CONSISTENT APPROACHES FOR STRUCTURAL IDENTIFICATION OF NONLINEAR SYSTEMS AND THEIR IMPLEMENTATION

A.F Pashchenko[1]
Institute of Control Sciences RAS Moscow, Russia

ABSTRACT
This paper is devoted to present the approach of the selection of input variables in the model, based on a maximum and generalized correlation moments, software development and simulation modeling.

KEY WORDS: maximum coefficient of correlation, identification, simulation

MSC: 62J02

RESUMEN
Este trabajo es dedicado a presentar el enfoque basado en momentos maximales generalizados para la selección de las variables de entrada en el modelo, software y modelación de la simulación

1. INTRODUCTION

The choice of model structure is one of the basic moments in the formulation of the problem of identification. It creates a sufficient effect on the accuracy of the identification problem, methods and computing procedures. Universal approaches to the choice of the model structure obtained are very little.

In real systems, when there is no exact description of the facility and the processes occurring in it, the input and output variables of the model are treated as random variables or random functions. When studying the dependence between random variables its necessary to determine not only the equation of relation between them, but the degree, closeness of that relation. This is a fundamental difference from the deterministic approach, where there is a functional single-valued dependence between the variables.

When analyzing and modeling of nonlinear systems and statistical dependences mathematical tools of the classical correlation functions often gives underestimated evaluations of the statistical relationships between random variables and functions, and sometimes leads to simply incorrect results.

One of the most frequently used numerical characteristics of a linear relationship between output Y and input X random variables is the correlation coefficient

$$\rho_{yx}^2 = M[(Y - m_y)(X - m_x)]/\sigma_y\sigma_x \tag{1}$$

where σ_y σ_x - the average square deviation of the random variables Y and X respectively; $m_y = MY$ - the mathematical expectation of Y; $m_x = MX$ - the mathematical expectation of X .

But ρ_{yx} is not an exhaustive measure of dependence. It can turn into zero even for dependent Y and X and not be equal to 1 for the values having the functional nonlinear dependence. In the case of non-linear relationships between Y and X, ρ_{yx} gives underestimated value of the degree of relationship, and sometimes does not show this relationship at all. It is easy to show that even in the linear case correlation coefficient often does not reflect the real degree of the relationship between input and output variables. This is explained by the fact that the classical correlation functions and correlation moments (in the terminology of Kolmogorov) are not consistent characteristics of the dependence between the random elements.

In this paper we develop the ideas of G. Gebeleyn, O.V. Sarmanov, A. Renyi [1-4] and use a class of generalized (functional) correlation functions and statistical moments, which limiting cases on the one hand are the classical correlation functions and moments, and on the other hand - the maximum correlation function and the correlation ratio [2,3]

[1]

2. MODIFIED ALGORITHMS FOR CHOOSING THE INPUT VARIABLES OF THE MODEL

As already noted, the choice of model structure is one of the key moments in the construction of mathematical relationships and patterns. The choice of the structure sufficiently determines the methods, algorithms and computational procedures used in identification.

The target of modeling is usually to build a model of a system that satisfies some given accuracy or measure of identity:

$$Q^{\phi}_{y\bar{x}_q} \geq d, \tag{2}$$

where *d* - a given number.

Mathematical model of a system can be constructed by the maximum likelihood of the condition (2):

$$A^* = \arg\max_{\{A\}} P\{Q^{\phi}_{y\bar{x}_q} \geq d\},$$

or by the maximum of measure of identity

$$(A, B, C_i, i = 1,..., q) = \arg\max_{\{A,B,C\}} \; Q^{\phi}_{y\bar{x}_q}(t, s), \tag{3}$$

$$(A, B, C_i, G) = \arg\max_{\{A,B,C,G\}} \; Q^{\phi}_{y\bar{x}_q}(t, s), \tag{4}$$

where *G* - the index set of input variables included in the model $i \in G$. The *s* parameter is in general a vector quantity.

Usually the researchers apply two different approaches when solving one of the major problems of structure selection – the choice of input variables:

1. To construct a more accurate predictive or adequate to the object under study model tend to include in the model as much as possible amount of input variables. This allows to increase an accuracy of prediction of the output variable or output parameters and precision of control. It is assumed that this reduces the uncertainty that arises due to the neglect of factors affecting the output variable.
2. Since the inclusion of a large number of input variables require significant additional costs associated with obtaining information on these variables (for example, extra dimensions) and its subsequent validation, in practice researchers seek to ensure that the equation of the model included as little variables as possible. In favor of the second requirement is also indicated by the fact that the inclusion of a large number of input variables, even weakly correlated with each other, often leads to badly-conditioning of the correlation matrix and, consequently, by the incorrectness of inverse problems - inappropriate solutions to the problem of identification.

Let's consider the problem of selecting of informative variables for the generalized regression equation of the form

$$B(Y(t)) = \sum_{i=1}^{n} a_i C_i(X_i(s)) \tag{5}$$

where *Y (t)* - output of the object at time *t,* $Y \in R^1$, $t \in [0, T]$; B in general a non-linear operator (for example $B \in L^2$ - the Hilbert space of square integrable functions). $X_i(s)$ - *i*-th input signal of the object at time *s,* $X_i \in R^1$, $s \in [0, T]$; C_i non-linear operators, *i = 1, ..., n;* a_i - coefficients of the linear part of the object. As the input signals can be taken the meanings of the output signals at the moments previous to the time *t.* The introduction of time points *t* and *s* allows us to consider both static and dynamic cases. Where it does not matter, the time moments will be omitted. For simplicity we shall consider the static case.

Functions *(operators), B* and *C* represent the eigenfunctions of the stochastic kernel $p(y, \bar{x})[p(\bar{x})p(y)]^{-1/2}$ where $\bar{x} = (x_1, x_2, ...x_n)^T$ - vector of dimension *n.* In the case when *B* and *C* correspond to the maximum eigenvalue of the stochastic kernel, i.e. maximum correlation coefficient (the maximum of the correlation function in the dynamic case), we have the maximum arithmetization of space of input and output random

signals [1,2] In particular, when simulating the linear systems, functions *(operators) B* and *C* are identical transformations ($By = y$; $C\bar{x} = \bar{x}$ or, in general case, linear transformations.

Thus, the problem of selecting of informative variables consists in reducing the dimension of the input variables of the mathematical model of the object with the performance of requirements to the adequacy of a model or accuracy of the prediction.

Currently, there are many methods for selecting informative variables, such as group arguments method, method of all possible regressions, backward elimination method, forward selection method, stagewise regression analysis, factor analysis, various modifications and combinations of these methods and many others. These methods often yield the same results, but in general case, even when solving the same problems, their solutions are different.

We shall point out also that these solutions are not consistent, and in some cases lead to errors because correlation functions and correlation coefficients used in these methods are not consistent measures of dependence.

Let's consider a modification of the classical method of forward selection (inclusion), based on generalized and maximum correlation coefficients.

The classical method of variable inclusion is a method for selecting informative variables, i.e input variables included in the regression model one by one as long as the regression equation is satisfactory in terms of selected criteria. In contrast to the classical method of inclusion, which aims to build a linear model, we consider the modification of it, based on generalized and maximum correlation coefficients. The algorithm of the method is as follows.

Step 1. We calculate the maximum correlation coefficients $R^{\max}$ between each input variable and output variable of the object.
Step 2. We selects the input variable X_i, which has the largest absolute value of the maximum correlation coefficient with the output variable *Y*. Let's assume that it is X_1. If it is some other variable, we can enumerate them again.
Step 3. We find the equation of a generalized regression *Y* on X_1

$$B_1(Y) = A(C_1(X_1)) \text{ or } \hat{Y} = B_1^{-1}A(C_1(X_1)) \ .$$

In future we assume that there exists an inverse operator B^{-1}

Step 4. We compute the generalized partial correlation coefficients between all the remaining variables $X_2, X_3, \ldots, X_n$ and *Y*. From the mathematical point of view, this is equivalent to finding the correlation between the remainders of the regression $\hat{Y} = B^{-1}A(C_1(X_1))$ and remainders from the other regression $\hat{X}_j = F_j(X_1), (\hat{X}_j = B_j^{-1}AC_j(X_1))$
Step 5. We select the input value X_j, which has the highest partial correlation coefficient with the value of *Y*. Let's assume that this is X_2.
Step 6. We find the second generalized regression equation $\hat{Y} = B_2^{-1}A(C_1X_1, C_2X_2)$
Step 7. If the generalized regression equation received in step 6 satisfies a given accuracy $D(\hat{Y}) < d_{зад}$ the process of selecting of informative variables ends.

If the required accuracy of the model is not reached, the process of selecting of informative variables continues.

After selecting the *m* input variables $X_1, \ldots, X$, generalized partial correlation coefficients reflect the correlation between the residuals of the regression $\hat{Y} = F(C_1X_1, C_2X_2, \ldots, C_mX_m)$ and the residuals of the regressions $\hat{X}_j = F_j(C_1X_1, \ldots, C_iX_i, \ldots, C_mX_m), \quad i \neq j$

Note 1. Selection algorithm terminates either when the required accuracy has been achieved, or after the choice of a given (limited with *m)* number of variables.

Note 2. After the introduction of each new variable in the regression to test the statistical significance can be examined also the following statistical quantities:

1. The square of a generalized multiple correlation coefficient.
2. Private *F-test* for the included variable. The purpose of this - to find out whether the introduced variable makes a significant contribution to the total variation. When the value of the private *F-criterion* relating to the entered variable becomes insignificant, the process ends.

In the same way has been designed a modified stepwise regression method for selecting variables in the model. Like in the classical linear case, this method is an improved version of the modified method of forward selection of variables discussed above. The modification is not only including the use of generalized and maximum correlation coefficients, but additional investigation at every step of the variables included in the model in previous steps. It is known that the variable that was the best for the introduction in the model at an early stage in the following steps may become unnecessary because of its dependence with other variables included in the model. To test the significance of an included (or excluded) variable we calculate partial F-criteria for each variable from the regression equation and compare it with the value of F-distribution corresponding to the selected percentage point. This makes it possible to estimate the contribution into the model of each variable. It is assumed that this variable is introduced into the model last, not considering that in fact it could be introduced at earlier stages. Any variables that make a non-valuable contribution are excluded from the model. The process of selection of the variables can be continued as long as no variables are not being excluded from the equation and not being added to it.All this process can be represented as the stepped algorithm given below.

Step 1. Calculating the maximum correlation coefficients of input variables (factors) with the output variable. Finding the corresponding transformation B_i and C_i , where i corresponds to the i - th input variable.

Step 2. We calculate the generalized (functional) correlation coefficients and construct the correlation matrix consisting of generalized and maximum correlation coefficients.

Step 3. Selecting the input variable most strongly correlated with the output variable, i.e variable having the largest absolute correlation coefficient with the output variable. Let it be X_1. We shall note that if this variable has a different serial number, we can replace them and rename the numbering.

Step 4. As the next variable to be included in the regression model, we select the variable X_k which is characterized by the highest partial correlation coefficient with the output variable. Let it be, considering the comments on step 3, the input variable X_2 .

Step 5. We obtain the regression model $\hat{Y} = f(C_1X_1, C_2X_2)$. We investigate the contribution of the variable X_1(or, rather, C_1X_1which would have occurred if in the model was initially included a variable X_2nd then X_1 To do this, calculate the value of the private F-test and determine a statistical significance of the magnitude X_1. If X_1 is statistically significant, it remains in the model. If it is not statistically significant, it is excluded from the model. Let X_2 be statistically significant and included in the model.

Step 6. In accordance with the stepwise method, for the next variable to be included in the model, we select the input variable, which has the highest partial correlation coefficient with the output variable (assuming that the variables X_1 and X_2 are already included in the generalized regression model). We assume, considering the comments in step 3, that it is a variable X_3

Step 7. We obtain the regression model in the form $\hat{Y} = f(X_1, X_2, X_3)$ In this step, we define partial F-test for variables X_1 and X_2 in order to find out, shall we leave them in the regression equation or not.

The process of selecting informative variables continues as long as the adding or deleting the variables in the regression model process does not stop. The process of inclusion or exclusion of informative variables is completed when the remaining variables are statistically insignificant for the F-criterion or all of the measured variables are statistically significant and should be included in the final mathematical model.

Also have been examined the modification of the stagewise regression method, based on a maximum and generalized correlation coefficients.

Like in the classical stagewise regression method, the basic idea is as follows. First we construct a regression equation for Y related to the variable X_k most strongly correlated with Y, i.evariable having the highest maximum correlation coefficient with Y $\left|R_{YX_k}^{\max}\right| \geq \left|R_{YX_j}^{mzx}\right|$, $j = 1,2,...,N$ where N number of observations. We calculate the residuals between Y and the value $\hat{Y}$ by the regression equation.

These residuals are considered as the value of the new response, and the regression dependence of a new response from one of the remaining input variables that is most strongly correlated with the new response is constructed.

This process continues until any desired stage. The final regression equation for this method can be obtained by consequent substitutions of regression equations derived in the previous step *t -1* in the equation for the *t-th* stage until the final solution.

Remark. This method is less accurate than the wealthy or the generalized least squares method. However, this method has other advantages:

- it allows you to choose sequentially input variables that most strongly influence on the output variable;
- the method allows you to select control variables for inclusion in the model for the target of further use for the control or process optimization. In the future, you can use the method in its classical form to reduce the dimension of the space of input variables;
- as generalized least squares method has better accuracy and prognostic characteristics, the modified stageise regression method can be used for selection of informative variables and the choice of the structure of the regression equation. And then the selected structure of the equations and the selected informative variables can be used to obtain the regression equation by the method of least squares.

3. SOFTWARE IMPLEMENTATION AND SIMULATION

With the help of software product developed to calculate the eigenvalues and the maximum and the generalized correlation coefficient was conducted simulation experiments and comparative analysis of classical linear and a modified approach to the selection of significant variables in the model.

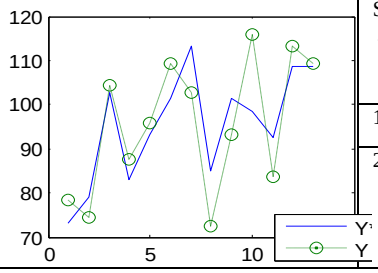
The most interesting question for us – does the $R^{\max}$ determine the statistical dependence between random variables in cases where a linear correlation does not catch it? Stepwise procedure of inclusion of variables is considering the partial pair linear correlation (i.e., purged of the influence of other variables) as a criterion for inclusion of a variable in the model. Let's consider examples of systems and interdependencies for which the methods of classical correlation theory are either inapplicable or giving large errors when using them.

To construct the regression equations we use a linear function:

$$Y^* = \sum_{i=1,m} a_i \cdot X^{(i)} \tag{6}$$

and the method of least squares.

Table 1

Type of dependence	Schedule of quality of the selection model at the last step	**Stepwise regression method of including of variable**				
Hald data: n = 13; m = 4 (The dependence is unknown)		Step, №	The variance of error	Fisher F-criteria (significance 0.05)		The variables included in the equation
				Tabulated	Calculated	
		1	73.65	4.84	22.79	$X^{(4)}$
		2	4.82	4.1	229.5	$X^{(1)}, X^{(2)}$

We shall compare on the same samples stepwise method of including and its modification, in which the criterion of partial correlation is replaced by the maximum correlation. As an example, we consider the data given by

Hald, cited in [5] of heat release per gram of cement (*Y*, in calories), depending on its composition (*X*-components of the clinker in % by weight of clinker).

Stepwise regression method of variable inclusion
Main characteristics of stepwise regression are given in Table 1.

The resulting equation looks as follows:

$$Y^* = 52.58 + 1.47 \cdot X^{(1)} + 0.66 \cdot X^{(2)} \quad (7)$$

As can be seen, selected model is very efficient and involves only two variables. In addition, F-criteria shows its high significance. Model (7) is built in 2 steps, because the remaining variables $X^{(3)}, X^{(4)}$ did not bring a significant estimation of the private F-criteria.

Modification of the stepwise method

We calculate $R^{\max}$ for each pair $\left\{X^{(j)}, Y\right\}$ For comparison, in the 3rd column of the table shown the coefficients of the classical correlation *r:*

Table 2

The compared variables	$R^{\max}$	Linear correlation coefficient *(r)*
$\left(X^{(1)}, Y\right)$	-0.5	0.73
$\left(X^{(2)}, Y\right)$	1.0	0.81
$\left(X^{(3)}, Y\right)$	-0.914	-0.53
$\left(X^{(4)}, Y\right)$	-0.901	-0.82

Visible dependence between $R^{\max}$ and r is not detected.

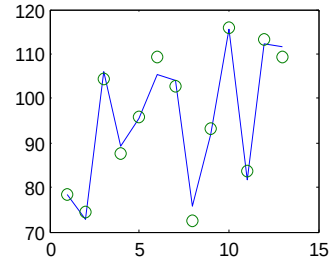
To construct the regression we will consistently add into the equation variables $X^{(j)}$ in descending order of their corresponding absolute value of $R^{\max}$: $X^{(2)} \to X^{(3)} \to X^{(4)} \to X^{(1)}$ (Table 3).

The resulting equation is:

$$Y^* = 62.4 + 1.55 \cdot X^{(1)} + 0.51 \cdot X^{(2)} + 0.1 \cdot X^{(3)} - 0.14 \cdot X^{(4)} \quad (8)$$

F-criteria shows the significance of the equation at all steps.

Table 3

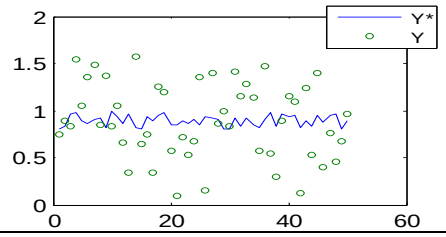
Type of dependence	Schedule of quality of the selection model at the last step	**Modification of Stepwise regression method**			
		Step, №	The variance of error	Fisher F-criteria (significance 0.05)	The variables included in the equation
				Tabulated / Calculated	
Hald data: n= 13; m = 4 (The dependence is unknown)		1	75.52	4.84 / 21.96	$X^{(2)}$
		2	34.62	4.1 / 27.68	$X^{(2)}, X^{(3)}$
		3	6.15	3.86 / 107.3	$X^{(2)}, X^{(3)}, X^{(4)}$
		4	3.98	3.83 / 111.47	$X^{(2)}, X^{(3)}, X^{(4)}, X^{(1)}$

Let's consider one more simulation experiment. Generate the following selection of variables N (0,1): $X^{(1)}, X^{(2)}, X^{(3)}$ of 50 values of each. And consider the following system:

$$Y = \left| X^{(1)} \right| + \left| X^{(2)} \right| \tag{9}$$

Since this function is additive, then we can assume that its linear regression approach will be effective. Variable $X^{(3)}$ is introduced just as an additional factor. In this case, the standard and modified procedures of stepwise regression gave the following results (Table 4,5):

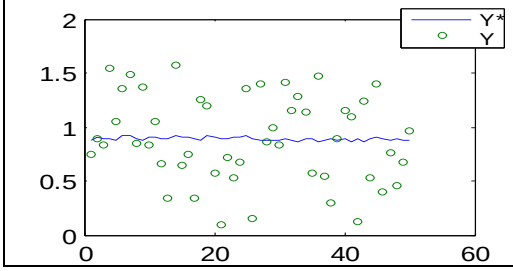
Table 4

Schedule of quality of the selection model at the last step	**Stepwise regression method of selection of variables**				
	Step, №	The dispersion of error	Fisher F-criteria (significance 0.05)		The variables included in the equation
			Tabulated	Calculated	
	1	0.157	4.04	1.02	$X^{(3)}$

On the given tables formally we cannot give preference to any of the regressions: they are both insignificant, both reaching almost the same level of variance error, both cost-effective (if a second procedure to perform only in 1 step).

However, there are qualitative differences. The first scheme included variable $X^{(3)}$ in the equation with linear correlation r = 0.14. But first, this relationship can be considered significant only with great reserve, and, secondly, $X^{(3)}$ - is not included in the function (9).

Table 5

Schedule of quality of the selection model at the last step	**Modification of Stepwise method**				
	Step, №	The variance of error	Fisher F-criteria (significance 0.05)		The variables included in the equation
			Tabulated	Calculated	
	1	0.1607	4.04	0.04	$X^{(1)}$
	2	0.1605	3.17	0.03	$X^{(1)}, X^{(2)}$
	3	0.157	2.8	0.35	$X^{(1)}, X^{(2)}, X^{(3)}$

In other words, the inclusion of this variable in the equation - a mere formality. Which can not be said for the modified method: it turned out that the order of variable inclusion was found to reflect the real dependence (9). In other words, the modified method makes it possible not to go through all possible combinations of variables, but just add a few of the most important factors to be sure that they really represent the nature of the process. And to further reduce the variance of the remainders can continue to add the observed input variables.

Thus, based on a sufficiently large number of simulation experiments (about 100 samples from 10 different functions), we can say that the maximum correlation is actually better detecting the stochastic relationship between random variables than the standard linear correlation.

RECEIVED MAY, 2012
REVISED NOVEMBER, 2012

REFERENCES

[1] PRANGISHVILI I.V., PASHCHENKO F.F. and BUSYGIN B.P. (2001): **System Laws and Regularities in Electrodynamics, Nature and Society.** Nauka, Moscow.

[2] SARMANOV O.V. (1958): The maximum correlation coefficient (symmetric case). **Doklady RAS**,.120, 715-718.

[3] SARMANOV O.V. (1960): Own correlation functions and their application in the theory of stationary Markov's processes. **Doklady RAS**, 132, 769-772.

[4] RENYJ A. (1959): On measures of dependence. **Acta. Math., Acad. Sei. Hung**. 10.

[5] DRAPER, N. and H. SMITH (1987): **Applied Regression Analysis**. Vol 2. Finances and Statistics, Moscow.

[6]PASHCHENKO A.F. (2004): Selection of informative variables in the problem of structural identification of biotechnological systems. **Journal of MASI.**, 7 , 41-48.