A Novel Information-Theoretic Approach to System Identification

Kirill Chernyshov
V.A. Trapeznikov Institute of Control Sciences
Moscow, Russia
e-mail: myau@ipu.ru

ABSTRACT
The aim of the paper is to present a general approach to the identification of nonlinear stochastic systems based on information-theoretic measures of dependence. In the paper, an identification problem statement using an information-theoretic criterion under rather general conditions is proposed. It is based on a parameterized description of the model of a system under study. Such a problem statement leads finally to a problem of the finite dimensional optimization. As a result, a constructive procedure of the model parameter identification is derived. It possesses a high level of generality and does not involve unrealistic a priori assumptions that degenerate the entity of the initial identification problem statement like those ones presented in some referenced literature sources and revised in the present paper.

Keywords
Entropy, Gaussian distributions, Information theory, Jensen-Tsallis divergence, Joint probability, System Identification

1. PRELIMINARIES
Measures of the comparison of continuous multivariate probabilistic distributions, say $g_1(z)$ and $g_2(z)$, $z \in R^{\nu}$, are well known as measures of divergence, among which the Kullback-Leibler divergence

$$D_{KL}\left(g_1 \| g_2\right) = \int\limits_{R^{\nu}} g_1(z) \ln\left(\frac{g_1(z)}{g_2(z)}\right) dz$$

is, perhaps, the most widely known and applicable. Measures of divergence may be considered as a performance index within various theoretical and practical problems. In particular, the Kullback-Leibler divergence leads to the expression for the Shannon mutual information $I\{X_1, X_2\}$ of, as $\nu = 2$, two random values X_1 and X_2, when one probability distribution density in $D_{KL}\left(g_1 \| g_2\right)$, namely $g_1(z)$, is the joint probability distribution density $p_{12}(x_1, x_2)$ of these random values, and the second one, $g_2(z)$, is the product of the marginal distribution densities of X_1: $p_1(x_1)$, and X_2: $p_2(x_2)$. Accordingly, the corresponding Kullback-Leibler divergence $D_{KL}\left(p_{12} \| (p_1 p_2)\right)$ leads to the information-theoretic performance index that may be considered as a basis for constructing a system identification criterion defining thus the information-theoretic system identification approach:

$$D_{KL}\left(p_{12} \| (p_1 p_2)\right) = I\{X_1, X_2\} =$$
$$= \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} p_{12}(x_1, x_2) \ln \frac{p_{12}(x_1, x_2)}{p_1(x_1) p_2(x_2)} dx_1 dx_2 =$$
$$= \mathbf{E}\left(\ln \frac{p_{12}(x_1, x_2)}{p_1(x_1) p_2(x_2)} \right),$$

where $\mathbf{E}(\cdot)$ stands for the mathematical expectation.

Regarding namely the system identification, in paper [1] the identification problem statement is restricted by consideration of the class of linear Gaussian systems and naturally leads to applying the following relationship for the mutual information $I_{Gauss}(Y, X)$ of the multivariate Gaussian distribution

$$I_{Gauss}(Y, X) = -\frac{1}{2} \ln\left(\frac{\det(Q_{\mathbf{ZZ}})}{\det(Q_{\mathbf{YY}}) \det(Q_{\mathbf{XX}})} \right). \qquad (1)$$

In formula (1), the following notations were used: **Z** stands for the Gaussian random vector with the covariance matrix $Q_{\mathbf{ZZ}}$, $\dim \mathbf{Z} = n + m$, with $\mathbf{Z} = \left(\mathbf{X}^T \ \mathbf{Y}^T\right)^T$, where $\dim \mathbf{X} = n$, $\dim \mathbf{Y} = m$, and $Q_{\mathbf{XX}}$, $Q_{\mathbf{YY}}$ are the covariance matrices of the random vectors **X** and **Y** respectively. In turn, the aim of paper [1] was to demonstrate an equivalence of a number of criteria of identification and control for linear Gaussian systems.

Papers [2-5], thesis [6], and tutorial [7] consider the Shannon mutual information $I\{y(t), y_M(t)\}$ of system output process $y(t)$ and model output process $y_M(t)$ as an identification criterion to derive the required model. Such a criterion is to be directly maximized, and the output model variable is just considered as the maximization argument: $I\{y(t), y_M(t)\} =$

$= \mathbf{E}\left\{ \log\left(\frac{p_{SM}(y, y_M)}{p_S(y) p_M(y_M)} \right) \right\} \to \max\limits_{y_M}$, where $p_{SM}(y, y_M)$ is the joint probability distribution density of the output system process $y(t)$ and output model process $y_M(t)$, and $p_S(y)$, $p_M(y_M)$ are corresponding marginal probability distribution densities of these processes.

The approach proposed in [2-7] cannot be considered as a constructive one, because it is initially based on either a requirement that the joint probability distribution density $p_{SM}(y, y_M)$ of the system output process $y(t)$ and model output process $y_M(t)$ is to be preliminary known, or the above system and model output processes may be observed.

But both these assumptions cannot be actual. In fact, one may advert to a widely used representation of the stochastic system identification criterion in the form $\mathbf{E}\left\{\rho\left[y(t), y_M(t)\right]\right\} \to \underset{y_M(t)}{ext}$, where ρ is a priori given loss function. Then, in [2-7] such a loss function ρ is just not given, since, for the case, it is of the form $\rho\left[y(t), y_M(t)\right] = \log\frac{p_{SM}(y, y_M)}{p_S(y))p_M(y_M)}$ and involves both marginal, $p_S(y)$, $p_M(y_M)$, and (what is of special importance) *joint*, $p_{SM}(y, y_M)$, probability distribution densities of the system and model output processes respectively.

At the same time, the fact, that this joint probability distribution density $p_{SM}(y, y_M)$ is initially known within the problem statement, assumes such an amount of a priori knowledge, under which the identification problem is already to lose its sense at all: the joint distribution of the system and model output processes is a final result of influence of many factors (system and model structure, statistical properties of the input processes, etc.). In particular, one can write the following formal expression for the joint probability distribution density $p_{SM}(Y, Y_M)$ of the system and model output variables, which is implied by the relationship for the joint probability distribution density of a transformation of a random vector [8]:

$$p_{Y,Y_M}(Y, Y_M) = \int\limits_{(z_{n+1},\varphi(X_1,\ldots,X_n))\in C}\cdots\int C(S_{n-1})dS_{n-1},$$

where

$$C(S_{n-1}) = \frac{p_{X_1,\ldots,X_n,Y}(z_1,\ldots,z_n,z_{n+1})}{\sqrt{\sum\limits_{i_1<i_2}\left[\frac{D(z_{n+1},\varphi)}{D(z_{i_1},z_{i_2})}\right]^2}}.$$

The above formula is written for the system model represented as $Y_M = \varphi(X_1,\ldots,X_n)$, where $X_1,\ldots,X_n$ are the (generalized) system input variables, Y is the system output variable, $p_{X_1,\ldots,X_n,Y}(z_1,\ldots,z_n,z_{n+1})$ is the joint probability distribution density of the system input and output variables. In the right hand side, the integration is over the $(n-1)$-dimensional surface determined by the system of equations

$$\begin{cases}\varphi(z_1,\ldots,z_n) = Y_M\\ z_{n+1} = Y\end{cases},$$

and

$$\frac{D(z_{n+1},\varphi)}{D(z_{i_1},z_{i_2})} = \begin{vmatrix}\frac{\partial z_{n+1}}{\partial z_{i_1}} & \frac{\partial z_{n+1}}{\partial z_{i_2}}\\ \frac{\partial\varphi}{\partial z_{i_1}} & \frac{\partial\varphi}{\partial z_{i_2}}\end{vmatrix}$$

is the Jacobean of the functions z_{n+1},φ over the variables z_{i_1}, z_{i_2}.

However, just postulating a concrete kind of the joint probability distribution density of the output variables of system and model has been used as a basis for analytical inferences in [2-7], which assume the joint probability distribution of the system and model output processes to be Gaussian, what directly leads the initial identification problem to the problem of maximizing the correlation coefficient of the output processes of the system and model. From a substantial point of view, the assumption that the joint probability distribution of output variables of the system and model is Gaussian is equivalent to that, for instance, if a new method of matrix inversion is proposed followed by an assumption that the matrix subject to inversion should be diagonal.

In particular, from the two above formulae the well-known fact follows that the joint probability distribution of the system and model output variables is Gaussian, if the distribution of $p_{X_1,\ldots,X_n,Y}(z_1,\ldots,z_n,z_{n+1})$ is Gaussian, and the function $\varphi(X_1,\ldots,X_n)$ describing the system model is linear. So, in any more general case there is no basement for a priori assumption the joint probability distribution of the system and model output processes to be Gaussian. Such an assumption would be just an artificial simplification of the initial identification problem statement, leading to emasculation of its entity.

It should also be noted that the assumption that the joint probability distribution of system and model output processes is Gaussian is always not valid, for instance, under identification of the identity transformer. In fact, let the input X have the standard Gaussian distribution, i.e., $P\{X < x\} = \Phi(x)$, the system output variable $Y \equiv X$; the model output variable $Y_M \equiv X$; the joint probability distribution of the system and model output variables is of the form:

$$P\{Y < y; Y_M < y_M\} = P\{X < y; X < y_M\} =$$
$$= P\{X < \min(y, y_M)\} = \Phi(\min(y, y_M)).$$

Hence, the joint probability distribution density $p_{SM}(y, y_M)$ of the system and model output variables is not Gaussian.

As to those seldom partial cases, when the assumption that the joint probability distribution density is Gaussian is valid (if the property is implied by the system and model structure, probabilistic properties of input signals, etc.) reasonability of such an approach is quite questionable since, for this case, it is enough to apply ordinary least square criterion (for the joint Gaussian distribution the maximal correlation is well known to be linear and to coincide with the ordinary one).

Shannon mutual information, commonly referred also as Shannon relative entropy, is based on involvement of corresponding Shannon entropies. At the same time, along with Shannon entropy, a number of other ways of defining the entropy of a (multivariate) random value are known. So, for a v-dimensional random value Z with multivariate probability distribution density $g(z)$, Tsallis entropy of the order α is defined as [9]

$$T_\alpha(Z) = \frac{1}{\alpha-1}\left(1 - \int\limits_{R^v}(g(z))^\alpha dz\right),\ \alpha > 0, \alpha \neq 1. \quad (2)$$

Meanwhile, as α tends to 1, $T_\alpha(Z)$ tends to Shannon entropy, and, thus Shannon entropy may be considered as Tsallis entropy one of the "order 1".

Like Shannon entropy, Tsallis entropy $T_\alpha(Z)$ of a v-dimensional random value Z with multivariate probability distribution density $g(z)$ may be considered with regard to

this probability distribution density $g(z)$, and within the case it will be denoted as $T_\alpha(g)$. Thus, $T_\alpha(Z)$ and $T_\alpha(g)$ should be considered as mathematically equivalent designations, both defined by formula (2). The need of such a remark will be explained by the considerations below, concerned with a corresponding measure of divergence of probability distributions, which involves Tsallis entropy (2) as a basis.

Namely, using the convexity property of Tsallis entropy, in the literature, say [10], Jensen-Tsallis divergence $D_\alpha^{JT}\left(g_1 \| g_2\right)$ of the order α of two probability distribution densities $g_1(z)$ and $g_2(z)$ is defined as follows

$$D_\alpha^{JT}\left(g_1 \| g_2\right) = T_\alpha\left(\frac{g_1 + g_2}{2}\right) - \frac{T_\alpha(g_1) + T_\alpha(g_2)}{2}, \quad \alpha > 0, \alpha \neq 1. \tag{3}$$

Meanwhile, $D_\alpha^{JT}\left(g_1 \| g_2\right)$ is non-negative and vanishes if and only if $g_1(z) = g_2(z)$. Accordingly, as $\nu = 2$ and Z is composed of two random values X_1 and X_2, when one probability distribution density in $D_\alpha^{JT}\left(g_1 \| g_2\right)$ in (3), namely $g_1(z)$, is the joint probability distribution density of these random values $p_{12}(x_1, x_2)$, and the second one, $g_2(z)$, is the production of the marginal distribution densities of X_1: $p_1(x_1)$, and X_2: $p_2(x_2)$, corresponding Jensen-Tsallis divergence $D_\alpha^{JT}\left(p_{12} \| p_1 p_2\right)$ of the order α leads to consistent, in the sense of A.N. Kolmogorov terminology, information-theoretic measure of dependence $I_\alpha^{JT}\{X_1, X_2\}$ of two random values X_1 and X_2.

From computational point of view, especially under necessity of estimation by use of sample date, Tsallis entropy should be considered as more attractive than that of Shannon, since Shannon entropy involves "integral of logarithm", while Tsallis entropy does not involve logarithm at all, what is considerably computationally simpler.

Meanwhile, selecting a particular value of the order α is of importance since the larger the order is, the more complicated a computational procedure becomes. These considerations of the computational and analytical issues of the value of the order α of Tsallis entropy imply reasonability of achieving a "compromise", within which the parameter value $\alpha = 2$ looks most attractive. The value $\alpha = 2$ in formula (3) corresponds to the quadratic Jensen-Tsallis divergence and mutual information correspondingly. Thus, as $\alpha = 2$ from formulae (2) and (3) it directly follows

$$I_2^{JT}\{X_1, X_2\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(p_{12}(x_1, x_2) - p_1(x_1)p_2(x_2)\right)^2 dx_1 dx_2. \tag{4}$$

2. PROBLEM STATEMENT AND SOLUTION

As the main inference from the considerations of the above Section, a natural question arises, if there exist constructive ways of using the information-theoretic criterion, which would not be based on the restrictive assumptions of the kind considered. If so, obviously such an approach cannot be based on direct analytical involving the information-theoretic criterion since it is a functional of the unknown marginal $p_S(y)$, $p_M\left(y_M(\theta)\right)$, and joint $p_{SM}\left(y, y_M(\theta)\right)$ probability distribution densities of the system, $y(t)$, and model, $y_M(t;\theta)$, output processes (all notations regarding the probability distribution densities completely correspond to the ones in the preceding Section). Hence, a feature of the constructive method is to apply some appropriate sample data estimates of the information-theoretic criterion instead of the analytical one.

Let us consider a widely used in the theory and practice of identification class of non-linear discrete time system described by a linear-in-parameters mapping

$$y_M(t;\theta) = \theta^T \phi(t), \tag{5}$$

$\theta = (\theta_1, \ldots, \theta_n)^T$. Components of the column-vector $\phi(t) = (\phi_1(t), \ldots, \phi_n(t))^T$ are some known functions of preceding values of the input process of the system, as well as, generically, preceding values of the output system process. Thus, equation (5) is applicable to describe a broad class of (generically) non-linear dynamic systems.

Within the problem statement, the model parameters, that is the column-vector θ components are subject to identification in accordance to the information theoretic criterion

$$I_2^{JT}\{y(t), y_M(t;\theta)\} \xrightarrow[\theta]{} \sup, \tag{6}$$

with simultaneous substitution of the analytical expression for the quadratic Jensen-Tsallis mutual information with a suitable estimate

$$\hat{I}_2^{JT}\left(y^{(1)}, \ldots, y^{(N)}; \phi^{(1)}, \ldots, \phi^{(N)}\right) = f(\theta) \tag{7}$$

obtained basing on observations of sample data $\phi^{(1)}, \ldots, \phi^{(N)}$, $y^{(1)}, \ldots, y^{(N)}$ of the (generalized) input, $\phi(t)$, and output, $y(t)$, processes of the system. In (7), $\phi^{(i)} = \left(\phi_1^{(i)}, \ldots, \phi_n^{(i)}\right)^T$, $i = 1, \ldots, N$.

Then within the approach, the initial problem of identification of stochastic system model (4) with information-theoretic criterion (6), (7) under availability of sample values of the input and output processes gives rise to a problem of finite dimensional optimization

$$f(\theta) \xrightarrow[\theta]{} \sup,$$

to solve which deriving an explicit analytical expression for the function $f(\theta)$ in (7) is required. In turn, such deriving is based on applying a corresponding technique of estimating the quadratic Jensen-Tsallis mutual information.

The function $f(\theta)$ may be obtained in various ways concerned with estimating joint and marginal distribution densities of the input and output processes, based on sample data. For the cases, the Rosenblatt kernel type density estimates [11] are widely used.

Thus, in accordance with formulae (4) and (7) one may write

$$\hat{I}_2^{JT}\left(y^{(1)}, \ldots, y^{(N)}; \phi^{(1)}, \ldots, \phi^{(N)}\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\mathrm{A}_{SM}^{(N)}\left(y, y_M(\theta)\right)\right)^2 dy dy_M(\theta), \tag{8a}$$

where

$$\mathrm{A}_{SM}^{(N)}(y, y_M(\theta)) = \\ = p_{SM}^{(N)}(y, y_M(\theta)) - p_S^{(N)}(y) p_M^{(N)}(y_M(\theta)). \quad (8b)$$

In turn, in (8)

$$p_S^{(N)}(y) = \frac{1}{Nh_N} \sum_{i=1}^{N} K_1\left(\frac{y - y^{(i)}}{h_N} \right), \quad (9)$$

$$p_M^{(N)}(y_M(\theta)) = \frac{1}{Nh_N} \sum_{i=1}^{N} K_2\left(\frac{y_M(\theta) - \theta^T \phi^{(i)}}{h_N} \right), \quad (10)$$

$$p_{SM}^{(N)}(y, y_M(\theta)) = \\ = \frac{1}{Nh_N^2} \sum_{i=1}^{N} K_1\left(\frac{y - y^{(i)}}{h_N} \right) K_2\left(\frac{y_M(\theta) - \theta^T \phi^{(i)}}{h_N} \right). \quad (11)$$

In expressions (9) to (11), $\{h_N\}$ is a sequence of positive real numbers converging to zero; $K_j(\cdot)$, $j = 1,2$ are positive bounded kernels on R^1, meeting conventional conditions imposed on kernels under non-parametric density estimation.

Under assumption on the initial system subject to identification of the form that $(y(t), y_M(t;\theta))$ to be strongly mixing random processes [12], and suitable integrability conditions imposed on the kernels $K_j(\cdot)$, $j = 1,2$, and densities $p_S(y)$, $p_M(y_M(\theta))$, and $p_{SM}(y, y_M(\theta))$ (formulae (3) to (7) in [13]), estimate (8) has the following mean square risk

$$\mathbf{E}\left(\hat{I}_2^{JT\left(y^{(1)},\ldots,y^{(N)};\phi^{(1)},\ldots,\phi^{(N)} \right)} - I_2^{JT}\{y(t), y_M(t;\theta)\} \right)^2 = \\ = O\left(N^{-1} h_N^{-4} + h_N^2 \right).$$

3. DISCUSSION

The significance of application in system identification of information-theoretic criteria as a measure of dependence is motivated by the evidence that stochastic system identification is always based on utilizing measures of dependence of random values or processes, both within representation of a system under study either by use of an input/output mapping, or within state-space description framework. Among the measures of dependence, conventional correlation and covariance ones are the most widely known and used. Their application directly follows from the identification problem statement itself, when it is based on the conventional mean square criterion. The main advantage of these measures is convenience of their use involving both a possibility of deriving explicit analytical expressions to determine required system characteristics, and relative simplicity of constructing their estimates involving those of based on observation of dependent data. However, the main disadvantage of the measures of dependence based on the linear correlation is their ability to vanish even if there exists some deterministic dependence between random values. Even more so is with regard to a stochastic dependence. Just to overcome such a disadvantage, use of more complicated, nonlinear, as the information-theoretic ones, measures of dependence was involved in the system identification.

The key feature of the present paper approach was applying a consistent measure of dependence. In accordance with the A.N. Kolmogorov terminology, a measure of dependence between two random values is referred as consistent if it vanishes if and only if the random values are stochastically independent. With regard to the system identification with the information theoretic criterion presented, vanishing the Jensen-Tsallis mutual information would indicate the system under study to be not identifiable at all.

REFERENCES

[1] Stoorvogel, A.A. and J.H. van Schuppen. “System identification with information theoretic criteria”, In: *Identification, adaptation, learning / Ed. by S. Bittanti and G. Picci*. Springer, 1996, pp. 289-338.

[2] Durgaryan, I.S. and F.F. Pashchenko. “An Information Theory Approach to Identification”, *IFAC Proceedings Volumes*, 1982, vol. 15, no. 4, pp. 731-735.

[3] Durgaryan, I.S. and F.F. Pashchenko. “Identification of Objects by the Maximal Information Criterion”, *Automation and Remote Control*, 2001, vol. 62, no. 7, pp. 1104-1114.

[4] Pashchenko, F.F. “Determining and modeling regularities via experimental data”, In: Prangishvili, I.V., Pashchenko, F.F., and B.P. Busygin. *System Laws and Regularities in Electrodynamics, Nature, and Society*, Chapter 7, Nauka Publ., Moscow, 2001, pp. 411-521. (in Russian)

[5] Durgaryan, I.S., Pashchenko, F.F., Pikina, G.A., and A.F. Pashchenko. “Information method of consistent identification of objects”, *8th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2013, pp. 1325-1330. Digital Object Identifier: 10.1109/ICIEA.2013.6566572.

[6] Pashchenko, F.F. “The method of functional transformations and its application within problems of modeling and identification of systems”, *Doctoral Thesis*, V.A. Trapeznikov Institute of Control Sciences, Moscow, 2001, 114 p. (in Russian)

[7] Pashchenko, F.F. *Introduction to consistent methods of systems modeling. Identification of nonlinear systems*, Finansy i statistika Publ., Moscow, 2007, 288 p. ISBN 978-5-279-03042-2 (in Russian)

[8] Korolyuk, V.S., Portenko, N.I., Skorokhod, A.V., and A.F. Turbin. *Handbook on Probability Theory and Mathematical Statistics*, Nauka Publ., Moscow, 1985, 640 p. (in Russian)

[9] Tsallis, C. “Possible generalization of Boltzmann-Gibbs statistics”, *Journal of Statistical Physics*, 1988, vol. 52, no. 1, pp. 479–487.

[10] Anguloa, J.C., Antolín, J., López-Rosa, S., and R.O. Esquivel. “Jensen–Tsallis divergence and atomic dissimilarity for ionized systems in conjugated spaces”, *Physica A*, 2011, vol. 390, pp. 769-780.

[11] Rosenblatt, M. “Remarks on some nonparametric estimates of a density function”, *Ann. Math. Statist.*, 1956, vol. 27, pp. 832-835.

[12] Rosenblatt, M. “A central limit theorem and a strong mixing condition”, *Proc. Nat. Acad. Sci., U.S.A.*, 1956, vol. 42, pp. 43-47.

[13] Mokkadem, A. “Estimation of the entropy and information of absolutely continuous random variables”, *IEEE Transactions on Information Theory*, 1989, vol. IT-35, pp. 193-196.