar{w}_i^{\mathbf{T}}(\tau)$.

Therefore, formula (6), above reasoning, and formula (16) suggest that Hellinger-Tsallis mutual information (6) $I^{HT}(\hat{v}_i(t; \mathbf{W}), u_j(s); \tau)$ on input and output model processes (10) (in other words, bivariate Gaussian random vector $(\hat{v}_i(t; \mathbf{W}) \ u_j(t-\tau))^{\mathbf{T}}$ with covariance matrix defined by (11)) is expressed as

$$\begin{aligned}
 I^{HT}(\hat{v}_i(t; \mathbf{W}), u_j(s); \tau) &= \\
 &= 1 - 2 \frac{\sqrt{\det(C_{(\hat{v}_i, u_j)}(\tau))}}{\sqrt{3 + \det(C_{(\hat{v}_i, u_j)}(\tau))}}, \quad t - s = \tau = 1, 2, \dots,
 \end{aligned}$$

which, in turn, based on condition (11), results in

$$I^{HT}(v_i(t), u_j(s); \tau) = 1 - 2 \frac{\sqrt{1 - \omega_{ij}^2(\tau)}}{\sqrt{4 - \omega_{ij}^2(\tau)}}, \quad (17)$$

$$t - s = \tau = 1, 2, \dots$$

Expression (17), in turn, directly results in

$$I^{HT}(v_i(t), u_j(s); \tau) = \frac{2 \sqrt{\left((I_{ij}(\tau))^2 + \sqrt{1 - 3(I_{ij}(\tau))^2 - 1} - 2 \right)}}{I_{ij}(\tau)}, \quad (18)$$

$$I_{ij}(\tau) = \left(I^{HT}(v_i(t), u_j(s); \tau) - 1 \right)^2,$$

$$t - s = \tau = 1, 2, \dots$$

Then, based on condition (11), the required expressions for rows $\bar{w}_i(\tau)$ of the matrix-valued weight function coefficients $W(\tau), \tau = 1, 2, \dots$ of model (10) appear as follows:

$$\bar{w}_i^T(\tau) = C_U^{-1} \mathbf{I}^{HT}(v_i(t), U(s); \tau), \quad (19)$$

$$t - s = \tau = 1, 2, \dots,$$

where

$$\mathbf{I}^{HT}(v_i(t), U(s); \tau) = \begin{pmatrix} \text{sign}(\mathbf{m}_{v_i|u_1}(\tau)) \times I^{HT}(v_i(t), u_1(s); \tau) \\ \vdots \\ \text{sign}(\mathbf{m}_{v_i|u_j}(\tau)) \times I^{HT}(v_i(t), u_j(s); \tau) \\ \vdots \\ \text{sign}(\mathbf{m}_{v_i|u_m}(\tau)) \times I^{HT}(v_i(t), u_m(s); \tau) \end{pmatrix}, \quad (20)$$

$$t - s = \tau = 1, 2, \dots$$

In formulas (19) and (20) $\mathbf{m}_{v_i|u_j}(\tau)$ is a regression of $v_i(t)$ on $u_j(t - \tau)$; $\text{sign}(x) = 1$ as $x \geq 0$, $\text{sign}(x) = -1$ as $x < 0$ is a corresponding regression function sign that represents “relative orientation” of input and output process components, while value $I^{HT}(v_i(t), u_j(s); \tau)$ in (19), (20) is always non-negative.

Moreover, the measure of dependence $I^{HT}(v_i(t), u_j(s); \tau)$ defined by expressions (6) and (18) meets all Rényi (1959) axioms for measures of random values dependence. Meanwhile, the calculations here are much simpler than in the case of the maximum correlation coefficient (Rényi (1959, Sarmanov, 1963a,b).

The behavior of measure of dependence (18) as a function of $I^{HT}(v_i(t), u_j(s); \tau)$ is shown in Figure 2.

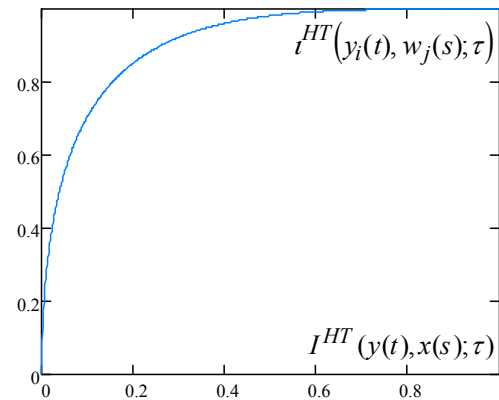


Figure 2: Behavior of measure of dependence $I^{HT}(v_i(t), u_j(s); \tau)$ (18) as a function of $I^{HT}(y_i(t), w_j(s); \tau)$.

Therefore, vanishing of the weight function coefficients of linearized model (10) within a nonlinear system characterized by probability distribution densities (7) is equivalent to vanishing of Hellinger-Tsallis mutual information (6) on input and output processes within the system in question. The latter, in turn, is possible if and only if these processes are stochastically independent. The direct consequence of the above is that vanishing of all weight function coefficients of linearized model (10) indicates that the original nonlinear system is unidentifiable. Meanwhile, as stated above, there are examples of traditional measures of dependence vanishing if there is stochastic dependence between the system variables.

4 ZERO CORRELATION OF INPUT AND OUTPUT VARIABLES: EXAMPLE

There are multiple examples where the use of conventional correlation methods within the framework of the models obtained fails to provide satisfactory results. Among such systems, it is possible to distinguish those where input and output processes dependence is characterized by probability distribution densities that belong to O. V. Sarmanov distribution class (Sarmanov, 1967, Kotz et al., 2000) and expressed as

$$p_{v,u;\lambda}(v, u) = p_u(u)p_v(v)(1 + \lambda \phi_1(u)\phi_2(v)), \quad (21a)$$

with marginal probability distribution densities $p_u(u)$ and $p_v(v)$,

$$\int p_u(u)\phi_1(u)du = 0, \quad \int p_v(v)\phi_2(v)dv = 0, \quad (21b)$$

where parameter λ meets the condition:

$$1 + \lambda \phi_1(u)\phi_2(v) \geq 0. \quad (21c)$$

Correlation coefficient and correlation ratio for probability distribution densities (21) are equal to zero. Let's consider the following probability distribution density that belongs to O. V. Sarmanov distribution class:

$$p_\lambda(v,u) = \frac{e^{-\frac{v^2+u^2}{2}}}{2\pi} \left(1 + \lambda \left(2e^{-\frac{3}{2}v^2} - 1 \right) \left(2e^{-\frac{3}{2}u^2} - 1 \right) \right), \quad (22)$$

$$-1 \leq \lambda \leq 1.$$

Its marginal probability distribution densities are Laplacian ones.

Meanwhile, the maximum correlation coefficient for probability distribution density (22), hereinafter designated as $S_{vu}(\lambda)$, is expressed as follows:

$$S_{vu}(\lambda) = \left(\frac{4}{\sqrt{7}} - 1 \right) |\lambda|.$$

The value of parameter λ has a significant impact on the form of probability distribution density (22). Figure 3 depicts probability distribution density (22) for certain values of parameter λ .

Let us assume that in (7), the joint probability distribution density is $p_\lambda(v,u)$ of (22). Then, the Hellinger-Tsallis mutual information for the probability density function $p_\lambda(v,u)$ may be a corresponding function of parameter λ to be designated as $I^{HT}(\lambda)$. Consequently, measure of dependence (18) for probability distribution density $p_\lambda(v,u)$ (22) will be designated as $\iota^{HT}(\lambda)$, respectively.

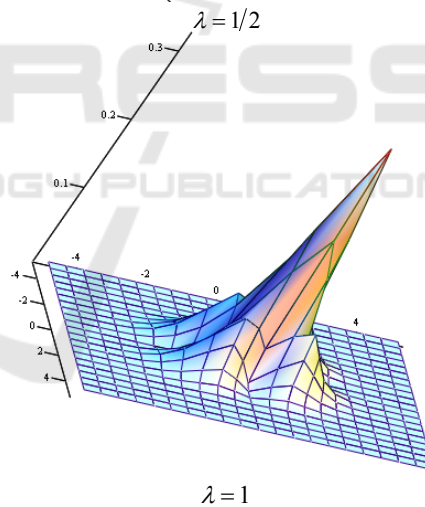
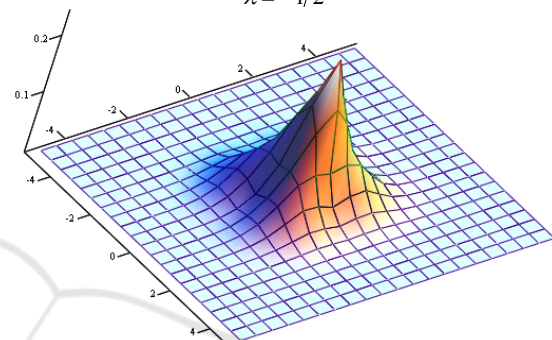
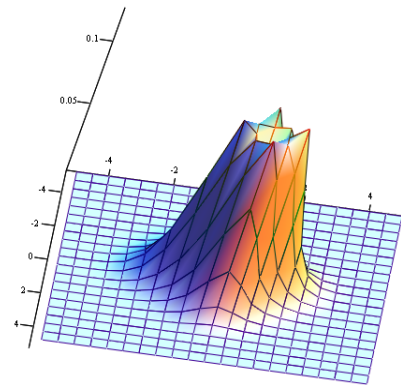
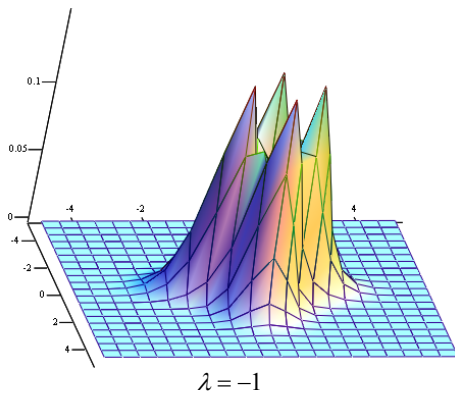


Figure 3: Shape of probability distribution density (22) for certain values of parameter λ .

Figure 4 depicts the behavior of $\iota^{HT}(\lambda)$ as a function of parameter λ of probability distribution density (22) compared with the corresponding values of the maximum correlation coefficient $S_{vu}(\lambda)$. $\iota^{HT}(\lambda)$ clearly shows the dependence between random values, which basically matches (formally, even to a greater extent) the maximum correlation.

For example, if stochastic dependence (7) between the components of the output process, $v_i(t)$, and the input process, $u_j(s)$, of a nonlinear stochastic system

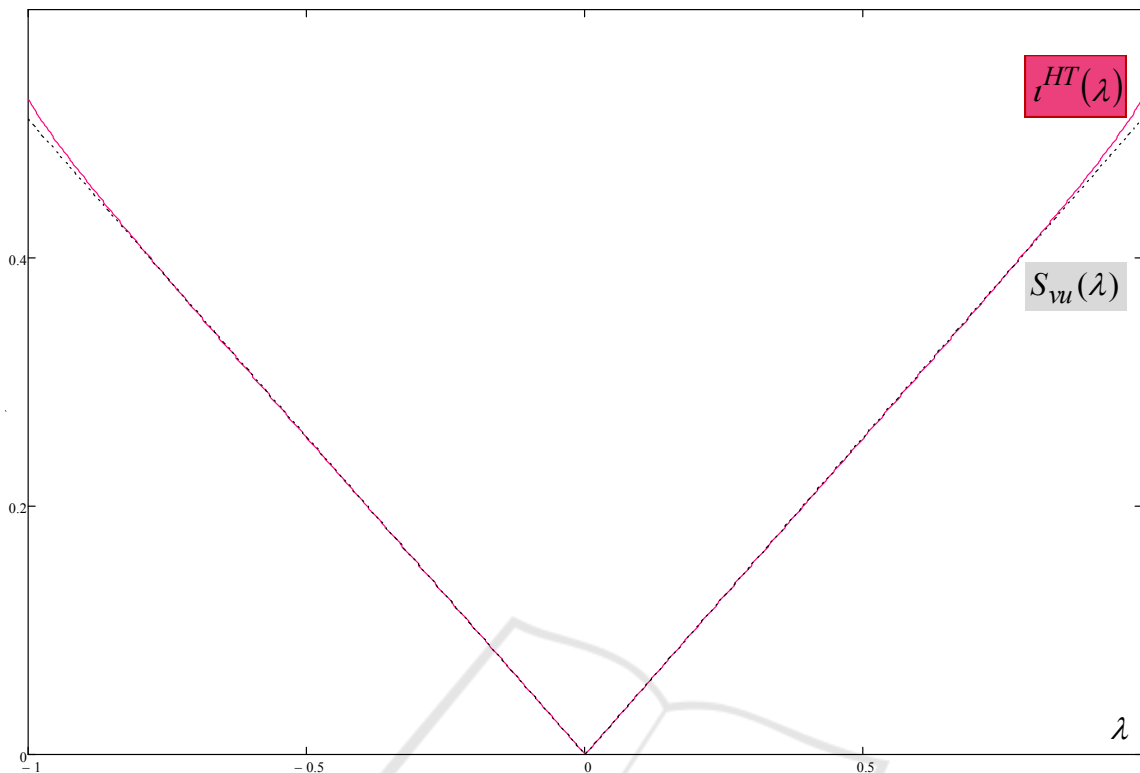


Figure 4: Behavior of $t^{HT}(\lambda)$ (unbroken line) as a function of parameter λ of probability distribution density (22) compared with the corresponding values of the maximum correlation coefficient $S_{vu}(\lambda)$ (broken line).

is defined by a probability distribution density (of which the researcher is, naturally, not aware) of type (21) with parameter $\lambda = \lambda_{ij}(\tau), \tau = t - s$, then the use of conventional correlation methods within the framework of model (10) being constructed will result in a representation of the output model process as a null equation, which is excluded within the framework of the proposed information-theoretic approach.

5 HELLINGER-TSALLIS MUTUAL INFORMATION ESTIMATION

As far as the problem of obtaining estimates for weight function coefficients (19) of linearized model (10) by using the data from sample observation of the input and output process values in a system characterized by joint probability distribution densities (7) is concerned, a need arises for a corresponding estimation of the Hellinger-Tsallis mutual information (6); this type of problems allows for a direct application of the Sklar (1959) theorem

regarding the representation of joint probability distribution densities through their copula functions. In particular, for joint probability distribution density $p_{vu}(v, u)$ of random values V and U with corresponding marginal probability distribution densities $p_v(v), p_u(u)$, the following expansion is valid:

$$p_{vu}(v, u) = c(P_v(v), P_u(u))p_v(v)p_u(u), \quad (23)$$

where

$$P_v(v) = \int_{-\infty}^v p_v(y)dy, \quad P_u(u) = \int_{-\infty}^u p_u(x)dx$$

are the functions of marginal probability distribution densities of random values V and U , and $c(P_v(v), P_u(u))$ is the copula density function (for copulas, refer to book of Nelsen (2006) and other sources).

In accordance with representation (23), Hellinger-Tsallis mutual information (6) is expressed as follows:

$$\begin{aligned}
 I^{HT}(v_i(t), u_j(s); \tau) &= 1 - \mathbf{E} \left\{ \sqrt{\frac{P_v(v)P_u(u)}{P_{vu}(v,u; \tau)}} \right\} = \\
 &= 1 - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sqrt{c(P_v(v), P_u(u); \tau)} p_v(v) p_u(u) dv du = (24) \\
 &= 1 - \int_0^1 \int_0^1 \sqrt{c(P_v(v), P_u(u); \tau)} dP_v(v) dP_u(u).
 \end{aligned}$$

Representation (24) enables the application of the mutual information estimation (according to Shannon type) (Zeng and Durrani, 2011). Meanwhile, an example of Hellinger-Tsallis mutual information shown in formula (24) within the context under consideration is even more simple since the copula density function based on representation (23) in the case of Shannon mutual information still includes the logarithm of the copula density function:

$$c(P_v(v), P_u(u); \tau) \ln(c(P_v(v), P_u(u); \tau)).$$

Meanwhile, complications inherent in the division operations may be avoided through the use of methods for copula density function estimation instead of probability density function estimation.

6 ANISOTROPIC NORM BASED ON THE HELLINGER-TSALLIS DIVERGENCE

The approach in question is based on the assumption that the m -dimensional process of the input system $U(s)$ is 1) white noise, and 2) is Gaussian. These assumptions may be checked by using an anisotropic norm as a quantitative measure. Anisotropic norm for a random vector was shown in (Vladimirov et al., 1995, 1999) based on the Kullback-Leibler divergence, which automatically results in its nonsymmetry.

Namely, the definition of the anisotropic norm for random m -dimensional vector \mathbf{U} with covariance matrix C and probability distribution density $p_U(U)$, which is based on Kullback-Leibler divergence, and is expressed as follows:

$$\|\mathbf{U}\|_a = \min_{\nu > 0} \int_{R^m} p_U(U) \ln \frac{p_U(U)}{G_\nu(U)} dU,$$

where $G_\nu(U)$ is the probability distribution density of an n -dimensional Gaussian random vector with a so-called a scalar covariance matrix $\nu \cdot \mathbf{I}_m$, where \mathbf{I}_m is

a unit $m \times m$ -matrix:

$$G_\nu(U) = \frac{1}{\sqrt{(2\pi\nu)^m}} e^{-\frac{\|\mathbf{U}\|^2}{2\nu}}.$$

The key feature and benefit of using Kullback-Leibler divergence to define the anisotropic norm $\|\mathbf{U}\|_a$ in that form is the possibility to solve *explicitly* the optimization problem that is behind definition of the anisotropic norm $\|\mathbf{U}\|_a$, since such a solution is determined by equation

$$\frac{d}{d\nu} \int_{R^m} p_U(U) \ln \frac{p_U(U)}{G_\nu(U)} dU = 0,$$

where optimum value ν is expressed as

$$\nu = \frac{\mathbf{E}(\|\mathbf{U}\|^2)}{m}.$$

Under this value of ν , the above definition of the anisotropic norm $\|\mathbf{U}\|_a$ immediately takes on its closed form

$$\|\mathbf{U}\|_a = \frac{m}{2} \ln \left(2\pi e \left(\frac{\mathbf{E}(\|\mathbf{U}\|^2)}{m} \right) \right) - S_C(\mathbf{U}),$$

where $S_C(\mathbf{U})$ is Shannon entropy of m -dimensional random vector \mathbf{U} with the covariance matrix C :

$$S_C(\mathbf{U}) = - \int_{R^m} p_U(U) \ln(p_U(U)) dU.$$

It should be noted that this simple solution is achieved solely through introducing logarithm of exponent in definition of the anisotropic norm $\|\mathbf{U}\|_a$.

Again, definition of Vladimirov et al. (1995, 1999) is based on a comparison between this Gaussian probability distribution density of random vectors with a scalar covariance matrix. The latter may be treated as an excessive limitation. In turn, Chernyshov (2018) has proposed an approach to define anisotropic norms, which would be both symmetrical and vanishing for any Gaussian vector with a particular focus on Hellinger-Tsallis divergence as naturally assuming unit interval values by default. Therefore, such anisotropic norm of an m -dimensional random vector $U = U(s)$ with a probability distribution density $p_U(U)$ with covariance matrix C_U is expressed as

$$\|U(s)\|_a^{HT} = D^{HT} \left(p_U \|G^{C_U}\right), \quad (25)$$

where G^{C_U} is the probability distribution density for Gaussian m -dimensional vector with the same covariance matrix C_U . It is evident that $\|U(s)\|_a^{HT} = 0$ for the Gaussian property of $U(s)$.

Then, in order to characterize Gaussian and white noise properties (mutual independence) typical of an (infinite) sequence of signals, the mean anisotropic norm (Vladimirov et al., 2006) is defined by a corresponding anisotropic norm for a single random vector. In other words, assume

$$U(s), \quad s = 1, 2, \dots \quad (26)$$

with a sequence of m -dimensional random vectors, and

$$U_N = \left(U^T(1), \dots, U^T(N) \right)^T. \quad (27)$$

Then the mean anisotropic norm for sequence (26) is defined by Vladimirov et al. (2006) as

$$\|U\|_a = \lim_{N \rightarrow \infty} \frac{\|U_N\|_a}{N}, \quad (28)$$

with $\|U_N\|_a$ being understood in terms of definition of Vladimirov et al. (1995, 1999).

In turn, as per definition (25), the mean anisotropic norm for vector sequence (26) is naturally defined as

$$\|U\|_a^{HT} = \lim_{N \rightarrow \infty} \|U_N\|_a^{HT}, \quad (29)$$

where

$$\|U_N\|_a^{HT} = D^{HT} \left(p_{U_N} \|G^{C(U_N^G)}\right). \quad (30)$$

In turn, p_{U_N} in (30) designates the probability distribution density for $(m \cdot N)$ -dimensional random vector (27), and U_N^G is a $(m \cdot N)$ -dimensional Gaussian random vector with covariance matrix

$$C(U_N^G) = \begin{pmatrix} C_U & \mathbf{0}_{m \times m} & \dots & \mathbf{0}_{m \times m} \\ \mathbf{0}_{m \times m} & C_U & & \vdots \\ \vdots & & \ddots & \mathbf{0}_{m \times m} \\ \mathbf{0}_{m \times m} & \dots & \mathbf{0}_{m \times m} & C_U \end{pmatrix}, \quad (31)$$

where, as above, C_U is covariance matrix $U(s)$, $s = 1, 2, \dots$ in (26).

Therefore, we may arrive at the conclusion that if the input process $U(s)$, $s = 1, 2, \dots$ meets the

conditions of statistical linearization problem statement, the maximum possible value ν of the mean anisotropic norm for the input process $U(s)$, $s = 1, 2, \dots$, defined in (29)-(31), must be set. In other words, if

$$\|U\|_a^{HT} \leq \nu, \quad (32)$$

then the input process $U(s)$ $s = 1, 2, \dots$ meets the conditions of the original problem statement, otherwise, the conditions are not met. Meanwhile, if a type (32) condition is introduced, it is essential that the anisotropic norm is defined by Hellinger-Tsallis divergence and its values fall within the unit interval.

This approach is by all means purely theoretic as it suggests that the corresponding probability distribution density

$$p_{U_N} = p_{U_N} \left(U^T(1), \dots, U^T(N) \right)$$

is known for any value of N , which is usually not the case, and the value of $\|U\|_a^{HT}$ is to be found by sample observation of the input system $U(s)$ process. In practice, cases when such direct estimation is possible are rare due to an enormous volume of the sampled data to be handled given that N is relatively large.

On the other hand, within the framework of definitions (29)-(31), condition (32) is essentially a test for both white noise (mutual independence) and Gaussian properties. Thus, condition (32) will be met if for any *two* input vectors in the system, e. g. for $U(i)$ and $U(j)$, in particular case (30), i. e. for type $\|U_2\|_a^{HT}$, the following condition is met:

$$\|U_2\|_a^{HT} = D^{HT} \left(p_{U_2} \|G^{C(U_2^G)}\right) \leq \nu, \quad (33)$$

where

$$U_2 = \left(U^T(i), U^T(j) \right)^T. \quad (34)$$

Condition (33) is, naturally, more strict than (32); however, it is much easier to check. Therefore, Parzen-Rosenblatt kernel density estimates and respective methods based on approach (Mokkadem, 1989) to estimate Shannon mutual information are used to build estimates of $\|U_2\|_a^{HT}$ in (33) via sample observation of $2m$ -dimensional random vector (34). Meanwhile, regarding Hellinger-Tsallis divergence the estimation procedure becomes considerably simpler than that of Shannon mutual information case namely due to the absence of the necessity to make the limit transfer concerned with presence of the integral of logarithm in Shannon mutual information, with

being absent the logarithm in Hellinger-Tsallis divergence.

7 CONCLUSIONS

This paper treats the problem of statistical linearization for nonlinear multidimensional dynamical stochastic systems described by input-output representation with an input process of a Gaussian white noise type as a construction of equivalent linear input-output model as per the information-theoretic criterion based on Hellinger-Tsallis mutual information (6). The latter resulted in equations enabling the determination of the linearized model weight matrix elements, which define them as a function of Hellinger-Tsallis mutual information, while vanishing of mutual information is equivalent to vanishing of the respective weight matrix elements. Meanwhile, this is equivalent to independence of the respective components of the input and output processes of the initial system under study, which, in turn, is indicative of the identifiability of such a system.

REFERENCES

- Chernyshov, K. R., 2018. "The Anisotropic Norm of Signals: Towards Possible Definitions", *IFAC-PapersOnLine*, vol. 51, no. 32, pp. 169-174.
- Hellinger, E. D., 1907. "Die Orthogonalinvarianten quadratischer Formen von unendlich vielen Variablen", *Thesis of the university of Göttingen*, 84 p.
- Kotz, S., Balakrishnan, N., and N. L. Johnson, 2000. *Continuous Multivariate Distributions. Volume 1. Models and Applications* / Second Edition, Wiley, New York, 752 p.
- Mokkadem, A., 1989. "Estimation of the entropy and information of absolutely continuous random variables", *IEEE Transactions on Information Theory*, vol. IT-35, pp. 193-196.
- Nelsen, R. G., 2006. *An Introduction to Copulas* / Second Edition, Springer Science+Business Media, New York, 2006, 274 p.
- Rajbman, N. S., 1981. "Extensions to nonlinear and minimax approaches", *Trends and Progress in System Identification*, ed. P. Eykhoff, Pergamon Press, Oxford, pp. 185-237.
- Rényi, A. 1959. "On measures of dependence", *Acta Math. Hung.*, vol. 10, no. 3-4, pp. 441-451.
- Sarmanov, O. V., 1963a. "The maximum correlation coefficient (nonsymmetric case)", *Sel. Trans. Math. Statist. Probability*, vol. 4, pp. 207-210.
- Sarmanov, O. V., 1963b. "Maximum correlation coefficient (symmetric case)", *Select. Transl. Math. Stat. Probab.* vol. 4, pp. 271-275.
- Sarmanov, O. V., 1967. "Remarks on uncorrelated Gaussian dependent random variables", *Theory Probab. Appl.*, vol. 12, no. 1, pp. 124-126.
- Sklar, A., 1959. "Fonctions de répartition à n dimensions et leurs marges", *Publ. Inst. Statist. Univ. Paris* 8, pp. 229-231.
- Tsallis, C., 2009. *Introduction to Nonextensive Statistical Mechanics. Approaching a Complex World*, Springer Science+Business Media, New York, 388 p.
- Vladimirov, I. G., Diamond, P., and P. Kloeden, 2006. "Anisotropy-based robust performance analysis of finite horizon linear discrete time varying systems", *Automation and Remote Control*, vol. 67, no. 8, pp. 1265-1282.
- Vladimirov, I. G., Kurdjukov, A. P., and A. V. Semyonov, 1995. "Anisotropy of Signals and the Entropy of Linear Stationary Systems", *Doklady Math.*, vol. 51, pp. 388-390.
- Vladimirov, I. G., Kurdjukov, A. P., and A. V. Semyonov, 1999. "Asymptotics of the anisotropic norm of linear discrete-time-invariant systems", *Automation and Remote Control*, vol. 60, no. 3, pp. 359-366.
- Zeng, X. and T. S. Durrani, 2011. "Estimation of mutual information using copula density function", *Electronics Letters*, vol. 47, no. 8, pp. 493-494.