Analysis of the properties of OLS-estimates in parameter identification of distributed dynamic processes

M G Matveev and E A Sirota
Computer science department, Voronezh State University, Universitetskaya pl. 1, Voronezh, Russia
E-mail: mgmatveev@yandex.ru, atoris@list.ru

Abstract. This work is devoted to analysis and research of OLS-estimates properties when identifying parameters of distributed dynamic processes. The study showed that at a low level of observation errors (1% and lower), the use of direct OLS estimates to identify parameters of distributed dynamic processes gives satisfactory results. At the same time, the displacement value is always slightly higher than the value of the standard deviation of the parameter estimate, which does not allow to neglect the displacement, especially at a high and average level of observation errors. The method of obtaining so-called alternative OLS-estimates is also proposed, which allows to reduce multicollinearity in sample statistics and at any level of observation errors significantly reduce standard error of parameter estimation.

1. Introduction
One of the most difficult problems of parametric identification is the estimation of changing parameters of models of distributed dynamic processes based on statistical methods [1, 2]. The scope is simplified if we assume that the parameters are constant in time and change only in spatial coordinates.

However, there remains the problem of choosing a method of evaluation that would provide acceptable values of statistical properties of the evaluation: standard error and bias. The range of estimation methods is determined by the method of obtaining statistical data. We consider a passive method consisting of observation the behavior of a scalar field value $y$ in nodes of a regular grid. In this case, the following methods are usually used for parametric estimation: the method of statistical moments, the method of maximum likelihood or the method of least squares, well-proven when working with time series of observations [3]. Object of our study is the method of least squares (OLS) - the most effective and easiest to use. The goal of the article is to study the possibility of using OLS for parametric identification of models of distributed dynamic processes.

2. Materials and methods
If the processes are properly described by linear differential equations, it is convenient to change to difference equations to obtain estimates. For example, for a homogeneous convective diffusion equation with one spatial variable, the reduced difference equation for the $i$-th node at $k + 1$ time will be as follows [4]:

$$y_i^{k+1} = a_1 y_{i-1}^k + a_2 y_i^k + a_3 y_{i+1}^k,$$

with given initial and boundary conditions:
\[ y_i^0 = c_i, \quad y_{i-1}^k = b_{i-1}^k, \quad y_{i+1}^k = b_{i+1}^k, \quad \forall k, \]

where \( t \) - discrete values of the spatial coordinate, \( k \) - discrete time; in general, the necessary index \( i \) of the parameters \( a = (a_1; a_2; a_3) \) is omitted henceforth for simplicity; condition \( a_1 + a_2 + a_3 = 1 \) ensures conservativeness of the scheme. Here should be recalled, that the scheme is called conservative if it reflects on the grid the same conservation laws that were presented in the original differential problem [5].

The values of the variable \( y_i^k \) are measured at each node \( i \) with an error \( \xi_i^k \) generated by a random white noise process. The measured value will be denoted by \( x_i^k = y_i^k + \xi_i^k \).

Then the expressions (1) and (2) can be rewritten in the form of an autoregressive dependence, describing non-stationary time series:

\[
x_i^{k+1} = a_1(x_{i-1}^k - \xi_{i-1}^k) + a_2(x_i^k - \xi_i^k) + a_3(x_{i+1}^k - \xi_{i+1}^k) + \xi_i^{k+1} = a^T x^k + \omega_i, \quad \omega_i = \xi_i^{k+1} - a^T \xi^k.
\]

Values of the initial and boundary conditions are also determined by the measurement results. As a rule, time series in adjacent nodes of the approximating grid are strongly correlated with each other. Together with non-stationarity, this causes problems of obtaining OLS estimates of a with reasonable statistical properties. In particular, in this case the parameter estimates will be biased [3]. Unbiased estimates can be obtained using the methods of stationarization [6], related procedures and criteria for determining the type of process, significantly complicating the algorithms for finding estimates compared to the direct application of OLS. Therefore, it seems relevant to investigate the limits of applicability of direct use of OLS, especially considering that in a number of our previous studies of non-stationary processes of temperature fields changes in the atmosphere [7, 8, 9, 10] OLS demonstrated quite acceptable results.

The presence of bias, in general, does not always determine the rejection of OLS, because the quality of the estimation is given by the combination of the standard error of the estimation and its bias. In this regard, it is relevant to study these properties when evaluating autoregression parameters (3) with conditions (2) at various levels of observation interference. We will also be interested in how the properties of the estimates change when changes the degree of correlation between the regressors.

We will investigate the relation of the standard error of parameter estimation, \( \sigma \) and displacement, \( \Delta \) on the level of observation interference, \( \xi \) and the pair correlation coefficient \( r_{ij} \) between the \( i \)-th and \( j \)-th time series at the corresponding nodes. Statistical data for the study were obtained in the form of an analytical solution of the homogeneous equation of convective diffusion [4]

\[
\frac{\partial y}{\partial t} + v \frac{\partial y}{\partial x} = D \frac{\partial^2 y}{\partial x^2}
\]

\[ y(0, l) = \varphi(l), \]

\[ y(t, l_{\text{min}}) = f_1(t), \quad y(t, l_{\text{max}}) = f_2(t), \]

\[ y(t, l) = \exp\left(\frac{v}{2D}(l - \frac{vt}{2})\right) (e^{-\delta t} \sin(l) + e^{-4\delta t} \sin(2l) + e^{-9\delta t} \sin(3l)).\]

The solution \( y(t, l) \) was discretized at the nodes of a regular grid \( t_k, l_i \). In each node an additive noise \( \xi_i^k \), which intensity varied during the study, was added to the obtained value \( y_i^k \).

The following procedure was used to estimate the effect of series multicollinearity on the standard error and estimation bias. Since time series in adjacent nodes are strongly correlated, it can be assumed that there is a linear dependence of the levels of the corresponding time series on a convex linear combination

\[ y_i^k = \beta_1 y_{i-1}^k + \beta_2 y_{i+1}^k, \]

where
\[ y_i^k = x_i^k - \xi_i^k, \beta_1 + \beta_2 = 1 \]

Thus
\[ x_i^k = \beta_1 x_{i-1}^k + \beta_2 x_{i+1}^k + \varepsilon^k, \]  
(5)

where
\[ \varepsilon^k = \xi_i^k - \beta_1 \xi_{i-1}^k - \beta_2 \xi_{i+1}^k. \]

Substitute expression (5) in (3), we obtain:
\[ x_i^k = (a_1 + a_2 \beta_1) x_{i-1}^k + (a_2 \beta_2 + a_3) x_{i+1}^k + \xi_i^{k+1} - \xi_{\sum}^k = \theta_1 x_{i-1}^k + \theta_2 x_{i+1}^k + \Delta \xi, \]  
(6)

where \( \xi_{\sum}^k \) - convex linear combination of the interferences \( \xi_{i-1}^k, \xi_i^k, \xi_{i+1}^k. \)

Expression (6) allows to construct a system of linear equations with respect to parameters \( a_1, a_2, a_3: \)
\[
\begin{align*}
    a_1 + a_2 \beta_1 &= \theta_1, \\
    a_2 \beta_2 + a_3 &= \theta_2, \\
    a_1 + a_2 + a_3 &= 1.
\end{align*}
\]  
(7)

In the future, the parameters’ estimates \( a \) obtained by the solution of system (7) will be called alternative OLS estimates, as opposed to direct OLS estimates \( \hat{a} \). Estimations of parameters \( \beta_1 \) and \( \beta_2 \) can be obtained in the following form
\[ \hat{\beta} = (X^T X)^{-1} X^T x_k, \]  
(8)

where \( X \) is the matrix of observations in the nodes \( i - 1 \) and \( i + 1; x_k \) is the vector of observations in the node \( i \) at time \( k. \)

Estimates of the parameters \( \theta_1 \) and \( \theta_2 \) are obtained as
\[ \hat{\theta} = (X^T X)^{-1} X^T x_{k+1}, \]  
(9)

where \( x_{k+1} \) is the vector of observations in a node \( i \) at time \( k + 1. \)

The estimates (8) and (9) will be biased. The value of the offset is specified by the corresponding expressions:
\[ \Delta \beta = (X^T X)^{-1} X^T \varepsilon \]  
and
\[ \Delta \theta = (X^T X)^{-1} X^T \Delta \xi. \]  
(10)

Alternative OLS estimates of the parameter vector \( a \) will differ from the corresponding estimates obtained using direct OLS. The difference will mainly be due to the different degree of correlation of regressors, since the correlation of non-adjacent regressors in (5) and (6) will be lower than the correlation of adjacent regressors in (3). Direct OLS uses three regressors with high pair correlation values, and an alternative approach uses two regressors, whose correlation is objectively lower. The alternative approach corresponds to a transition from the original, ill-conditioned system of equations to a better-conditioned equivalent system.

3. Results and discussion

The sample statistics required for the model study was obtained using the analytical solution (4) of the differential equation (5). The analytical solution was discretized in time and spatial coordinate. The difference equation (1) was obtained with the following parameter values: \( a_1 = 0.2583, a_2 = 0.5, a_3 = 0.2417. \)

In accordance with the research methodology, an additive observation noise was added to the values of the variable in the corresponding grid nodes, obtained using a Gaussian-type independent random number generator with zero expectation and unit variance. The intensity of the interference was set at
the levels: $c = 0.1$, $c = 0.01$, $c = 0.0025$, $c = 0.001$. The rate of interference was 40%, respectively; 4%; 1% and 0.4% of the RMS amplitude of the useful signal, which allowed to divide the obtained statistics into data with high (40%), medium (4%) and low (1% and 0.4%) levels of interference.

Change of multicollinearity was achieved by the choice of the method of obtaining solutions for problems of parameters identification of either direct OLS estimates for equation (3), alternatives to the OLS estimates in equations (7). The correlation matrix of regressors with elements averaged over the grid nodes has the following form:

$$
\begin{bmatrix}
  x_{i-1} & x_i & x_{i+1} \\
  1 & 0.98 & 0.76 \\
  0.98 & 1 & 0.97 \\
  0.76 & 0.97 & 1
\end{bmatrix}
$$

It is easy to see that the correlation level in adjacent nodes differs from the corresponding value in non-contiguous nodes by about 20%.

To obtain reliable results of estimates of bias and standard error, allowing comparison of numerical values, the experiments in all modes were repeated 1000 times and their results were averaged. Direct OLS estimates of parameters $c$ as well as alternative OLS estimates are given in table 1 for different levels of observation interference. To analyze the effectiveness of the considered methods, table 2 and table 3 show the values of the offsets for each method, along with the standard errors of the estimated parameters.

| Table 1. Comparison of OLS estimates at different interference intensities. |
|---------------------------------------------------------------|
| $c$ | Direct OLS-estimates | $\geq 0.97$ | Alternative OLS-estimates | $= 0.76$ |
|     | $a_1$ | $a_2$ | $a_3$ | $\sum a$ | $a_1$ | $a_2$ | $a_3$ | $\sum a$ |
| 0.1  | 0.3958 | 0.3293 | 0.2661 | 0.9912 | 0.3904 | 0.3310 | 0.2786 | 1 |
| 0.01 | 0.3520 | 0.3375 | 0.3100 | 0.9995 | 0.3483 | 0.3399 | 0.3118 | 1 |
| 0.0025 | 0.2732 | 0.4730 | 0.2540 | 1.0002 | 0.2657 | 0.4889 | 0.2454 | 1 |
| 0.001 | 0.2613 | 0.4951 | 0.2434 | 0.9998 | 0.2594 | 0.4992 | 0.2415 | 1 |
| True values | 0.2583 | 0.5000 | 0.2417 | 1 | 0.2583 | 0.5000 | 0.2417 | 1 |

Analysis of table 1 shows that reasonable parameter estimates in both approaches are achieved only at low noise levels. But the fundamental property of conservatism persists with high accuracy at all levels of interference, even in direct OLS estimates, where the ratio is not explicitly embedded in the solution of the identification problem.

Tables 2 and 3 provide an opportunity for a more detailed analysis of the results.

| Table 2. Estimates of the average value of the displacement of the considered methods at different noise intensities. |
|---------------------------------------------------------------|
| $c$ | Offset of direct OLS estimates, $r_{xx} \geq 0.97$ | Offset of alternative OLS estimates, $r_{xx} = 0.76$ |
|     | $\Delta a_1$ | $\Delta a_2$ | $\Delta a_3$ | $\Delta a_1$ | $\Delta a_2$ | $\Delta a_3$ |
| 0.1  | 0.1375 | 0.1707 | 0.0244 | 0.1321 | 0.1690 | 0.0369 |
| 0.01 | 0.0937 | 0.1625 | 0.0683 | 0.0900 | 0.1601 | 0.0701 |
| 0.0025 | -0.0140 | 0.0270 | 0.0125 | 0.0074 | 0.0111 | -0.0037 |
| 0.001 | -0.0030 | 0.0049 | -0.0017 | -0.0010 | 0.0008 | 0.0002 |
Table 3. Values of the average standard errors of the estimated parameters of the considered methods at different noise intensities.

| c    | Mean standard error values | Standard errors of direct OLS estimates | Standard errors of alternative OLS evaluations |
|------|---------------------------|----------------------------------------|-----------------------------------------------|
|      |                           | \( \delta_{a1} \) | \( \delta_{a2} \) | \( \delta_{a3} \) | \( \delta_{a1} \) | \( \delta_{a2} \) | \( \delta_{a3} \) |
| 0.1  |                           | 0.0265        | 0.0274        | 0.0263        | 0.0027        | 0.0028        | 0.0026        |
| 0.01 |                           | 0.0157        | 0.0256        | 0.0112        | 0.0015        | 0.0024        | 0.0011        |
| 0.0025 |                          | 0.0047        | 0.0077        | 0.0034        | 0.0012        | 0.0009        | 0.0012        |
| 0.001|                           | 0.0018        | 0.0029        | 0.0013        | 0.0002        | 0.0003        | 0.0001        |

It is convenient to carry out the analysis of the obtained results based on two sources of bias and standard error: the presence of observation error and multicollinearity.

The bias of direct OLS estimates in the presence of correlation of regressors with observation errors is determined as follows

\[
\hat{a} = (X^T X)^{-1} X^T x = (X^T X)^{-1} X^T (Xa + \omega) = a + (X^T X)^{-1} X^T \omega, \tag{11}
\]

\[
E(\hat{a}) = a + E(\left( X^T X \right)^{-1} X^T \omega), \tag{12}
\]

where \( E(\left( X^T X \right)^{-1} X^T \omega) \) - the value of the offset of the direct OLS-estimate.

An approximate expression of the bias of alternative OLS estimates can be obtained considering (10) as follows

\[
\hat{a} = \hat{\beta}^{-1} \hat{\beta} = \hat{\beta}^{-1}(X^T X)^{-1} X^T (X\theta + \Delta \xi) \approx a + \hat{\beta}^{-1}(X^T X)^{-1} X^T \Delta \xi, \tag{13}
\]

\[
E(\hat{a}) = a + E(\left[ \hat{\beta}^{-1}(X^T X)^{-1} X^T \Delta \xi \right]), \tag{14}
\]

where \( E(\left[ \hat{\beta}^{-1}(X^T X)^{-1} X^T \Delta \xi \right]) \) - approximate offset value of alternative estimation; \( \hat{\beta} \) - the matrix of the system (8), where the nondegeneracy is due to the biased estimates \( \beta \).

Expressions (13) and (15) show that the magnitude of the bias depends on two factors: the interference of observation, \( \xi_{ik} \), and the error of the matrix inversion \( (X^T X) \) associated with its poor conditionality. The conditionality numbers of this matrix for various observation interferences are given in Table 4.

Table 4. Condition numbers of the matrix \( X^T X \).

| c    | For direct OLS estimates | For alternative OLS estimates |
|------|--------------------------|------------------------------|
| 0.1  | 40                       | 16                           |
| 0.01 | 100                      | 60                           |
| 0.0025 | 4.1 \cdot 10^5         | 2.1 \cdot 10^5               |
| 0.001 | 8.6 \cdot 10^5          | 3.2 \cdot 10^3               |

The data of Table 4 show that a significant change in the number of conditionality of the matrix \( X^T X \) due to the application of an alternative approach is achieved only at low level of interference. As a consequence, at high and medium levels of interference, the change in the number of conditionality practically does not affect the value of the bias (Table 2). Hence, it can be concluded, that at high and medium levels of observation interference, they have a dominant influence on the magnitude of the bias estimates \( \alpha \) as in both approaches. At low noise, the error of the matrix \( X^T X \) begins to play the dominant role in the formation of the bias. In this case, usage of alternative OLS estimates significantly reduces the bias (Table 2).

The standard error of parameter \( a_i \) estimation is defined as follows

\[
\delta_{ai} = \sigma_{\xi} \sqrt{\frac{1}{n_i}}, \tag{15}
\]
where \( x_{ii} \) is the corresponding element of the main diagonal of the matrix \((X^TX)^{-1}\). That means, the standard error depends on the same factors as the bias—the mean square deviation of the observation interference, \( \sigma_\xi \) and the matrix reversal error \((X^TX)\), the analysis of table 3 shows a clear advantage of alternative OLS estimates at all levels of observation errors.

4. Conclusion
The study showed that at a low level of observation errors (1% and below), usage of direct OLS estimates to identify the parameters of distributed dynamic processes gives reasonable results. In this case, the value of the bias is always slightly greater than the standard deviation of the parameter estimate, which does not allow to neglect the bias, especially at high and average levels of observation errors. The use of alternative OLS estimates allows to reduce multicollinearity in sample statistics and at any level of observation errors significantly reduces the standard error of parameter estimation. Alternative OLS estimates have a positive effect on the displacement value only at a low level of observation errors.

Direct OLS estimates of parameters can be improved by using alternative OLS estimates, but only at a low level of observation errors.

References
[1] Guo L Z, Billing S A and Coca D 2010 Identification of partial differential equation models for a class of multiscale spatio-temporal dynamical systems Instr. J. Contr. 83 40-8
[2] Xun X, Cao J, Mallick B, Carrol R and Maity A J 2013 Parameter Estimation of Partial Differential Equation Models J. Am. Stat. Assoc. 108 1009-20
[3] Nosko V P 2011 Econometrics vol 1 ed Yu V Luiso (Moscow: Delo)
[4] Matveev M G, Kopytin A V and Sirota E A 2018 Parameters identification of a distributed dynamical model using combined approach J. Phys. Conf. Ser. 1096 012068
[5] Samarskii A A 2001 The Theory of Difference Schemes (Boca Raton: CRC Press)
[6] Grebenyuk E A 2005 Methods of analyzing the nonstationary time series with implicit changes in their properties Autom. Remote Control 66 1871-96
[7] Corduas M and Piccolo D 2008 Time series clustering and classification by the autoregressive metric Comput. Stat. Data Anal. 52 1860-72
[8] Matveev M G, Kopytin A V, Sirota E A and Kopytina E A 2017 Modeling of Nonstationary Distributed Processes on the Basis of Multidimensional Time Series Procedia Engineering 201 511-6
[9] Isermann R and Münchhof M 2011 Least Squares Parameter Estimation for Dynamic Processes Identification of Dynamic Systems (Berlin Heidelberg: Springer) chapter 9 pp 223-290
[10] Matveev M G, Sirota E A, Semenov M E and Kopytin A V 2017 Verification of the convective diffusion process based on the analysis of multidimensional time series CEUR Workshop Proc. 2022 354-8
[11] Kini B V and Sekhar C C 2013 Large margin mixture of AR models for time series classification Applied Soft Computing 13 361-71