Robust Minimax Adaptive Approach to Regression Problems in Interval Computations

Georgiy L. Shevlyakov

I. Introduction

One of the basic approaches to the synthesis of robust estimation procedures is the minimax principle. In this case, in a given class of densities, the least favorable one (which minimizes the so-called Fisher information, see definitions below) is determined. The unknown parameters of a regression model are then estimated by means of the maximum likelihood method for this density [1].

As a result, if we know the set of possible deviations of the actual probability distribution from the model that we are currently using, then we can construct *robust* statistical procedures, i.e., procedures which are stable with respect to (possible) deviations from apriori distribution model.

The robust minimax procedures provide a guaranteed level of the estimator's accuracy (measured by the supremum of an asymptotic variance) for any density distribution of a given class.

The form of the solution obtained by the minimax approach essentially depends upon the characteristics of a distribution class. In many real-life cases, the only information that we have about the distribution consists of its tolerance interval. Thus, it is important for quality control applications to design and implement robust methods that correspond to the class of all distributions with given tolerance intervals.

In this paper, we describe minimax $M-$estimators of a location parameter and regression parameters designed for this class. The properties of the new estimators and of their adaptive versions are studied in asymptotics and in a finite sample size case.

II. Robust Minimax $M-$Estimators

Consider the following linear regression model

$$X = \Phi\Theta + E, \tag{1}$$

where:

- $X = (x_1, ..., x_n)^T$ is a vector of observed values;
- $\Theta = (\theta_1, ..., \theta_m)^T$ is a vector of unknown parameters;
- $\Phi = (\phi_{ij})$, $1 \le i \le n$, $1 \le j \le m$, is a given model matrix, and
- $E = (e_1, ..., e_n)^T$ is a vector of (unobservable) random errors with an (unknown) density f belonging to a certain (known) class $\mathcal{F}$.

The $M-estimate$ of a regression parameter Θ is defined as [1]

$$\hat{\Theta}_n = \arg\min_{\Theta} \sum_{i=1}^{n} \rho(x_i - \sum_{j=1}^{m} \theta_j \phi_{ij}), \tag{2}$$

or

$$\sum_{i=1}^{n} \psi(x_i - \sum_{j=1}^{m} \hat{\theta}_j \phi_{ij})\phi_{ik} = 0, \quad k = 1, ..., m, \tag{3}$$

where $\rho(u)$ is a function (called a *loss function*), and $\psi(u) = \rho'(u)$. It is usually required that a function ψ belong to a certain class Ψ.

The idea of the minimax approach is, for a given class $\mathcal{F}$, to choose an $M-$estimate for which the worst case least square error is the smallest possible. According to [1], this choice of estimate can be done as follows: first, we determine the least favorable density f for the class $\mathcal{F}$ by minimizing the Fisher information $I(f)$ for a location parameter:

$$f^* = \arg\min_f I(f), \quad I(f) = \int_{-\infty}^{\infty} (f'/f)^2 f dx, \tag{4}$$

and then design the optimum maximum likelihood estimate (MLE) for this density f^*, i.e., an $M-$estimate with the following choice of a loss function (and its derivative):

$$\rho^*(u) = -\log f^*(u), \quad \psi^*(u) = -(f^*)'(u)/f^*(u). \tag{5}$$

Under some regularity and convexity conditions on the classes $\mathcal{F}$ and Ψ [2], the asymptotic covariance matrix $V(\psi, f)$ has the saddle point (ψ^*, f^*) with the desired minimax property:

$$V(\psi^*, f) \le V(\psi^*, f^*) \le V(\psi, f^*).$$

The author is with Department of Mathematics, St. Petersburg State Technical University, St. Petersburg, 195251, Russia.

The left part of this inequality is important for practice: it means that the use of $\psi = \psi^*$ provides the guaranteed level of the accuracy of robust minimax estimates.

The above-mentioned conditions on f include conditions of nonnegativeness, symmetry, etc.:

$$f(x) \geq 0, \;\; f(-x) = f(x), \;\; \int_{-\infty}^{\infty} f(x)\, dx = 1. \quad (6)$$

Symmetry is a rather restrictive condition but it provides the minimax property of robust estimates and hence is widely used in robustness research [1].

Depending on additional restrictions on the class $\mathcal{F}$, different forms of the density f^* and the appropriate loss functions ρ^* may result.

It follows from [1] that, for the class of approximately finite densities

$$\mathcal{F} = \{f : \int_{-l}^{l} f(x)\, dx = 1 - \beta\} \; , \; 0 \leq \beta \leq 1 \quad (7)$$

the least favorable density has the exponential tails:

$$f^*(x;\; A, B, C, D, l) = \begin{cases} A\cos^2(Bx), & \mid x \mid \leq l, \\ C\exp(-D|x|), & \mid x \mid > l, \end{cases} \quad (8)$$

where the values $A = A(\beta, l)$, $B = B(\beta, l)$, $C = C(\beta, l)$, and $D = D(\beta, l)$ are chosen to satisfy the following conditions:

- conditions (6);
- the condition that $f^* \in \mathcal{F}$; and
- the condition that the resulting (piecewise defined) function f^* is smooth at the point $x = l$.

These conditions can be described by the following equations:

$$\int_{-\infty}^{\infty} f^*(x)\, dx = 1 \; , \; \int_{-l}^{l} f^*(x)\, dx = 1 - \beta,$$

$$f^*(l{-}0) = f^*(l{+}0) \; , \; (f^*(l{-}0))' = (f^*(l{+}0))' \; . \quad (9)$$

The exponential tails of the least favorable density (8) and the corresponding form of the robust loss function $\rho^* = |x|$ for $|x| > l$ imply that the observed data with $|x| \geq l$ are simply ignored ("rejected") when we apply this method. This rejection is similar to the least modules method [1]. The smaller l, the more data is rejected.

The following result can be easily proven.

Theorem 1 *The least favorable density (4) for the class with inequality constraint on a tolerance interval*

$$\mathcal{F}_1 = \{f : \; F^{-1}(1 - \beta/2) - F^{-1}(\beta/2) \leq b\},$$

$$F(x) = \int_{-\infty}^{x} f(x)\, dx \quad (10)$$

is of the form (8) with $l = b/2$.

III. The Least Favorable Density for the Class with Constraints on Tolerance Intervals

Consider the class $\mathcal{F}_2$ with the inequality constraints on tolerance intervals:

$$\mathcal{F}_2 = \{f : \; F^{-1}(1 - \beta_1/2) - F^{-1}(\beta_1/2) \leq b_1,$$

$$F^{-1}(1 - \beta_2/2) - F^{-1}(\beta_2/2) \leq b_2\} \quad (11)$$

with $0 \leq \beta_2 \leq \beta_1 \leq 1, \; b_1 \leq b_2$. The following result is valid in this case.

Theorem 2 *The least favorable density (4) for the class $\mathcal{F}_2$ is of the form:*

$$f^*(x) = \begin{cases} f^*(x;\; A_1, B_1, C_1, D_1, b_1/2) \text{ if } b_2/b_1 \leq k_1, \\ f^*(x;\; A, B, C, D, b) \text{ if } k_1 < b_2/b_1 \leq k_2, \\ f^*(x;\; A_2, B_2, C_2, D_2, b_2/2) \text{ if } b_2/b_1 > k_2, \end{cases} \quad (12)$$

where:

- *the function $f^*(x;\; A, B, C, D, b)$ is defined by the equations (8);*
- *the values of the parameters $A_1 \div D_1$ are determined as $A_1 = A(\beta_1, b_1/2)$, $B_1 = B(\beta_1, b_1/2)$, $C_1 = C(\beta_1, b_1/2)$, $D_1 = D(\beta_1, b_1/2)$;*
- *the values of the parameters $A_2 \div D_2$ are determined as $A_2 = A(\beta_2, b_2/2)$, $B_2 = B(\beta_2, b_2/2)$, $C_2 = C(\beta_2, b_2/2)$, $D_2 = D(\beta_2, b_2/2)$;*
- *the values of the the parameters A, B, C, D and b $(b_1/2 < b < b_2/2)$ are determined from the equations*

$$\int_{-\infty}^{\infty} f^*(x;\; A, B, C, D, b)\, dx = 1,$$

$$\int_{-b_1/2}^{b_1/2} f^*(x;\; A, B, C, D, b)\, dx = 1 - \beta_1,$$

$$\int_{-b_2/2}^{b_2/2} f^*(x;\; A, B, C, D, b)\, dx = 1 - \beta_2,$$

$$f^*(b{-}0;\; A, B, C, D, b) = f^*(b{+}0;\; A, B, C, D, b),$$

$$(f^*)'(b - 0;\; A, B, C, D, b) =$$

$$(f^*)'(b + 0;\; A, B, C, D, b);$$

- *and, finally, the values of the switching parameters k_1 and k_2 of the solution (12) are derived from the equations*

$$\int_{0}^{k_1 b_2/2} f^*(x;\; A_1, B_1, C_1, D_1, \frac{b_1}{2})\, dx = \frac{1 - \beta_2}{2};$$

$$\int_{0}^{k_2 b_1/2} f^*(x;\; A_2, B_2, C_2, C_2, \frac{b_2}{2})\, dx = \frac{1 - \beta_1}{2}.$$

Various properties of the solution (12) are connected with the degree in which the constraints are taken into account:

- in the first zone ($b_2/b_1 \leq k_1$), only the first restriction matters;
- in the third zone ($b_2/b_1 > k_2$) only the second restriction is essential;
- in the intermediate zone, both restrictions are used.

From (5) and (12), we can conclude that:

- for relatively small distribution dispersion (in the first zone), the "mild" robust algorithm based on $f^*(x;\ A_1, B_1, C_1, D_1, b_1/2)$ is optimal;
- for (relatively) large distribution dispersion (in the third zone), the hard robust algorithm (with the hard rejection of sample elements) based on $f^*(x;\ A_2, B_2, C_2, D_2, b_2/2)$ is optimal;
- in the middle zone, a compromise between these algorithms is the best solution.

IV. Robust Adaptive $M-$Estimators

Considering the problems of designing robust estimators we have assumed the availability of apriori information on the characteristics of a distribution class. However, in practice, these characteristics are often apriori unknown, and can only be determined during the data processing itself. It is therefore necessary to develop estimators which can adapt to the ever increasing volume of data, and correspondingly correct the characteristics of the distribution class.

In the adaptive procedure, the characteristics b_1 and b_2 of the class $\mathcal{F}_2$ (defined by equation (11)) are estimated and then used in the robust minimax algorithm processing the same sample $x_1, ..., x_n$. With not very large sample sizes, such an approach is heuristic and the simplest for the examination by Monte Carlo technique.

Consider the adaptive algorithm of robust estimation of regression parameters (2), called *ARLI - regression*:

(i) Choose initial hard robust estimates for Θ : $\hat{\Theta}_{HR}$ from (2) using the least modules method with the loss function

$$\rho(x) = |x|.$$

(ii) Evaluate the errors estimates

$$\hat{e}_i = x_i - \sum_{j=1}^{m} \hat{\theta}_{j\,HR}\,\phi_{ij}\,, \quad i = 1, ..., n.$$

(iii) Evaluate the estimates of the characteristics of the class $\mathcal{F}_2$:

$$\hat{b}_i = \hat{e}_{(k_i)} - \hat{e}_{(l_i)},\ i = 1, 2$$

,

$$l_i = [\beta_i n/2], k_i = n - l_i.$$

(iv) Use the robust minimax algorithm (2) with the loss function $\rho^* = -\log f^*$ according to the formulas (5) and (12).

The behavior of the ARLI-regression algorithm with respect to finite size samples ($n = 20 \div 100$ with step 10 and $100 \div 1000$ with step 100) with $\beta_1 = 0.9$, $\beta_2 = 0.1$ was studied by Monte Carlo technique. The results of simulation show high efficiency of the ARLI-algorithm under the mixture (contamination) models of data distributions [1].

References

[1] P. J. Huber, **Robust Statistics** Wiley, New York, 1981.

[2] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions", *Proc. of the 5-th Berkeley Symp. on Math. Stat. and Prob.*, 1967, Berkeley Univ. California Press, Vol. 1, pp. 221–223.