INFORMATION THEORETIC IDENTIFICATION CRITERIA: APPROACHES AND ALTERNATIVES

Kirill Chernyshov

Institute of Control Sciences
65 Profsoyuznaya, Moscow 117997, Russia
E-mail: myau@ipu.rssi.ru

Abstract. The aim of the paper is to present a conceptual approach to identification of nonlinear stochastic systems based on information measures of dependence. In the paper, an identification problem statement using the information criterion under rather general conditions is proposed. It is based on a parameterized description of the system model under study combined with a corresponding method of estimation of the mutual information of the system's and model's output variables. 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 unreal a priori assumptions degenerating the entity of the initial identification problem statement like those ones presented in some referenced literature sources and revised in the present paper.

Keywords: Identification algorithms, Information theory, Nonlinear systems, Entropy, Global optimization, Parameter estimation, Probability density function, Gaussian distributions, Genetic algorithms, Criterion functions.

1. INTRODUCTION

Naturally, solving an identification problem always imply 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 measure of dependence, conventional correlation and covariance once are the most widely used. Theirs application is directly implied from the problem statement itself, based on the mean squared criterion. A main advantage of the measures is convenience of theirs use involving both a possibility of deriving explicit analytical expressions to determine the required characteristics and relative simplicity of constructing theirs 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 exist 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 (Sarmanov and Zakharov, 1960). Among the measures, the maximal correlation coefficient, Shannon mutual information, contingency coefficient (Rényi, 1959, Sarmanov and Zakharov, 1960) 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 function

$$S_{yw}(v) = \sup_{\{B\},\{C\}} \frac{\mathbf{cov}(B(y(t)), C(w(s)))}{\sqrt{\mathbf{D}(B(y(t)))\mathbf{D}(C(w(s)))}}$$

where the supremum is being taken over all the Borel measurable functions *B*, *C*.

- Shannon mutual information (function)

$$I_{yw}(\tau) = \mathbf{M}\left\{ \ln\left(\frac{p_{21}(y,w,\tau)}{p_1(w)p_2(y)} \right) \right\};$$

- the contingency function

$$\Delta^2_{yw}(\tau) = \mathbf{M}\left\{ \frac{(p_{21}(y,w,\tau) - p_1(w)p_2(y))^2}{p_{21}(y,w,\tau)p_1(w)p_2(y)} \right\}.$$

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.

Throughout the paper, the symbols $\mathbf{M}(\bullet)$, $\mathbf{D}(\bullet)$, and $\mathbf{cov}(\bullet,\bullet)$ respectively stand for the mathematical expectation, variance, conditional expectation, and covariance. Also, $p_1(w), p_2(y)$, $p_{21}(y,w,\tau)$ stand respectively for the marginal and joint distribution densities of the stationary and joint stationary random processes $y(t), w(s), \tau = t - s$.

Based on the mutual information $I_{yw}(\tau)$ which takes its values in $[0,\infty)$, the normalized quantity $\iota_{yw}(\tau)$ proposed originally by {Linfoot, 1957) and defined as

$$\iota_{yw}(\tau) = \sqrt{1 - e^{-2I_{yw}(\tau)}}$$

takes its values in $[0,1]$, and also is a consistent measure of dependence. It has been shown by Chernyshov (2002b) that a problem of statistical input/output linearization of a nonlinear system driven by the Gaussian white noise process in accordance to the criterion of coincidence of the mutual information of the system output and input processes and model output and input processes just leads to using the quantity $\iota_{yw}(\tau)$ to derive the weight coefficients of the weight function of the linearized model.

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

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

In turn, the information theoretic criterion gives rise to applying the mutual information. Recent examples of such an approach are presented in (Durgaryan and Pashchenko, 2001, Pashchenko 2001, Stoorvogel and van Schuppen, 1996). Identification relevant problems solved by use of the information criteria are presented by Basak (2002), Principe et al. (2000), Uchida and Yoshida (2001).

2. REVISING EXISTING APPROACHES

In the paper of Stoorvogel and van Schuppen (1996), 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

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

In the formula the following notations are introduced: $\mathbf{Z}$ 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 the paper Stoorvogel and van Schuppen (1996) is to demonstrate an equivalence of a number of criteria of identification and control for the linear Gaussian systems.

Papers of Durgaryan and Pashchenko (2001), Pashchenko (2001) consider the mutual Shannon information $I\{y(t), y_M(t)\}$ of model output variable Y_M and system output variable Y as an identification criterion to derive the required model. Such a criterion is to be maximized, and the model's output variable is just considered as the maximization argument:

$$I\{y(t), y_M(t)\} =$$
$$= \mathbf{M}\left\{ \log\left(\frac{p_{SM}(y, y_M)}{p_S(y)p_M(y_M)} \right) \right\} \to \max_{y_M}.$$

The approach proposed by Durgaryan and Pashchenko (2001), Pashchenko (2001) thus cannot be considered as a constructive one, because it initially is based either on a requirement that the joint distribution density $p_{SM}(y, y_M)$ of the model output process $y_M(t)$ and system output process $y(t)$ are to be preliminary known or the above model and system output processes are able to be observed (Chernyshov, 2002a, 2002b, 2003a, 2003b). But this can not be the case. In fact, one may advert to a widely used representation of the stochastic system identification criterion in the form

$$\mathbf{M}\left\{ \rho\left[y(t), y_M(t) \right] \right\} \to \underset{y_M(t)}{ext},$$

where as above **M** stands for the mathematical expectation, and ρ is a priori given loss function. Then, in the papers of Durgaryan and Pashchenko (2001), Pashchenko (2001) 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 especial importance) *joint* $p_{SM}(y,y_M)$ distribution densities of the system and model output processes respectively.

At the same time, the fact, that this joint distribution density is initially known within the problem statement, assumes such an amount of a priori knowledge under which the identification problem is already to loose its sense: the joint distribution of the model and system output processes is a final result of many factors (system and model structure, statistical properties of the inputs processes, etc.) (Chernyshov, 2002a, 2002b, 2003a, 2003b). In particular, one can write the following formal expression for the joint distribution density $p_{SM}(Y,Y_M)$ 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 (Korolyuk et al., 1985):

$$p_{Y,Y_M}(Y,Y_M)=\int\overset{n-1}{\cdots}\int_{(z_{n+1},\varphi(X_1,\dots,X_n))\in C} p_{X_1,\dots,X_n,Y}(z_1,\dots,z_n,z_{n+1})\frac{dS_{n-1}}{\sqrt{\sum\limits_{i_1<i_2}\left[\dfrac{D(z_{n+1},\varphi)}{D(z_{i_1},z_{i_2})}\right]^2}}\,.$$

The above formula is written for the system model represented as

$$Y_M=\varphi(X_1,\dots,X_n)$$

where $X_1,\dots,X_n$ are the (generalized) system input variables, Y is the system output variable, $p_{X_1,\dots,X_n,Y}(z_1,\dots,z_n,z_{n+1})$ 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

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

and

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

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

However, just postulating a concrete kind of the joint distribution density of the output variables of model and system has been used as a basis for analytical inferences of Durgaryan and Pashchenko (2001), Pashchenko (2001). Specifically, Durgaryan and Pashchenko (2001), Pashchenko (2001) assume the joint distribution of the model and system output processes to be the Gaussian one, what directly gives rice the initial identification problem to the problem of maximizing the correlation coefficient of the output processes of the model and system. From a substantial point of view, the assumption that the joint distribution of output variables 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 (Chernyshov, 2002a, 2002b, 2003a, 2003b).

In particular, from the two above formulae the well-known fact follows that the joint distribution of the system's and model's output variables is the Gaussian one if the distribution of $p_{X_1,\dots,X_n,Y}(z_1,\dots,z_n,z_{n+1})$ is the Gaussian one and the function $\varphi(X_1,\dots,X_n)$ describing the system model is linear. So, at any more general case there is no basement for a priori assumption the joint distribution of the system input and output processes to be the Gaussian one. Such an assumption would be just an artificial simplification of the initial identification problem statement, leading to emasculation of its entity. One also should be noted 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's output variable $Y\equiv X$; the model's output variable $Y_M\equiv X$; the joint distribution of the system's and model's output variables is of the form:

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

Hence, the joint distribution density $p_{SM}(y,y_M)$ of the system's and model's output variables is not the Gaussian one (Chernyshov, 2003a, 2003b).

As to those seldom partial 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) (Chernyshov, 2002a, 2002b, 2003a, 2003b).

3. PROBLEM STATEMENT

Thus, a natural question arises, if there exist a constructive way of using the information criterion, which wouldn't be based on the restrictive assumptions of the kind considered. If so, obviously such an approach can not be based on direct analytical involving the information criterion since it is a functional of the unknown marginal $p_S(y)$, $p_M(y_M(\theta))$ and joint $p_{SM}(y, y_M(\theta))$ distribution densities of the system's $y(t)$ and model's $y_M(t;\theta)$ output processes. Hence, a feature of the constructive method is to apply some appropriate sample data estimates of the information criterion instead of the analytical one.

Consider for sake of simplicity commonly used in system identification class of nonlinear discrete-time input/output systems which are linear in theirs parameters $\theta = (\theta_1, \dots, \theta_n)^T$

$$y_M(t;\theta) = \theta^T \phi(t).$$

Components of the column-vector $\phi(t) = (\phi_1(t), \dots, \phi_n(t))^T$ are some known functions of the preceding values of the system's input process as well as (generically) those of the system's output process. Within the problem statement, the system's parameters, i.e. the components of the column-vector θ, are subject to identification in accordance to the above information criterion

$$I\{y(t), y_M(t;\theta)\} \xrightarrow[\theta]{} \sup$$

with simultaneous substitution of the analytical expression of the mutual information

$$I\{y(t), y_M(t;\theta)\} = \int_{-\infty}^{\infty}\int_{\infty}^{\infty}\left\{\ln \frac{p_{SM}(y, y_M(\theta))}{p_S(y)p_M(y_M(\theta))}\right\} p_{SM}(y, y_M(\theta))dy dy_M(\theta)$$

by an appropriate estimate

$$\hat{I}_{y^{(1)},\dots,y^{(N)};\phi^{(1)},\dots,\phi^{(N)}}\{\theta\} = f(\theta)$$

obtained by observation of the sample values $\phi^{(1)},\dots,\phi^{(N)}$, $y^{(1)},\dots,y^{(N)}$ of the system's (generalized) input $\phi(t)$ and output $y(t)$ processes. Here $\phi^{(i)} = \left(\phi_1^{(i)},\dots,\phi_n^{(i)}\right)^T$, $i = 1,\dots,N$. Then, under such an approach, the initial stochastic system identification problem with the information criterion, given a sample of the input and output data, leads to a final dimensional deterministic optimization problem

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

to solve which an explicit analytical representation of the function $f(\theta)$ is to be derived first of all. The function derivation is based on applying on a suitable technique of the mutual entropy estimation.

4. REDUCING TO FINITE-DIMENSIONAL OPTIMIZATION

Obtaining the function $f(\theta)$ may be implemented in various manners relating to estimation of the joint and marginal distribution densities of the output and input processes of system by sampled data; and Rosenblatt (1956b) kernel-type density estimates are commonly used within the problem. Generically, estimating mutual information is implemented via estimation of the corresponding mutual entropy and the marginal ones. As well known, the mutual information of any two random processes, under the present consideration, the system's and model's output processes $y(t)$ and $y_M(t;\theta)$ respectively, is expressed via their marginal and mutual entropies in the following manner:

$$I\{y(t), y_M(t;\theta)\} = H_S + H_M(\theta) - H_{SM}(\theta),$$

where

$$H_S = -\int_{-\infty}^{\infty} (\ln p_S(y)) p_S(y) dy,$$

$$H_M(\theta) = -\int_{-\infty}^{\infty} (\ln p_M(y(\theta))) p_M(y(\theta)) dy(\theta),$$

$$H_{SM}(\theta) =$$

$$-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\ln p_{SM}(y, y(\theta))) p_{SM}(y, y(\theta)) dy dy(\theta)$$

are the marginal and mutual entropies of $y(t)$ and $y_M(t;\theta)$. Then, following to the approach of the paper of Mokkadem (1989) to obtain an estimate $\hat{I}_{y^{(1)},\dots,y^{(N)};\phi^{(1)},\dots,\phi^{(N)}}\{\theta\} = f(\theta)$ of the mutual information $I\{y(t), y_M(t;\theta)\}$ using N pairs of sampled observations of the random processes $y(t)$

and $\phi(t)$ kernel-type density estimates which are commonly used within such a kind of problems) the following relationships (in general) are natural to be applied (hereafter the super script (*N*) will stand for the corresponding estimate of a function over a sample of the length *N*):

$$\hat{I}_{y^{(1)},\ldots,y^{(N)};\phi^{(1)},\ldots,\phi^{(N)}}\{\theta\}=$$
$$=H_S^{(N)}+H_M^{(N)}(\theta)-H_{SM}^{(N)}(\theta)\,; \qquad (1)$$

$$H_S^{(N)}=-\frac{1}{h_N}\ln\int_{-\infty}^{\infty}\left(p_S^{(N)}(y)\right)^{h_N+1}dy\,, \qquad (2)$$

$$H_M^{(N)}(\theta)=$$
$$=-\frac{1}{h_N}\ln\int_{-\infty}^{\infty}\left(p_M^{(N)}(y_M(\theta))\right)^{h_N+1}dy_M(\theta)\,, \qquad (3)$$

$$H_{SM}^{(N)}(\theta)=-\frac{1}{h_N}\ln\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(p_{SM}^{(N)}(y,y_M(\theta))\right)^{h_N+1}dy\,dy_M(\theta)\,; \qquad (4)$$

$$p_S^{(N)}(y)=\frac{1}{Nh_N}\sum_{i=1}^{N}K_1\left(\frac{y-y^{(i)}}{h_N}\right), \qquad (5)$$

$$p_M^{(N)}(y_M(\theta))=$$
$$=\frac{1}{Nh_N}\sum_{i=1}^{N}K_2\left(\frac{y_M(\theta)-\theta^T\phi^{(i)}}{h_N}\right), \qquad (6)$$

$$p_{SM}^{(N)}(y,y_M(\theta))=$$
$$=\frac{1}{Nh_N^2}\sum_{i=1}^{N}K_1\left(\frac{y-y^{(i)}}{h_N}\right)K_2\left(\frac{y_M(\theta)-\theta^T\phi^{(i)}}{h_N}\right). \qquad (7)$$

In Equations (2) to (7), $\{h_N\}$ is a sequence of positive real numbers converging to zero; in Equations (5) to (7), $K_j(\cdot)$, $j=1,2$ are positive bounded kernels on $\mathbf{R}^1$.

In turn, within the identification problem statement, Equations (2) and (5) are not required since the marginal entropy of the system's output process $y(t)$ does not involve the unknown column-vector of the model parameters θ, what directly implies the following relationship

$$\arg\inf_{\theta}\hat{I}_{y^{(1)},\ldots,y^{(N)};\phi^{(1)},\ldots,\phi^{(N)}}\{\theta\}=\arg\inf_{\theta}\tilde{f}(\theta)\,,$$

where

$$\tilde{f}(\theta)=H_M^{(N)}(\theta)-H_{SM}^{(N)}(\theta)\,.$$

Thus, the required information type performance index subject to maximization is of the form

$$\tilde{f}(\theta)=\frac{1}{h_N}\ln\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{1}{Nh_N^2}\sum_{i=1}^{N}K_1\left(\frac{y-y^{(i)}}{h_N}\right)K_2\left(\frac{y_M(\theta)-\theta^T\phi^{(i)}}{h_N}\right)\right)^{h_N+1}dy\,dy_M(\theta)-$$
$$-\frac{1}{h_N}\ln\int_{-\infty}^{\infty}\left(\frac{1}{Nh_N}\sum_{i=1}^{N}K_2\left(\frac{y_M(\theta)-\theta^T\phi^{(i)}}{h_N}\right)\right)^{h_N+1}dy_M(\theta) \qquad (8)$$

to be supplemented by a scheme of numerical integration.

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 (Rosenblatt, 1956a), and suitable integrability conditions imposed on the kernels $K_j(\cdot)$, $j=1,2$, and densities $p_S(y)$, $p_M(y_M(\theta))$ $p_{SM}(y,y_M(\theta))$ (formulae (3) to (7) in Mokkadem (1989)), estimate (1) has the following mean squared risk

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

Of course, from an analytical point of view, expression (8) for the function $\tilde{f}(\theta)$ looks rather complex and the function may have several local maximums; and the natural way to solve such an optimization problem is applying the genetic algorithms (e.g. Baeck et al. (1997)) being an efficient tool for numerical function optimization.

5. CONCLUSIONS

A conceptual approach to input/output identification of nonlinear stochastic systems based on information measures of dependence has been presented. Within such an approach, an identification problem statement using a criterion consisting in maximization of mutual information of the system's and model's output variables under rather general conditions is proposed. In contrast to the approach derived, an alternative approach to application of the information methods of identification recently presented by Durgaryan and Pashchenko (2001), Pashchenko (2001) has been analyzed. The approach of these authors has been shown to be both nonconstructive and unrealistic due to imposing degenerating assumptions on the joint distribution density of the system's and model's output variables.

The present paper problem statement has been based on a parameterized description of the system model under study combined with a corresponding method of estimating the mutual information of the system and model output variables. 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 unreal a priori assumptions degenerating the entity of the initial identification problem statement like those ones presented in some referenced literature sources revised in the present paper.

REFERENCES

Baeck, T., D.B. Fogel and Z. Michalewicz (Eds.) (1997). “Handbook of Evolutionary Computation”, Bristol: Institute of Physics Publishing, 988 p.

Basak, I. (2002). “On the Use of Information Criteria in Analytic Hierarchy Process”, *European Journal of Operational Research*, **141**, no. 1, pp. 200-216.

Chernyshev, K.R. (2002a). “A Paper on the Identification of Stochastic Plants”, *Automation and Remote Control*, April 2002, **63**, issue 4, p. 687.

Chernyshev, K.R. (2002b). “Using Informational Measures of Dependence in Statistical Linearization”, *Automation and Remote Control*, September 2002, **63**, issue 9, pp. 1439-1447.

Chernyshov, K.R. (2003a). “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*. Moscow: Institute of Control Sciences, pp. 2660-2698. (in Russian)

Chernyshov, K.R. (2003b). “Questions of identification: consistent measures of dependence”, Moscow: Institute of Control Sciences, 60 p. (in Russian)

Durgaryan, I.S. and F.F. Pashchenko (2001). “Identification of Objects by the Maximal Information Criterion”, *Automation and Remote Control*, July 2001, **62**, issue 7, pp. 1104-1114.

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

Linfoot, E.H. (1957). “An information measure of correlation”, *Information and control*, **1**, pp. 85-89.

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

Pashchenko, F.F. (2001). “Determining and modeling regularities via experimental data”, In: *System Laws and Regularities in Electrodynamics, Nature, and Society* / Pranguishvili, I.V., F.F. Pashchenko and B.P. Boussyghine Moscow: Nauka Publ. Chapter 7, pp. 411-521. ISBN 5-02-013088-5 (in Russian)

Principe, J.C., Dongxin Xu, Qun Zhao and J.W. Fisher III (2000). “Learning From Examples with Information Theoretic Criteria”, *The Journal of VLSI Signal Processing*, **26**, no. 1/2, pp. 61-77.

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

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

Rosenblatt, M. (1956b). “Remark on some nonparametric estimates of a density function”, *Ann. Math. Statist.*, **27**, pp. 832-835.

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

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

Uchida, M. and N. Yoshida (2001). “Information Criteria in Model Selection for Mixing Processes”, *Statistical Inference for Stochastic Processes*, **4**, no. 1, pp. 73-98.