Reliable Computing (2005) 11: 433–442
DOI: 10.1007/s11155-005-0050-3

On Fitting Empirical Data under Interval Error

SERGEI I. ZHILIN
Mathematical Department, Altai State University, 90, Krasnoarmeiskii prosp., 656049 Barnaul, Russia, e-mail: sergei@asu.ru

(Received: 25 October 2004; accepted: 30 January 2005)

Abstract. This paper is devoted to the problem of fitting input-output data by a modeling function, linear in its parameters, in the presence of interval-bounded errors in output variable. A method for outlier detection is proposed. Another issue under consideration is the comparative simulation study of the well-known statistical point estimates (least squares, maximum likelihood) and point estimates calculated as the center of interval hull of uncertainty set. The results of the study allow us to draw the conclusion that non-statistical interval based estimation is a competitive alternative to statistical estimation in some cases.

1. Introduction

The statement of the problem of curve or surface fitting for empirical data under the assumption that the values of dependent variable contain bounded uncertainties, considered in this paper, originates from the idea of L. V. Kantorovich [4] and has been extensively studied e.g. in [1], [5], [8]–[11].

The essence of the problem is to construct a linear parameterized modeling function

$$y = \sum_{i=1}^{n} \beta_i x_i, \tag{1.1}$$

where $x \in \mathbb{R}^n$ is a vector of input variables, $\beta \in \mathbb{R}^n$ is a vector of parameters to be estimated, y is a scalar output variable.

The modeling function is constructed from empirical information in which the table of experimental data obtained in N observations,

$$T = \{(y_j, x_{1j}, \ldots, x_{nj}) \mid j = 1, 2, \ldots, N\},$$

plays the lead. It is assumed that the measurement errors of input variables x_i may be neglected, and the value of the output variable y in the j-th observation is measured with the error which belongs to the interval $[-\varepsilon_j, \varepsilon_j]$.

The fact that the errors of the output variable are bounded may be expressed in the form of the following bilateral inequalities:

$$y_j - \varepsilon_j \le \sum_{i=1}^{n} \beta_i x_{ij} \le y_j + \varepsilon_j, \qquad j = 1, \ldots, N. \tag{1.2}$$

All the values of the parameters vector $\beta = (\beta_1, \ldots, \beta_n)$ which satisfy every inequality (1.2) form the set B of possible values of the parameters also called an "uncertainty set."

Solutions of the linear programming problems

$$\underline{\beta}_i = \min_{\beta \in B} \beta_i, \qquad \overline{\beta}_i = \max_{\beta \in B} \beta_i, \qquad i = 1, 2, \ldots, n \tag{1.3}$$

provide us with the lower and upper bounds of the possible values of the model parameters. Cartesian product of the intervals $[\underline{\beta}_i, \overline{\beta}_i]$ gives the minimum bounding box of the uncertainty set B. The intervals $[\underline{\beta}_i, \overline{\beta}_i]$ are often used as interval estimates of the parameter β_i, and the middle points of these intervals may serve as sought-for point estimates

$$\hat{\beta}_i = (\underline{\beta}_i + \overline{\beta}_i) / 2. \tag{1.4}$$

Apart from the problem of estimation of the modeling function parameters β, the problem of interval and point estimation of the output variable y for some known values of the input variable x (forcasting problem) may be stated in respect to the uncertainty set B. The bounds of the interval estimate $[\underline{y}_i(x), \overline{y}_i(x)]$ can be found by solving the linear programming problems

$$\underline{y}_i(x) = \min_{\beta \in B} \sum_{i=1}^{n} \beta_i x_i, \qquad \overline{y}_i(x) = \max_{\beta \in B} \sum_{i=1}^{n} \beta_i x_i, \qquad i = 1, 2, \ldots, n. \tag{1.5}$$

Averaging the interval estimate bounds gives the point estimate $\hat{y}(x)$ of the output variable:

$$\hat{y}(x) = \frac{1}{2}\big(\underline{y}(x) + \overline{y}(x)\big). \tag{1.6}$$

However all the above estimation problems make sense only if the uncertainty set B is nonempty and bounded. The unboundedness of the set B can be found out in the rank analysis of the observation matrix $X = (x_{ij})_{N \times n}$, and it may be interpreted as a lack of empirical information for the construction of the model. The emptiness of B means inconsistency of the collected empirical information. The presence of outliers in the observation data is one of the possible reasons of contradictions. In this paper, a simple method for outliers detection is proposed.

Another question considered in the paper is a relationship between the interval-based estimates and traditional statistical estimates, namely maximum likelihood estimates and least squares estimates. We present a comparative study of the issue based on extensive simulation. In this case, an analytical comparison would be impossible because the estimation methods compared are based upon different systems of hypotheses, and this is why simulation is the only available approach [2]. Throughout this paper the estimates (1.4)–(1.6) are called *non-statistical* for brevity.

2. Outlier Detection

From practical viewpoint, a possibility to detect conflicts in collected empirical information is one of the most important properties of the non-statistical approach to parameter estimation described in the introduction. The indicator of contradiction is the fact that the uncertainty set is empty.

The main sources of contradictions are

- wrong hypothesis about modeling function structure;
- presence of outliers.

To choose the way for overcoming these troubles, one must perform a comprehensive data analysis. However, the results of such an analysis are determined by the information one can get, and the outlier detection method proposed in this section may be regarded as a tool to obtain information necessary for the analysis.

An outlier is, by definition, a peculiar, non-typical observation. It means that the outliers must be thoroughly examined in order to reveal the causes of their appearance. In some cases, an outlier gives information, which cannot be obtained from other observations because it is a result of a measurement under unusual combination of conditions. In such a situation, further extended investigation is necessary. However, more frequently outliers are the result of blunders during the measurement of the observed variables. In this case, an outlier should be rejected or must be taken into account with a relatively low weight.

The core idea of the outlier detection method proposed below is as follows. An outlier caused by a blunder may be treated as a value that is measured with an underestimated error, i.e. whose real measurement error is greater than the declared error. In order to correct the outlier, it is necessary to find the lower bound of its possible actual error, which makes the corrected observation consistent with the others. Comparing the values of the lower bound of possible real error to the values of the declared observation error allows us to make some inferences concerning the degree of inconsistency of the outlier with respect to the entire data set.

The lower bounds ε'_j of possible real errors providing a non-empty uncertainty set may be treated as the product of the declared errors ε_j and unknown scale coefficients w_j: $\varepsilon'_j = w_j\varepsilon_j$, $j = 1, 2, \ldots, N$. The desired values of the scale coefficients may be found as a solution of the following problem

$$\min_{\beta, w} \sum_{j=1}^{N} w_j, \tag{2.1}$$

$$y_j - w_j\varepsilon_j \le \sum_{j=1}^{N} \beta_j x_j \le y_j + w_j\varepsilon_j, \qquad w_j \ge 1, \;\; j = 1, 2, \ldots, N. \tag{2.2}$$

In the solution of (2.1)–(2.2), the values of j for which the resulting scale coefficients w_j are greater than the unity correspond to outliers. If the researcher has information that, for some observations, the errors are equal (for example, in the case they are

obtained using the same measurement facility and techniques), the equalities of the form $w_{j_1} = w_{j_2} = \cdots = w_{j_K}$ may be added to the constraints system (2.2) of the problem (2.1). If the researcher is sure about some declared error values, he can "freeze" the corresponding scale coefficients w_j setting them to ones.

A large number of the scale coefficients w_j greater than the unity in the solution of (2.1)–(2.2) may be caused by overestimating the precision of measurement instrument or a wrong choice of the modeling function structure.

To conclude, it should be noted that the proposed approach is closely related to the theory of improper linear programming problems correction [3] and may be regarded as one of possible ways to parameterize an improper linear programming problem in order to construct its approximation by a proper linear programming problem and to correct it at a minimal cost.

3. Experimental Comparison of Statistical and Non-Statistical Estimates

The main difference in the systems of basic hypotheses of the statistical and non-statistical (bounding) approaches to parameter estimation is the hypothesis on error structure.

In statistical approaches, the error is supposed to be a random variable with a distribution selected by the researcher. In practice, the distribution is often assumed Gaussian normal. It is well known that the least squares method (which is a particular form of the maximum likelihood method) provides the most qualitative estimates (consistent and efficient) in this case. However, the error normality assumption is not always justified [6], [7]. Furthermore, in most cases the researcher has no strong reason to take specific distribution as belonging to this or that predetermined parametric class.

The main principle of the non-statistical approach to data processing, which determines all the algorithms and conclusions, is that all elements of the error interval and therefore all elements of the uncertainty set B are equally possible and feasible [8].

We propose to clear up the descriptive strength of both statistical and non-statistical fitting methods through simulation, which is organized in the following way. Each iteration of the simulation consists of a model data generation step and a step that solves a point forecasting problem for the generated data using both statistical and non-statistical estimation methods. The deviations of the estimated values of the output variable from its real value (forecast errors) are accumulated through all the iterations in order to compute standard deviations of estimates obtained by each method. In every iteration, the model data are generated by adding simulated error (random value that belongs to the known distribution) to a value of some known modeling function for fixed values of the explanatory variables.

The point estimates are chosen as the main objects for comparison because of the possibility to similarly interpret the statistical and non-statistical point estimates

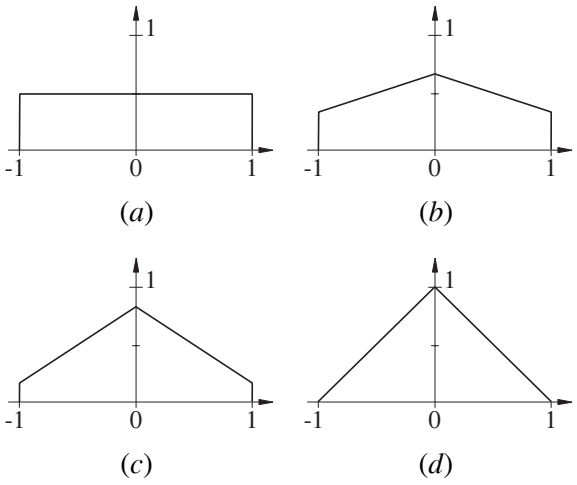

Figure 1. Probability density function (3.1) for $\varepsilon = 1$ and (*a*) $\alpha = 0$, (*b*) $\alpha = 1/3$, (*c*) $\alpha = 2/3$, (*d*) $\alpha = 1$.

(while it is too difficult in the case of interval estimates) and because they are of great importance in practice.

As for the distribution of simulated errors in the model data, it is of interest to examine situations best for each of the compared methods as well as some intermediate variants. Statistical methods provide good results when the distribution of error is unimodal, in particular, for the least squares method it should be normal. Apparently, the most adequate way to ensure the main principle of the non-statistical estimation approach in a statistical manner is to supply the uniform error distribution. Therefore, it is advisable to perform the comparative simulation study under normally and uniformly distributed error as well as for some intermediate distributions of the error.

3.1. Non-Statistical Estimates and Maximum Likelihood Estimates

To compare non-statistical and maximum likelihood estimates, we use the following parametric class of probability density functions (p.d.f.):

$$p_\alpha(x) = \begin{cases} \dfrac{\alpha}{\varepsilon^2}x + \dfrac{\alpha+1}{2\varepsilon}, & -\varepsilon \le x < 0, \\ -\dfrac{\alpha}{\varepsilon^2}x + \dfrac{\alpha+1}{2\varepsilon}, & 0 \le x \le \varepsilon, \end{cases} \tag{3.1}$$

where ε is the absolute value of the error bound, $\alpha \in [0, 1]$ is a parameter that determines the degree of proximity of (3.1) to the triangular p.d.f. For $\varepsilon = 1$, the functions (3.1) that correspond to the boundary and two intermediate values of α are plotted in Figure 1.

The simulation process is described by the following pseudo-code.

ALGORITHM 1.

input

$y = f(x, \beta)$ — function in the form (1.1)
β^* — exact values of the model parameters
x^* — argument value which the forecasting problem is to be solved for
ε — mid-width of the error interval
Q — multiplicity of the simulated observations
K — number of simulation iterations
M — step number for p.d.f. evolution from the uniform to triangular form

output

d_m^s, d_m^n — standard deviation of forecast errors for statistical (maximum likelihood) and non-statistical method respectively

begin

$y^* := f(x^*, \beta^*)$
$X :=$ GENERATE_EXPERIMENT_PLAN
for $m := 0$ **to** M **do**
$d_m^s := 0$
$d_m^n := 0$
$\alpha_m := \frac{m}{M}$
for $k := 1$ **to** K **do**

{ *Data generation* }
$j := 1$
for each row $x \in X$ **do**
for $q := 1$ **to** Q **do**
$e :=$ random value from interval $[-\varepsilon, \varepsilon]$ with p.d.f. $p_{\alpha_m}(x)$
$x_j := x$
$y_j := f(x_j, \beta^*) + e$
$j := j + 1$
end for
end for

{ *Estimation* }
$\hat{\beta}^s :=$ MLE(x, y)
$\hat{\beta}^n :=$ NONSTAT(x, y, ε)
$\hat{y}^s := f(x^*, \hat{\beta}^s)$
$\hat{y}^n := f(x^*, \hat{\beta}^n)$

{ *Accumulation of squared deviations* }
$d_m^s := d_m^s + (\hat{y}^s - y^*)^2$
$d_m^n := d_m^n + (\hat{y}^n - y^*)^2$
end for

```
        { Computation of standard deviation }
        d^s_m := d^s_m / K
        d^n_m := d^n_m / K
    end for
end
```

The subroutine MLE, used in Algorithm 1, returns maximum likelihood estimates for the data passed by its arguments. The implementation of the maximum likelihood method by this subroutine has the following feature. For p.d.f.'s that are close to the uniform p.d.f., the computation of the maximum likelihood estimate entails difficulties related to non-uniqueness of the maximum of the likelihood function. We overcome these difficulties by the regularization of the optimization problem, which assumes adding, to the likelihood function $L(\beta)$, a regularizing term in the form of $\delta|L(\beta)|(\beta - \beta^*)^2$, where $\delta < 0$ is the relative weight constant (in our simulation, $\delta = -0.01$), and β^* is a known vector of exact values of the modeling function parameters.

The subroutine NONSTAT computes non-statistical estimates according to (1.2)–(1.4).

The function GENERATE_EXPERIMENT_PLAN generates the values of the input variables in the simulated measurements table. In the performed simulation, a simple function with two input variables (one of them is dummy) and unity parameters, $y = f(x, \beta) = x + 1$, was used as modeling function, i.e. $\beta^* = (1, 1)$. The generated plan of experiment X consisted of the records $(x_{1i}, 1)$, where $x_{1i} = i$, $i = 1, 2, \ldots, 10$.

The other parameters of Algorithm 1 had the following values: $x^* = (5.5, 1)$; $\epsilon = 0.5$; $M = 10$; $K = 5000$; $Q = 3$.

The values of the resulting standard deviation of the statistical and non-statistical estimates depending on m are plotted in Figure 2.

The analysis of the simulation results allows us to draw the following conclusions. For the error p.d.f.'s that are close to the triangular one, the maximum likelihood estimate is more efficient than the non-statistical one. The situation can be naturally explained by the fact that the statistical estimation method uses additional information (in the form of error distribution) which the non-statistical method does not use. However, when the error distribution is close to the uniform distribution, the standard deviation of the non-statistical estimate of the predicted value noticeably decreases, while the deviation of the statistical one grows up and even exceeds it. This may be explained by the following: the error distribution that are close to the uniform one gives less information for the maximum likelihood method than the triangular one, but in such cases the basic non-statistical principle becomes more adequate. The zero standard deviation of the statistical forecast for $m = 0$ is determined by relatively large contribution of regularizing term to the likelihood function.

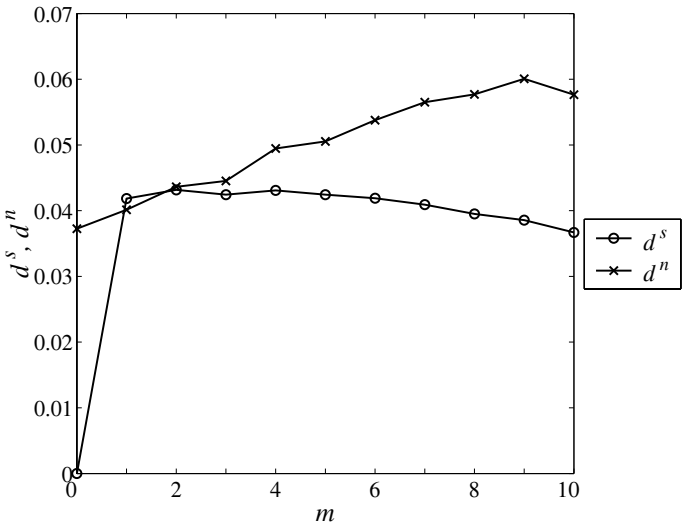

Figure 2. The standard deviation of the maximum likelihood and non-statistical forecasts from exact value for error distribution with p.d.f. $p_{\alpha_m}(x)$.

3.2. Non-Statistical Estimates and Least Squares Estimates

The procedure and parameters of the experimental comparison of the non-statistical forecast and the least squares forecast that we used are much the same as described in the previous subsection. The only exceptions are the error distribution and observation multiplicity.

Instead of $p_\alpha(x)$, the class $N_k(a, \sigma^2)$ of normal error distributions truncated on a level k is used in this experiment (i.e. the error is bounded by the interval $[a - k\sigma,\ a + k\sigma]$, where a is mathematical expectation, σ is mean square deviation). The values of a and σ were set equal to 0 and 1 respectively. The class parameter k plays the same part for $N_k(a, \sigma^2)$ as α for $p_\alpha(x)$. The value of k changes in the interval $[0.5, 3]$ with the step 0.25. So, the error distribution changes from nearly uniform to the nearly normal one as k increases.

In order to reveal the dynamics of the standard deviations of forecast errors depending on observations multiplicity the experiment is conducted for each of the following values of Q: 1, 3, and 9.

The results of the experiments are depicted in Figure 3.

Analyzing the standard deviations of the non-statistical and the least squares forecast errors depending on the truncation level k and observations multiplicity Q we can see that

- when the characteristics of the error distribution correspond to the hypotheses of the least squares method, the least squares estimates are more efficient, but in the range of nearly uniform distributions the efficiency of the non-statistical estimates becomes comparable;

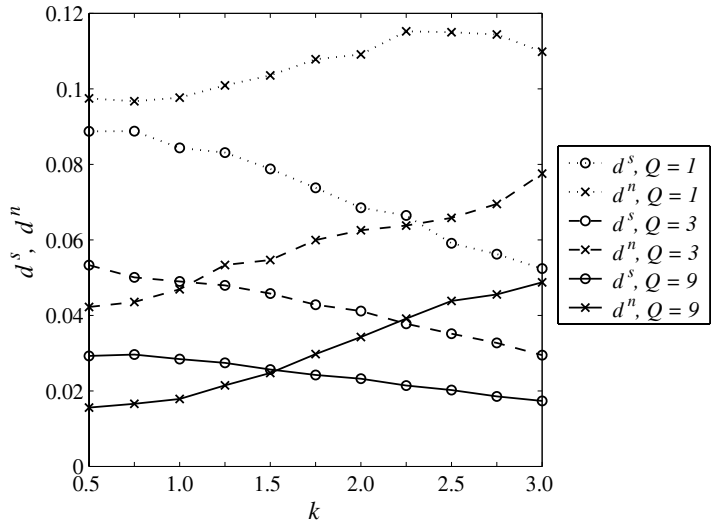

Figure 3. Standard deviation of the least squares estimates and non-statistical estimates depending on the truncation level of normal distribution of the error k and observation multiplicity Q.

- in the range of nearly uniform error distributions, as observation multiplicity grows up the efficiency of non-statistical estimates increases faster than that for the least squares estimates. This fact may be considered as an evidence that the non-statistical estimation method implicitly accumulates information about the error.

Therefore, the results of the simulation performed allow us to conclude that, in case of bounded errors and lack of information on error distributions, the non-statistical estimation approach may be a competitive alternative to the classical statistical methods of maximum likelihood and least squares. In spite of the fact that the non-statistical estimation method uses less empirical information than the statistical techniques, the efficiency of the point non-statistical estimates is comparable to that of statistical estimates, at least when the error distribution is close to uniform.

References

1. Belov, V. M., Sukhanov, V. A., Guzeev, V. V., and Unger, F. G.: Estimation of Linear Physicochemical Dependencies by Rectangle of Uncertainty Center Method, *Izvestia Vuzov. Fizika* **8** (1991), pp. 35–45 (in Russian).
2. Borodyuk, V. P.: Comment I to the Article by Voshchinin A. P., Bochkov A. F., Sotirov G. R. "A Method for Data Analysis in the Presence of Interval Non-Statistical Error," *Zavodskaya Laboratoriya* **7** (56) (1990), pp. 81–83 (in Russian).
3. Eremin, I. I.: *Contradictory Models of Optimal Planning*, Nauka, Moscow, 1988 (in Russian).
4. Kantorovich, L. V.: On Some New Approaches to Numerical Methods and Observation Processing, *Sib. Math. Zhurnal* **5** (3) (1962), pp. 701–709 (in Russian).

5. Milanese, M. and Belforte, G.: Estimation Theory and Uncertainty Intervals Evaluation in Presence of Unknown but Bounded Errors: Linear Families of Models and Estimators, *IEEE Transactions on Automatic Control* **2** (27) (1982), pp. 408–414.
6. Novitskii, P. V. and Zograph, I. A.: *Estimating the Measurement Errors*, Energoatomizdat, Leningrad, 1985 (in Russian).
7. Orlov, A. I.: How Often Is the Observation Normal? *Zavodskaya Laboratoriya* **7** (57) (1991), pp. 64–66 (in Russian).
8. Oskorbin, N. M., Maksimov, A. V., and Zhilin, S. I.: Construction and Analysis of Empirical Dependencies by Uncertainty Center Method, *Transactions of Altai State University* **1** (1998), pp. 35–38 (in Russian).
9. Spivak, S. I.: Detailed Analysis of Application of Linear Programming Methods to Estimation of Parameters of Kinetic Models, in: *Matematicheskie Problemy Khimii* **2** (1975), VC SO AN SSSR, Novosibirsk, pp. 35–42 (in Russian).
10. Rodionova, O. Ye. and Pomerantsev, A. L.: Antioxidants Activity Prediction Using DSC Measurements and SIC Data Processing, in: *Proceedings of Second Conference on Experimental Methods in Physics of Heterogeneous Condensed Media*, Barnaul, 2001, pp. 239–246.
11. Voshchinin, A. P., Bochkov, A. F., and Sotirov G. R.: A Method for Data Analysis in the Presence of Interval Non-Statistical Error, *Zavodskaya Laboratoriya* **7** (56) (1990), pp. 76–81 (in Russian).