Information Theory within System Identification: Revising Some Approaches

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

ABSTRACT
The paper presents methods to analyze approaches concerned with application of information theoretic techniques in such a branch of the control theory as system identification: application of the mutual (Shannon) information and attempts of generalization of the notion of entropy, as well as application of consistent measures of dependence based on the information-theoretic (Kulback-Leibler) divergence in system identification. Ways and methods, both analytical and simulation ones are presented.

Keywords: Entropy, Gaussian distributions, Information theory, Integrals, Joint probability, Nonlinear systems, Random variables, System Identification, Software tools

1. INTRODUCTION
Conventionally, solving an identification problem always implies using a measure of dependence of random values (processes) both within representation of the system under study by an input/output relationship and as a state-space description. Among the measures of dependence, conventional correlation and covariance once are the most widely used. Their application is directly implied from the problem statement itself, based on the mean squared criterion. A main advantage of the measures is convenience of their use involving both a possibility of deriving explicit analytical expressions to determine the required 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 linear correlation is the fact that these may vanish even provided that there exists a deterministic dependence between the pair of the investigated variables.
Just to overcome such a disadvantage, use of more complicated, nonlinear, measures of dependence has been involved into the system identification. A feature of the technique proposed in the paper is that it is based on application of a *consistent* measure of dependence. Following to Kolmogorov' terminology, a measure of stochastic dependence between two random variables is referred as consistent if it vanishes if and only if the random variables are stochastically independent. Among the measures, the maximal correlation coefficient, Shannon mutual information, contingency coefficient are commonly known. Under investigation of the random processes, the measures (coefficients) are substituted by the corresponding functions. Among the functions, being the consistent measures of dependence, the following ones are the most known: the maximal correlation and Shannon mutual information. However, calculating the maximal correlation function is known to be a significantly complicated iterative procedure. So, as suitable mathematical tools within the paper, the information/entropy based measures of dependence are used.

Application of consistent measures of dependence possesses some particularities and limitations. Within the scope, the Shannon mutual information looks more preferable than the maximal correlation whose calculation deals with necessity of using a complex iterative procedure of determining the first eigenvalue and the pair of the first eigenfunctions of the stochastic kernel

$$\frac{p_{21}(y,w)}{\sqrt{p_1(w)p_2(y)}}.$$

In the formula above, $p_1(w), p_2(y)$, $p_{21}(y,w)$ stand respectively for the marginal and joint distribution densities of the corresponding random values.

In turn, the information theoretic criterion gives rise to applying the mutual information. Recent examples of such an approach are presented at the ridge of the Millennia in [1-4], the present paper, involving results of [5-8], demonstrates ways and methods, both analytical and simulation ones, to be applied to analyze the information theoretic approaches applied within the system identification.

2. AN INFORMATION CRITERION WITHIN SYSTEM IDENTIFICATION
Problem 1. In [1-4], the mutual Shannon information $I\{Y,Y_M\}$ of model "output" Y_M and system "output" Y has been considered as an identification criterion to derive the required model. Such a criterion, which has been referred as the information one, is to be maximized, and the model "output" is just considered as the maximization argument:

$$I\{Y,Y_M\} = \\ = \mathbf{E}\left\{\log\left(p(y,y_M)\Big/p(y)p(y_M)\right)\right\} \to \max_{Y_M}. \qquad (1)$$

Here $p(y,y_M), p(y), p(y_M)$ stand for the joint and marginal distribution densities of the above Y and Y_M, correspondingly; and $\mathbf{E}$ stands for the mathematical expectation.
One may justify by reasoning that the approach of problem 1 is not constructive within system identification. Indeed, the approach initially is based either on a requirement that the joint distribution density $p(y, y_M)$ of the model "output" Y_M and system "output" Y is to be preliminary known (what is nonsense, in entity), or the above "outputs" are able to be observed. But this, the second way is not applicable because the problem is just to derive the model, and, hence, its "output" can naturally not be observed. As to the first way, it also cannot be considered

as acceptable, because it requires such an amount of a priori knowledge under which the identification problem already is to loose its sense: the joint distribution of model and system "outputs" is a final result of many factors (system and model structure, statistical properties of "inputs", etc.). ■

Problem 2. In criterion (1), postulating a concrete kind of the joint distribution density of the "outputs" of model and system has been used as a basis for analytical inferences, in [1-4] the joint distribution of the model and system "outputs" is assumed to be the Gaussian one, what directly gives rise of the initial identification problem to the problem of maximizing the correlation coefficient of the "outputs" of model and system.

$$p_{y,y^*}\left(y,y^*\right)=\int\limits_{(z_{n+1},\varphi(z_1,\ldots,z_n))\in C}\cdots\int p_{x_1,\ldots,x_n,y}\left(z_1,\ldots,z_n,z_{n+1}\right)\frac{dS_{n-1}}{\sqrt{\sum\limits_{i_1<i_2}\left[\frac{D\left(z_{n+1},\varphi\right)}{D\left(z_{i_1},z_{i_2}\right)}\right]^2}}.$$

The above formula is written for the system model represented as $Y_M=\varphi\left(X_1,\ldots,X_n\right)$ where $X_1,\ldots,X_n$ are the (generalized) system input variables, Y is the system output variable, $p_{X_1,\ldots,X_n,Y}\left(z_1,\ldots,z_n,z_{n+1}\right)$ is the joint 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

$$\begin{cases}\varphi\left(z_1,\ldots,z_n\right)=Y_M\\ z_{n+1}=Y\end{cases},\quad\text{and}\quad\frac{D\left(z_{n+1},\varphi\right)}{D\left(z_{i_1},z_{i_2}\right)}=\begin{vmatrix}\frac{\partial z_{n+1}}{\partial z_{i_1}} & \frac{\partial z_{n+1}}{\partial z_{i_2}}\\ \frac{\partial\varphi}{\partial z_{i_1}} & \frac{\partial\varphi}{\partial z_{i_2}}\end{vmatrix}$$

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

From the formulae above, in particular, a well known fact follows that the joint probability distribution density is Gaussian if the probability distribution density $p_{x_1,\ldots,x_n,y}\left(z_1,\ldots,z_n,z_{n+1}\right)$ is Gaussian and the function $\varphi\left(x_1,\ldots,x_n\right)$ describing the model is linear. So, in any of more general cases, for instance, considered within the de-scribed information-theoretic approach [1-4] the dispersional or dynamic models, there is no basis for the a priori assump-tion on the Gaussian nature of the joint probability distribu-tion of the model output and the output of "an arbitrary non-linear" plant. Such an assumption is just an evident simplifi-cation of the initial problem statement leading to degenerat-ing its entity.

One may also note that the assumption the joint distribution to be 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" $Y\equiv X$; the model "output" $Y_M\equiv X$; the joint distribution of the model and system "outputs" is of the form:

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

Hence, the joint distribution density $p\left(y,y_M\right)$ of the model and system "outputs" is not Gaussian.

One may justify by reasoning that the assumption described in problem 2 is not constructive within system identification. Indeed, from a substantial point of view, the assumption that the joint distribution of "outputs" of the model and system to be Gaussian is equivalent to that, for instance, if there would be proposed a new method of matrix inversion followed by an assumption that the matrix subject to inversion to be the diagonal one. In particular, one can write the following formal expression for the joint distribution density $p_{SM}\left(Y,Y_M\right)$ of the system's and model's output variables, which is implied by the relationship for the joint distribution density of a transformation of a random vector:

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

3. "GENERALIZATIONS" OF THE ENTROPY NOTION

Problem 3. In [1-3] one has introduced a number of definitions relating to the entropy notion (in the Shannon sense). These are:

- dynamic entropy

$$H^{\phi}\{Y\}=-\int_{-\infty}^{\infty}\left(p(B(y))\right)\log\left(l_Y\,p(B(y))\right)dy;\qquad(2)$$

- generalized dynamic entropy

$$H^{\phi 0}\{Y\}=$$
$$=-\int\limits_{\{B\}}\int_{-\infty}^{\infty}\left(p(B(y))\right)\log\left(l_Y\,p(B(y))\right)dy\,d\mu(B);\qquad(3)$$

- total entropy

$$H^{0}\{Y\}=H\{Y\}+H^{\phi 0}\{Y\};\qquad(4)$$

- maximal entropy

$$H^{\phi\max}\{Y\}=\max_{B}H^{\phi}\{Y\}.\qquad(5)$$

In formulae (2) to (5), l_Y "causes a reference mark on a scale of entropies" [1-3], and B is a nonlinear transformation. The elements By form the set of all states $\{BY\}$ which is the result of acting of arbitrary transformations B on the initial random value Y. Within such a framework, it is noted also that $H^{\phi\max}\{Y\}\le H^{\phi 0}\{Y\}$, and $H^{\phi\max}\{Y\}\le H^{0}\{Y\}$.

In turn, in [1-3] it is stated that the results of this textbook relating to the identification via information criterion (1), are valid both for the conventional entropy and the above considered generalized one. Meanwhile, in [1-3] no details are provided concerning such issues as existence of the values

$$H^{\phi}\{Y\},\ H^{\phi 0}\{Y\},\ H^{0}\{Y\},\ H^{\phi\max}\{Y\},$$

as well as a definition of the measure $\mu(B)$.

One may provide numerical examples demolishing the corresponding inferences of [1-3] with respect to the above presented generalizations of the entropy notion.

Indeed, the simplest way is just to demonstrate (2) to become infinite for a density $p(y)$ and for a transformation B of the random value Y.

In turn, the most obvious indicator of divergence of an improper integral is that the subintegral function does not tend to zero on the infinity.

As a tool to select the corresponding examples confirming the subintegral function in (2) (entering the sign "minus" under the sign of integral),

$$H^{\phi}\{Y\} = -\int_{-\infty}^{\infty} \left(p(B(y))\right)\log\left(l_Y p(B(y))\right)dy =$$

$$\int_{-\infty}^{\infty} \Phi(y)dy\,, \qquad (6)$$

meets the following conditions:

$$\Phi(y) \geq 0\ \forall y\,,\ \lim_{y\to\infty} \Phi(y) > 0\,, \qquad (7)$$

a corresponding computer package, such as MathCAD Professional, may be recommended to fit rapidly the transformation B for a preliminary selected density $p(y)$ by using the graphical representation.

Namely, let the random value Y have the standard Gaussian distribution density, and the nonlinear transformation B of this random value be chosen in the form

$$By \overset{def}{=} \begin{cases} ye^{-y^2}\sqrt{\ln\left(y^{-2}+1\right)}, \text{if } y \neq 0 \\ 0, \text{if } y = 0 \end{cases}.$$

For the case, the plot of the function in $\Phi(y)$ in (6) is of the kind presented in Figure 1 (in the end of the paper), and is practically a direct line that is parallel to the abscissa axis and situated at distance $\ln\sqrt{2\pi}/\sqrt{2\pi}$ from it. Obviously, for the case, $H^{\phi}\{Y\}$ is equal to infinity.

For the cases of

$$By \overset{def}{=} \begin{cases} y\sqrt{\ln\left(y^{-2}+1\right)}, \text{if } y \neq 0 \\ 0, \text{if } y = 0 \end{cases},\ By = \sin(y)\,,$$

$$By = \tan(y)\,,\ By = \arctan(y)$$

the corresponding functions $\Phi(y)$ in (6) meet conditions (7). Besides that, a visual analysis of the plots of these transformations gives a clear representation on the form of the transformations B for which (at the given distribution density) the dynamic entropy (2) becomes infinity (Figure 2). The same inference is valid for the function $\Phi(y)$ derived for the exponential distribution of the random value Y with the parameter $\lambda = 1$ and $By = \ln\left(y^{-2}+2\right)$. ■

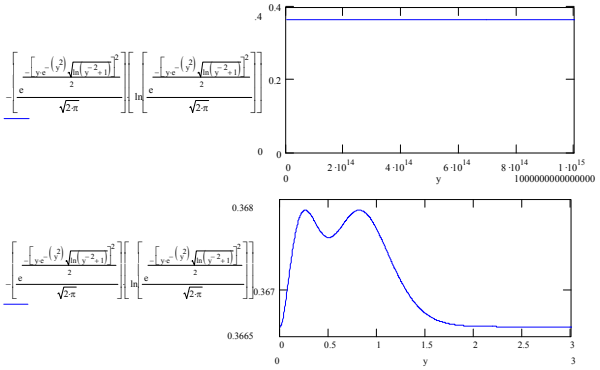

Figure 1: Towards infiniteness of $H^{\phi}\{Y\}$ under standard Gaussian distribution of Y and

$$By \overset{def}{=} \begin{cases} ye^{-y^2}\sqrt{\ln\left(y^{-2}+1\right)}, \text{if } y \neq 0 \\ 0, \text{if } y = 0 \end{cases}.$$

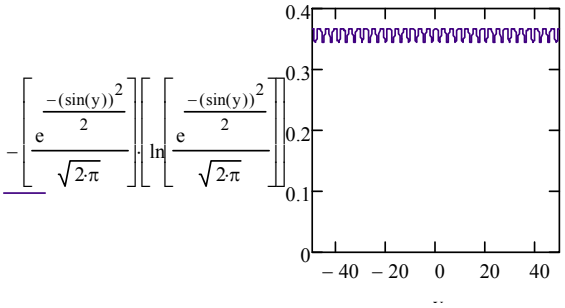

Figure 2a: Towards infiniteness of $H^{\phi}\{Y\}$ under standard Gaussian distribution of Y and $By = \sin(y)$.

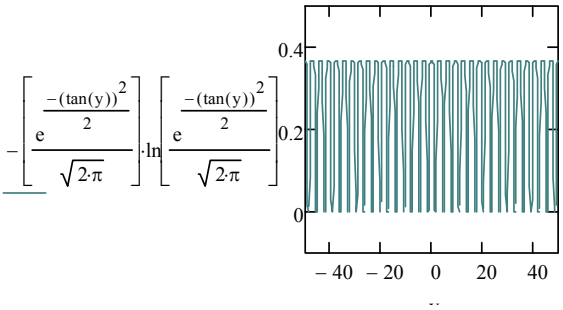

Figure 2b: Towards infiniteness of $H^{\phi}\{Y\}$ under standard Gaussian distribution of Y and $By = \tan(y)$.

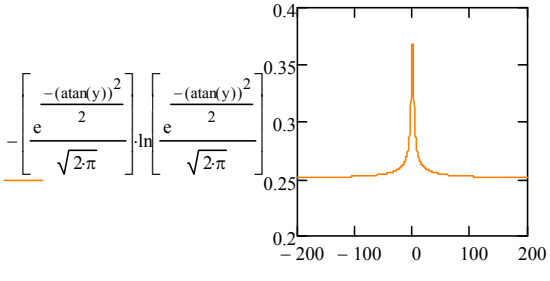

Figure 2c: Towards infiniteness of $H^{\phi}\{Y\}$ under standard Gaussian distribution of Y and $By = \arctan(y)$.

One may prove analytically that $H^{\phi\max}\{Y\}$ in (5) becomes infinity for any probability density $p(y)$.

Indeed, since in accordance to [1-3] B is an arbitrary transformation, restrict the domain of the search of the extremum in (5) to the transformations of the form $BY \cdot \alpha = Y, \quad \in \alpha R^1$. Then, based on formulae (2), (5),

$$H^{\phi\max}\{Y\} = \max_{B} H^{\phi}\{Y\} \geq$$

$$\geq \max_{\alpha \in R^1}\left\{-\int_{-\infty}^{\infty}\left(p(\alpha \cdot y)\right)\log\left(l_Y p(\alpha \cdot y)\right)dy\right\}.$$

Quod erat demonstrandum. ■

Numerical example. Let a Gaussian random value be transformed by multiplication by the scalar: $BY = \alpha \cdot Y, \quad \alpha \in [0;1]$. The plot of the subintegral expression, obtained under the transformation with simultaneous insertion of the sigh "minus" into the integral sign, is presented in Figure 3 for different magnitudes of $\alpha \in [0;1]$. One can easily be seen that the integral in (2) exists at any $\alpha \in ([0;1]]$, and the less α is, the larger the magnitude of $H^{\phi}\{Y\}$ is.

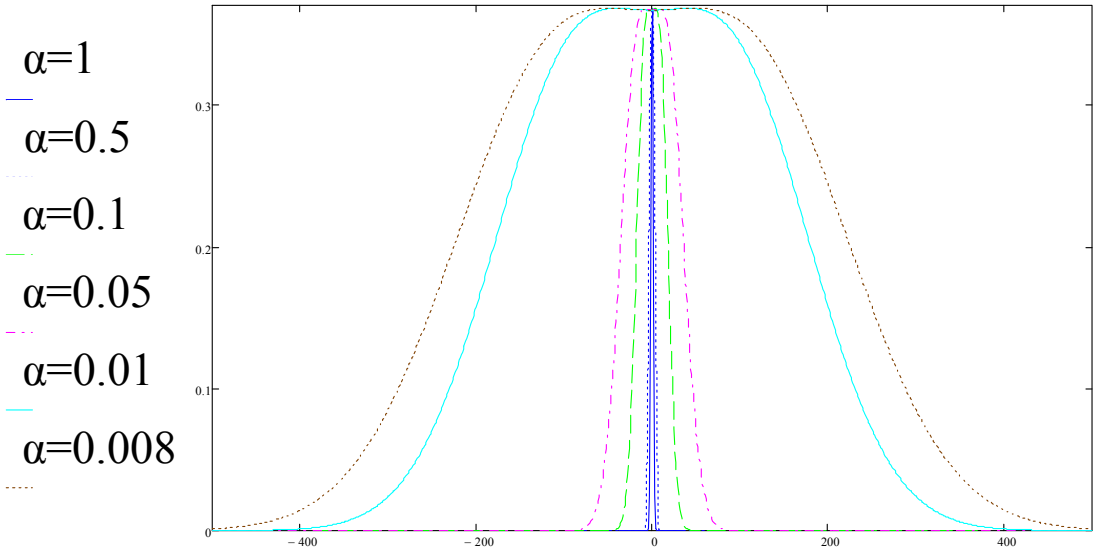

Figure 3: Plot of the subintegral expression in (2) with insertion of the sign "minus" into the integral sign for the standard Gaussian random value Y under its transformation of the form $BY=\alpha Y$ for different magnitudes of α from the interval [0, 1].

Under $\alpha = 0$ the integral in (2) diverges (the plot of the corresponding expression with introducing the sigh "minus" into the integral sign for $\alpha = 0$) is presented in Figure 4). Starting with which magnitude of α the notion of the "dynamic" entropy loses its sense?

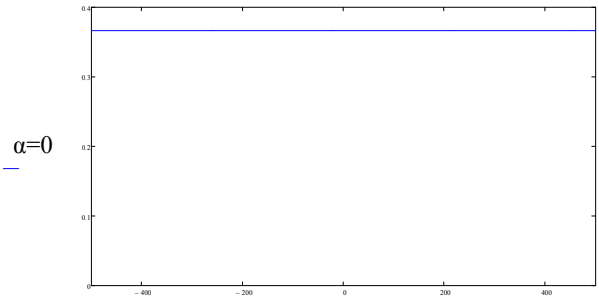

Figure 4: Plot of the subintegral expression in (2) with insertion of the sign "minus" into the integral sign for the standard Gaussian random value Y under its transformation of the form $BY=\alpha Y$ for the zero α.

REFERENCES

[1] Pashchenko, F.F. "Determining and modeling regularities via experimental data", In: *System Laws and Regularities in Electrodynamics, Nature, and Society*. Chapter 7, Nauka Publ., Moscow, 2001, 411-521, 2001. (in Russian)

[2] 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, 114 p., 2001. (in Russian)

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

[4] 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.

[5] Chernyshov, K.R. "An essay on some delusions in system identification", Proceedings of the II International conference "System Identification and Control Problems" SICPRO '03. Moscow, 29-31 January 2003. V.A. Trapeznikov Institute of Control Sciences, Moscow, 2003, pp. 2660-2698. (in Russian)

[6] Chernyshov, K.R. *Questions of identification: consistent measures of dependence*, Moscow, V.A. Trapeznikov Institute of Control Sciences, 60 p., 2003. (in Russian)

[7] Chernyshov, K.R. "Towards the support of the education process in systems modeling", *Quality. Innovations. Education*, no. 9, pp. 39-50, 2007. (in Russian)

[8] Chernyshov, K.R. "Stochastic systems and information-theoretic methods: an analysis of some approaches", In: *Proceedings of the 9th International Conference "System Identification and Control Problems" SICPRO '12.* Moscow, January 30 - February 2, 2012. V.A. Trapeznikov Institute of Control Sciences, Moscow, 2012, pp. 1140-1164. (in Russian)