Identification quality assessment of the complex object management in multicollinearity conditions

Marina Potanina and Svetlana Pisariuk

1Department of Management and Business Analysts, Sevastopol State University, Universitetskaya st. 33, 299053 Sevastopol, Russia

Abstract. The article presents the method of assessing the quality of the complex object management (COM) identification process in multicollinearity conditions. The main problem in the process of identifying complex object management is assessing its quality. When operating COM, multicollinearity often arises, which significantly complicates the task. Nowadays, a lot of evaluation methods have been developed and studied. Despite this, when solving practical tasks related to processing data from experiments about the management object, the accuracy of the MNS becomes insufficient. As a criterion for the optimality of the regression model, it is proposed to use the amount of forecast error in a given area. A method for finding the optimal regularization parameter for offset estimation of regression equation parameters is proposed. A scientific substantiation of the principles of assessing the quality of the COM identification process in multicollinearity conditions are resolved by building its predictive model using the method of offset assessment of regression equation parameters. In this case, a method is proposed for selecting the regularization parameter \( r \) based on the minimum root-mean-square error of the forecast obtained from this model. The simulation experiments using the technique, shown that the value of the optimal regularization parameter \( r_{opt} \) obtained is close to the experimental one. This confirms the correctness of the proposed approach. Thus, the quality of the predictive model under conditions of multicollinearity has been improved, taking into account the uncertainty of the model structure and the method of biased estimation of model parameters.

1 Introduction

The main problem in the process of identifying complex object management (COM) is assessing its quality. When operating COM, multicollinearity often arises, which significantly complicates the task. Partial mitigation of the negative impact of multicollinearity is possible by switching to biased methods for estimating regression models, for example, ridge regression. But with a mixed assessment of parameters, the central problem is the choice of the regularization parameter \( r \). At the moment, there is no...
recommended acceptable method for building a model for evaluating this indicator, taking into account the uncertainty of the model structure and the specifics of the offset assessment method itself.

The main difficulty is that both the offset and quadratic error of the forecast depend on unknown true values of the parameters.

The paper considers the development and research of a method for selecting the optimal regularization parameter for structural identification of the COM model based on the study of the loss function as a criterion of optimality for biased assessment of the regression model for an extreme.

2 Materials and methods

Let the COM be described by the output variable $y$ and the vector of independent input variables $x^T = [x_1, x_2, ..., x_k]$, where $k$ - is the number of input variables. The value of the variable vector $x$ can be accurately calculated, measured, predicted, and set in an experiment. Variable $y$ is an indicator of the quality of operation of an object or target function. The choice of the target function is determined by the specific conditions of the process under study, as well as the possibility of operational measurement, reliability and accessibility of information, universality and statistical efficiency. Let there also be a priori information about the areas of possible variable values: $y_1 \leq E\{y\} \leq y_2$, $x \subseteq W$, where $W$ is the set of possible values of the vector $x$, $E\{\cdot\}$ - a symbol of mathematical expectation.

Identification is performed using a configurable model of a certain structure, the parameters of which may change. Let the functional dependence between variables look like

$$y = \eta(x) + \varepsilon, \quad (1)$$

where $\eta(x)$ – is a true but unknown model and an independent case value with zero mean and variance $\sigma^2$.

We consider the case in which a given structure of a class of models makes it possible to partially order a set of models. In this case, model classes are represented as nestings of one into another:

$$S_1 \subset S_2 \subset ... \subset S_q, \quad S_j$$ - set of class $j$ models. Thus, linear (according to their parameters) models can be determined by their structure in accordance with the number of their terms. Then each class $S_j$ is determined by the model:

$$\hat{\eta}(\hat{x}, \hat{\alpha}) = f^T(\hat{x})\hat{\alpha} \quad (j = 1,2,...,q), \quad (2)$$

where $f^T(\hat{x}) = [f_1(\hat{x}), f_2(\hat{x}),... , f_n(\hat{x})]$ is the vector of known functions from the vector $\hat{x}$, $n$ – the number of model parameters, $j$, $q$ – the maximum possible order of the model, $\hat{\alpha} = [a_1, a_2, ..., a_n]$ the vector of unknown parameters.

A complex control object is characterized by a large number of indicators that often correlate with each other. The model structure of this type of object is usually not defined. There is a problem of choosing an evaluation method for building predictive regression models of complex objects. Currently, much attention is being paid to this problem, a number of authors are reflected various aspects of it in their works, such as [1-23 etc.]. To date, many evaluation methods have been developed and studied. For example, the method of smallest squares (MNS), the weighted method of smallest squares, ridge regression, etc. Among linear regression adjustment parameters based on observations, the most commonly used estimates (MNS) are those that are statistically optimal when a number of conditions are met.
Despite this, when solving practical tasks related to processing data from experiments about the management object, the accuracy of the MNS becomes insufficient. The effectiveness of estimates increases when using a priori information about the studied parameters of the probabilistic or deterministic nature of regression equations. Determining the most reliable parameter values is achieved through biased estimates, which reduces the mean squared error in comparison with the MNS estimate. This work examines experimental data with high multicollinearity. Negatively affecting the assessment of regression coefficients and occurring when the experiment is not planned. The problem of multicollinearity is solved in different ways. In this situation, the use of biased ones makes it possible to eliminate the disadvantage of MNS estimation. Also, “ridge” or ridge estimates are considered more stable than MNS estimates, they have a lower value of the root mean square error of the forecast.

Let’s write the regression model as

\[ Y = F\alpha + \varepsilon, \]  

(3)

Where \( \alpha \) – is the column vector of unknown regression coefficients, 
\( F \) – matrix of functions from controlled variables, 
\( \varepsilon \) – vector-column of remnants or accidental outrage. 
MNS estimates have the form:

\[ \hat{\alpha} = (F^TF)^{-1}F^TY \]  

(4)

When solving a system of normal equations, to stabilize the least squares estimates, we will use the matrix \( (F^TF + rI_n) \), instead of \( F^TF \), where \( r \) – is the regularization coefficient (a small positive number added to the diagonal elements of the \( F^TF \) matrix), and \( I_n \) is the identity matrix. These estimates are called ridge estimates and are presented in the form:

\[ \hat{\alpha} = (F^TF + rI)^{-1}F^TY \]  

(5)

Determining the regularization parameter \( r \) when using ridge estimates is the most difficult. Due to the fact that this operation is almost impossible to automate, since the bias and square error of the forecast \( R(r) \) depend on the unknown true values of the parameters. There are various approaches to determine the regularization parameter.

Hurl and Kennard showed that the quadratic error \( R(r) \) for a regularized estimate will always be smaller than for an OLS estimate if the inequality is met

\[ 0 < r < \sigma^2/\beta_{max}^2, \]  

(6)

where \( \beta = V\alpha \), and \( V \) – is an orthogonal \((k \times k)\) matrix made up of proprietary vectors of the \( X^TX \) information matrix; \( \alpha \) – unknown values of regression coefficients; \( \beta_{max} = (|\beta_1|, |\beta_2|, ..., |\beta_k|) \).

Theobold identified sufficient minimization conditions for a more general loss function \( R_\omega (r) = E(\alpha(r) – \alpha)^T\Omega(\alpha(r) – \alpha) \), where \( \Omega \) – there is an arbitrary function. These conditions are set by inequality:

\[ 0 < r < 2\sigma^2/a^T a \]  

(7)

Obichein received the necessary and sufficient conditions for fulfilling the inequality \( R(r) < R(0) \), i.e.

\[ 0 < r < 2/|x_k|, \]  

(8)

where \( x_k \) – is the negative own number of the matrix \( (X^TX)^{-1} – (aa^T/\sigma^2) \).
All these approaches require knowledge of the true values of variance and the vector of unknown coefficients. Substituting values or estimates instead makes the choice of r stochastic and does not guarantee a reduction in the root-mean-square error.

Hurl and Kennard, based on Formula (8), propose to choose the regularization coefficient equal to

$$r = \frac{s^2}{\beta_{\text{max}}^2},$$

(9)

where $\beta_{\text{max}}$ – is the maximum modulo element of the vector $\beta = V\alpha$, and $s^2$ – is the estimate $\sigma^2$. It can be seen that in this case, the values $\sigma^2$ and $\beta^2$ in (9) are simply replaced by their estimates. Hurl and Kennard suggest adding various additives to different $X'X$ diagonal elements that regularize parameters, but this does not always lead to a significant improvement. Unfortunately, this approach also requires information that is rarely available to the researcher.

### 3 Results and discussion

In this paper, the issues of assessing the quality of the COM identification process in multicollinearity conditions are resolved by building its predictive model using the method of offset assessment of regression equation parameters. In this case, a method is proposed for selecting the regularization parameter $r$ based on the minimum root-mean-square error of the forecast obtained from this model.

An algorithm is presented for studying the loss-to-excess function, proving the existence of a local minimum, and the possibility of calculating the optimal regularization parameter.

In the form of an application of the developed methodology, the problem of obtaining the best predictive model of a complex management object was considered. All $X$ are highly correlated, their number is large, this does not allow us to accurately identify the parameters needed in the forecast model, which in turn greatly affects the accuracy of the forecast.

The use of biased estimates of the coefficients of the predictive model allows you to take advantage of the advantages of ridge regression. As a result, the forecast obtained by this model has the smallest error in the forecast area.

At the same time, the planning of a computational experiment for parametric identification of the mathematical model of the object of study satisfies the limitations imposed by the measure of correlation of experimental data.

The conformity of the customized model to the object is evaluated by the identification quality criterion, which represents the average losses of the customized model

$$I(\alpha) = \int \left( y - \varphi(x,\alpha) \right)^2 P(x, y) dx dy,$$

(10)

where $P(x,y)$ is the joint density of the distribution of variables $x$ and $y$. The smaller the average losses, the higher the quality of identification [3].

If $x$ is a non-random vector variable taking only discrete values $x_u$ ($u=1,2,...,N$), which is the case in the case of a pre-planned experiment, then the value of the average risk can be written in the form:

$$I(\alpha) = \frac{1}{N} \sum_{u=1}^{N} \int \left( y - \eta(x_u) \right)^2 P(y | x_u) dy + \frac{1}{N} \sum_{u=1}^{N} (\eta(x_u) - \varphi(x_u,\alpha))^2.$$  (11)

The first part of this formula is related to the variance of the value, and the second part is related to the quality of the proposed model. If the variance $\sigma^2$ is constant, then the value of the average risk is given by the formula:
\[
I(\alpha) = \sigma^2 + \frac{1}{N} \sum_{n=1}^{N} (\eta(x_n) - \varphi(x_n, \alpha))^2.
\]  \hfill (12)

If the dependence \( \varphi(x, \alpha) \) is determined on the basis of sample values of \( Y \), then this value will also be random and it is necessary to consider the mathematical expectation of the average risk. To obtain optimal modeling results and determine the response function \( y \) in the area \( W_1 \), it is necessary to present the loss function of model \( j \) in the form:

\[
L(j) = \int_{\omega_1} E(y - \eta_j(x, \alpha_j))^2 dx / \int_{\omega_1} dx, \quad (j=1,2,...,q)
\]  \hfill (13)

For each model \( j \) of area \( W_1 \), the root mean square error of the forecast is determined, \( y \) is presented as a possible value of the dependent variable at point \( x \). If the variance \( \sigma^2 \) is a constant, then the loss function can be represented as:

\[
L(j) = \sigma^2 + \int_{\omega_1} (\eta(x) - \eta_j(x, \alpha_j))^2 dx / \int_{\omega_1} dx.
\]  \hfill (14)

The values of the experimental data at each point allow us to determine the approximating model.

\[
x_i^T = (x_{1i}, x_{2i},.., x_{ki}), \quad (i=1,2,....,N).
\]

Let the matrix \( D^T = [x_1, x_2, ..., x_N] \) be a matrix of a plan by size \( N \times k \) and \( Y^T = [y_1, y_2, ..., y_N] \) – be a vector of \( N \) observations over the response function.

Let

\[
\hat{a}_j = \Phi_j(X,Y) - \hfill (15)
\]

estimate \( a_j \), where \( \Phi_j \) – is a function of \( Y, \) plan \( D, \) and model class \( S_j. \)

If the model is linear in parameters and the ridge regression method is used, then

\[
\hat{a}_j = (F^TF + rI_n)^{-1}F^TY, \hfill (16)
\]

where \( F_j^T = [f_j(x_1), f_j(x_2), ..., f_j(x_N)] \) – is the matrix of vector \( f_j \) values at \( N \) experimental points; \( r \) – is the regularization coefficient (a small positive number added to the diagonal elements of the \( F^TF \) matrix) and \( I_n \) – is the identity matrix.

It is determined that the matrix \( F_j \) has full rank. In the case of ridge regression for model \( q \) over the area \( W_i \), where \( x \subseteq [-1,1] \) the loss function will have the form:

\[
L(q) = E \int_{w_1} x^T(a - \hat{a}(r)) (a - \hat{a}(r))^T x dx =
\]

\[
= \sigma^2 spXBjBx^T + a^T(I_q - X^TX)f(I_q - Bx^Tx)a, \hfill (17)
\]

where \( B = (X^TX + rI_q)^{-1} \) and \( J = \int_{w_1} xx^T dx / \int_{w_1} dx. \)

Formula (17) consists of two parts. The first, containing the true coefficients \( a^T(I_q - X^TX)f(I_q - Bx^TX)a \) – is the effect of the bias of the estimates, and the second part \( \sigma^2 spXBjBx^T \) contains the variance of the error. Obviously, it is possible to find such a value of the regularization parameter at which the quadratic error of the estimates can be significantly reduced due to their small bias.

The authors have developed an iterative procedure for finding the optimal regularization parameter \( r \) in multicollinearity conditions and determining unknown parameters of the regression model. It is assumed that if the loss function in the case of ridge regression \( ER(r) \) has a local minimum, then it corresponds to the optimal value of the regularization parameter.
In order to investigate the loss function $ER(r)$ at a local extremum, it is necessary to differentiate it and equate it to zero. Then, finding the roots, determine which extremum will be obtained at this point (minimum or maximum). If the function changes the sign in the extremum area from "−" to "+", then this is the minimum, if vice versa, then this is the maximum.

Having performed the necessary transformations and calculations, we come to the conclusion that the loss function in the case of ridge regression has a local minimum, which proves the existence of an optimal regularization parameter. And its value can be calculated by the formula:

$$r = \sigma^2 sp(XJX^T)/a^TJX^Txa.$$  \hspace{1cm} (18)

It is also possible to determine the necessary and sufficient conditions for the existence of a local extremum. Obviously, in addition to $r$, this expression depends on the following factors:
- from the experiment plan;
- from the true coefficients; with large coefficients, it is difficult to determine whether there will be a local minimum at all;
- from the variance of the error.

Consequently, in the future it is possible to introduce restrictions on obtaining a local minimum.

### 3.1 Example

Let there be a class of nested models $S_j$:

- Model 1: $y_1 = a_0 + a_1x_1$;
- Model 2: $y_2 = a_0 + a_1x_1 + a_2x_2$;
- Model 3: $y_3 = a_0 + a_1x_1 + a_2x_2 + a_3x_3$.

There is a strong multicollinearity. The maximum model from the class is analyzed first in the simulation experiment: $y_3 = a_0 + a_1x_1 + a_2x_2 + a_3x_3$, the coefficients of the model are taken true, variance $\sigma^2 = 1$. The regularization parameter $r$ varies from zero to two in increments of 0.1. Calculate the value of the loss function for each $r$.

Let the matrix of the experiment plan $X_i$ have the form:

- $X_1 = \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}$;
- $X_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ -1 & 1 \\ -1 & -1 \end{bmatrix}$;
- $X_3 = \begin{bmatrix} -1 & -1 \\ -1 & -1 \\ -1 & -1 \\ -1 & -1 \end{bmatrix}$.

Software has been developed in the language of MATLAB m-files, in which, using an iterative procedure, $r$ is found, at which the derivative $\frac{dEI}{dr}$ turns to zero (this proves the existence of a local minimum) and the value of the loss function at this point is calculated. In addition, the value of the optimal $r$ is calculated using the formula (18). The results obtained confirm that the loss function has a local minimum, which is the $r_{opt}$. Table 1 shows the results of finding the optimal value of the regularization parameter of the $r_{opt}$, the values of the corresponding loss functions depending on the experimental plan. As an illustrative example, Fig. 1 shows a graph of the dependence of the loss function on the value of the regularization coefficient $r$ for plan $X_1$.

### Table 1. Simulation experiment results for Model 3

<table>
<thead>
<tr>
<th>The type of plan</th>
<th>The value of the $r_{opt}$</th>
<th>the value of the $EI_{min}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1.00</td>
<td>0.267</td>
</tr>
</tbody>
</table>
Obviously, the greater the variance and the smaller the true regression coefficients, the greater the value of the optimal \( r \), i.e. the value of the optimal regularization parameter is directly proportional to the variance and inversely proportional to the square of the coefficients of the true model. It is not possible to find the optimal \( r \) for all plans, since the derivation of the formula for \( r \) is an optimization task. For plans whose points are very different from the boundary ones, it is proposed to use an iterative procedure.

Consider the confirmation that the developed procedure for finding the optimal regularization coefficient can work for incomplete models from a given class by example. Fig. 2 shows the dependence of the loss function on the regularization parameter for all nested models.
Fig. 2. Dependence of the loss function on the regularization parameter $r$ for the entire class of nested models.

The graph confirms the correctness of the developed procedure for finding the $r_{opt}$ for any model from the nested class. The value of the $r_{opt}$ regularization parameter is obtained for different models is different.

4 Conclusion

As a result of the application of the developed methodology, a study of the loss function of a complex control object was carried out to assess the existence of a local minimum, which proved and made it possible to determine the value of the optimal regularization parameter with a biased estimation of the parameters of regression equations. After the simulation experiments, it can be seen that the value of the optimal regularization parameter $r_{opt}$ obtained as a result of using the technique, the value of the optimal regularization parameter $r_{opt}$ is close to the experimental one. This confirms the correctness of the proposed approach.

Based on all of the above, it follows that the goal has been achieved. Thus, the quality of the predictive model under conditions of multicollinearity has been improved, taking into account the uncertainty of the model structure and the method of biased estimation of model parameters. All this confirms the correctness of the used and developed control methods, the created software in the MATLAB m-file language that implements the proposed methodology.

The subject of further research is supposed to study the relationship between the construction of a model of the COM by the method of biased estimation of regression equations and the regularization criterion Ivakhnenko A.G.

References