объединенный
институт
ядерных
исследований
дубна

E10-86-282

N.I.Chernov, G.A.Ososkov

JOINT ROBUST ESTIMATES OF LOCATION AND SCALE PARAMETERS

Submitted to "X Prague Conference on Information Theory, Statistical Decision Functions and Random Processes", CSSR, July, 1986

1986

1. Introduction. We consider the robust estimation of a location parameter or regression coefficients in some models arising in the particle track recognition problems of high energy physics. We use the gross-error model of the contaminated distribution of errors

$$f(x)=(1-\varepsilon)\varphi(x)+\varepsilon h(x) \qquad (1)$$

with $\varphi(x)=(2\pi\sigma^2)^{-1/2}\exp(-x^2/2\sigma^2)$ and some long-tailed noise distribution $h(x)$ specified below.

In the case of automatic scanning the experimental data obtained from particle track detectors consist of a useful ("good") part related to the track to be found, as well as of signals of background tracks, fiducials and other noise points. The noise points are usually uniformly distributed. This is the reason why we suppose $h(x)$ in (1) to be uniform: $h(x)=h_0$ in a sufficiently large interval I_h of the length $1/h_0 \gg \sigma$ and $\varepsilon > 1/2$ (even close to 1).

These models are usually explored by the pattern recognition or clustering methods. The robust estimates are also applicable in these cases, but with certain modifications or auxiliary means. We propose one of these modifications and show its high efficiency in the regression model by Monte-Carlo method. Such an approach may be useful for the robust estimation theory itself.

2. The choice of the weight function for M-estimation. It is convenient to begin with a one-parameter model of estimating the location parameter $a=Ex$ from a sample $x_1,x_2,\dots,x_n$, where $x\sim a+ +(1-\varepsilon)\varphi(x)+\varepsilon h(x)$, the functions φ, h are described above. We use Huber's M-estimates /3/

$$L(a,\sigma) = \sum_i \rho\left(\frac{x_i-a}{\sigma}\right) \to \inf_a \qquad (2)$$

or

$$a = \frac{\sum w_i x_i}{\sum w_i} \qquad (3)$$

with the weights $w_i=w((x_i-a)/\sigma)$, where $w(t)=\rho'(t)/t$ is the weight function of the estimator. The usual requirements on (2) are: the function $\rho(t)$ must be even, C^2-smooth, not decreasing for $t>0$, $\rho(0)=0$, $\rho(t)\sim t^2/2$ as $t\to 0$ (i.e. $w(t)\to 1$ as $t\to 0$), the function $w(t)\geqslant 0$ and does not increase for $t>0$, and the estimate (2) must be shift- and scale-invariant, as well as the estimate of σ if its value is unknown.

The problem of the choice of the function $\rho(t)$ (or $w(t)$) is widely discussed in the literature on robust statistics. Unbounded convex functions $\rho(t)$ provide the uniqueness of the estimate (2), its consistency, asymptotic normality in some models and a certain minimax efficiency /3,7/. But these estimates are practically unsuitable for heavy contaminated data models with $\varepsilon > 1/2$ and asymmetric, not unimodal function $h(t)$.

The M-estimators with bounded functions $\rho(t)$ are very robust in these cases, but there are many difficulties in their use. The first one is that there is almost no theoretical foundation for the use of such functions. In particular, they all are obtained by their authors heuristically. In any case there are certain objections against their application - /3/.

We shall demonstrate that the maximum likelihood estimation in the framework of our model straightforwardly leads to a bounded function $\rho(t)$ in (2). Evaluating the corresponding likelihood equation

$$\frac{\partial}{\partial a}\sum \ln \left(\frac{1-\varepsilon}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-a)^2}{2\sigma^2}} + \varepsilon h_o\right) = 0$$

we obtain $a=\sum w_i x_i/\sum w_i$, where $w_i=w((x_i-a)/\sigma)$ with the weight function

$$w(t)=w_U(t)=\frac{1+c}{1+ce^{t\ /2}} \tag{4}$$

with $c=\sqrt{2\pi}\,\sigma\, h_o\,\varepsilon/(1-\varepsilon)$ (the factor $1+c$ is introduced in (4) to fulfil $w(0)=1$). The weight function (4) corresponds to the bounded function

$$\rho(t)=(1+c)\ln\frac{c+1}{c+e^{-t\ /2}}\ . \tag{5}$$

The function (4) has no scale parameter ($w(t)\neq w_o(t/c)$). The only parameter c is the ratio of the mean number of noise observations within an interval of the length $\sqrt{2\pi}\,\sigma$ to the mean number of useful observations in the sample. It is determined by the contamination of data not in the whole range of the sample but within its essential part where all useful observations are practically concentrated (for instance, in the interval $(a-3\sigma, a+3\sigma)$). The value of c is often approximately known in experimental models.

The upper bound of (5) $(1+c)\ln(1+1/c)$ increases without limit as $c\to 0$ (with the noise diminishing). Hence the boundedness of this function is significant only for $c>0.1$ which corresponds to heavy contamination. Fig.1 shows the function (4) with $c=0.2$ compared to Tukey's bi-square weight /6/

$$w_T(t)=\left\{(1-(t/c_T)^2)^2 \text{ for } |t|<c_T \text{ and } 0 \text{ for } |t|\geqslant c_T\right\} \tag{6}$$

with $c_T=4$. These functions are close to each other, but (6) is more preferable due to faster computation. We shall suppose that $w(t)=w_T(t)$ in our further considerations.

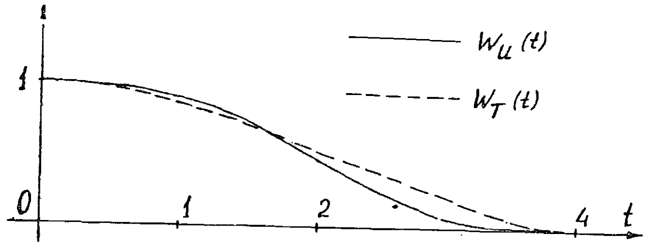

Fig. 1

3. Estimation of σ. Another problem caused by the use of bounded functions $\rho(t)$ is connected with the non-uniqueness of the estimate (2). The function $L(\underline{a},\sigma)$ often has several minima. It is difficult to find them all and to choose one of them for estimating a. The number of minima and their location depends on the value of σ, i.e., the problems of M-estimating a and σ are closely related. J.O.Ramsay /4/ also notes that separate procedures for estimating a and σ are unwise. Our approach is based upon the joint M-estimates of a and σ.

It is a difficult problem to estimate the parameter σ in our model. The common robust estimate $\hat{\sigma}=\text{const}\cdot\text{med}\left\{|x_i-a|\right\}$ is unavailable for $\varepsilon > 1/2$ as well as the other estimates based on the order statistics. From the likelihood equation for σ

$$\frac{\partial}{\partial \sigma}\sum \ln \left(\frac{1-\varepsilon}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-a)^2}{2\sigma^2}} + \varepsilon h_o\right) = 0$$

we obtain

$$\sigma^2 = \frac{\sum w_i (x_i-a)^2}{\sum w_i} \tag{7}$$

with w_i defined in (4). This estimate is applicable for $\varepsilon > 1/2$, too.

It was proposed by J.O.Ramsay /4/, S.A.Aivazyan et al./1/ with different functions w(t). It also satisfies Huber's definition of M-estimate of σ through the solution of the following equation /3/

$$\sum_i \chi\left(\frac{x_i-a}{\sigma}\right) = 0 \qquad (8)$$

with an even function $\chi(t)$. The estimate (7) corresponds to (8) when $\chi(t)=t^2w(t)-w(t)$. Therefore it is shift- and scale-invariant. Using the same weight function w(t) in (3) and (7) we can consider a- and σ-estimating as a single problem.

4. Some special geometric properties of M-estimates (3),(7). Let us consider the function $L(a,\sigma)$ in (2) as a two-parameter function. The set of local conditional minima of $L(a,\sigma)$ for all fixed $\sigma > 0$ form a finite collection of smooth curves in the semi-plane $\{(a,\sigma),\sigma>0\}$. Denote them $\gamma_1,\gamma_2,\dots,\gamma_m$. There is $\sigma_1 > 0$ such that the semi-plane $\{(a,\sigma),\sigma>\sigma_1\}$ contains only one of these curves which is infinite and has the asymptote $a=(x_1+x_2+\dots+x_n)/n$ as $\sigma\to\infty$ (we denote this curve by γ_1).

In fig.2 we give two examples of the surface $z=L(a,\sigma)$ for the samples each containing 20 points grouped into three and two clusters, correspondingly. On these surfaces one can see "ravines" (sometimes quite curved), the number of which increases as $\sigma\to 0$.

Fig.3 represents the curves $\gamma_1, \gamma_2,\dots,\gamma_m$ which are the "bottoms" of these ravines (the corresponding samples are marked with asterisks). These curves are always disconnected except the special case of symmetric samples like 3b, which lead to branching γ_1 into two different curves. The upper ends of the curves $\gamma_1,\gamma_2,\dots,\gamma_m$ usually lay on the axis $\sigma=0$, but there are some exceptions. At the lower ends of these curves $\partial^2L/\partial a^2=0$, i.e., the tangents at these points are parallel to the axis $\sigma=0$.

In order to find the estimate (7) on the curves $\gamma_1,\gamma_2,\dots,\gamma_m$ consider the function $M(a,\sigma)=\sigma^2\sum w_i-\sum w_i(x_i-a)^2$. It is easy to show that $M(a,\sigma)>0$ for the curve γ_1 for large σ ($\sigma>\text{const}$) and $M(a,\sigma)<$ <0 at the endpoints of these curves on the axis $\sigma=0$. However, it does not mean the existence of a solution $M(a,\sigma)=0$ in any of the curves $\gamma_1,\dots,\gamma_m$ and, morover, there are even examples of absence of such solutions on these curves.

This annoying situation can be overcome if we replace (7) by

$$\sigma^2 = \frac{\sum w_i(x_i-a)^2}{\sum \varkappa_i}, \qquad (9)$$

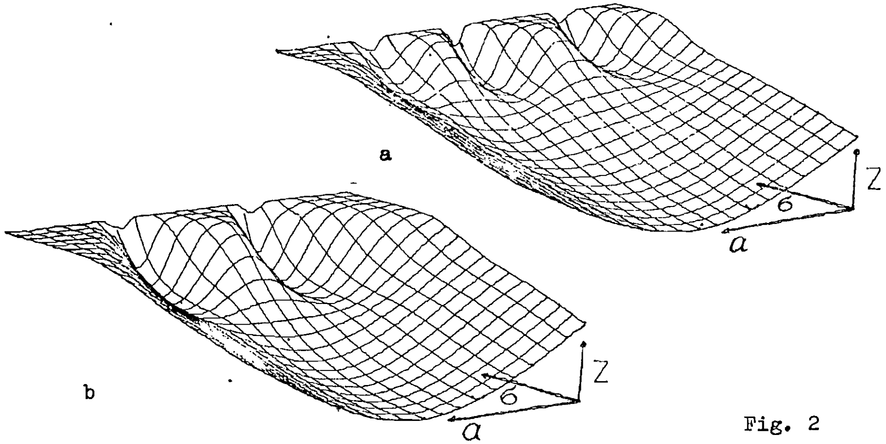

Fig. 2

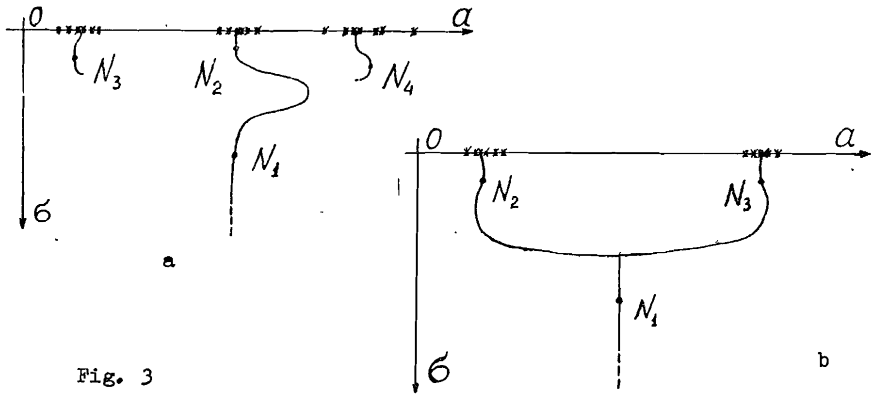

Fig. 3

where $\varkappa_i=\varkappa((x_i-a)/\sigma)$, $\varkappa(t)=\rho''(t)$. The idea of substituting $\sum\varkappa_i$ for $\sum w_i$ has come up in calculating the local $L(a,\sigma)$ minimum by Newton's iteration method

$$a = a_0-\frac{\partial L/\partial a}{\partial^2 L/\partial a^2} = a_0+\frac{\sum w_i(x_i-a_0)}{\sum\varkappa_i}$$

(a_0 is an initial value). Replacing $\sum\varkappa_i$ by $\sum w_i$ we obtain exactly (3). It is easy to check that $\varkappa(t)\leq w(t)$, hence the above substitution can slightly increase the estimate of σ. Nevertheless, it will be still robust because $\varkappa(t)\equiv 0$ if $w(t)\equiv 0$ and $\varkappa(t)\to 1$ as $t\to 0$. The estimate (9) also satisfies the definition (8), where $\chi(t)=t^2w(t)-\varkappa(t)$, hence it is shift- and scale-invariant.

The estimate (9) is more suitable for our purposes, because the behaviour of $M_1(a,\sigma)=\sigma^2\sum\varkappa_i-\sum w_i(x_i-a)^2$ is quite similiar to $M(a,\sigma)$ mentioned above, but with the important exception: $M_1(a,\sigma)<0$ at each endpoint of the curves $\gamma_1,\dots,\gamma_m$ (both for $\sigma=0$ and $\sigma>0$). Therefore the curve γ_1 contains at least one solution $M_1(a,\sigma)=0$. Moreover, there must be the solution with $\partial M_1/\partial\sigma>0$ as the necessary condition of the σ-estimate (otherwise the part of the sample in the interval $(a-c_T\sigma, a+c_T\sigma)$ is concentrated at the ends of the considered interval, where $\varkappa((x-a)/\sigma)<0$ which is in contradiction with the normal distribution of useful observations).

The obtained joint estimate (2),(9) can be defined by the set of equations containing the function $L(a,\sigma)$ only:

$$\begin{cases} \dfrac{\partial L}{\partial a}=0 \\ \dfrac{\partial^2 L}{\partial a^2}+\dfrac{1}{\sigma}\dfrac{\partial L}{\partial\sigma}=0 \end{cases} \qquad (10)$$

with the following conditions: $\partial^2L/\partial a^2>0$, $\partial/\partial\sigma\,(\partial^2L/\partial a^2 + 1/\sigma\;\partial L/\partial\sigma)>0$. Note that (10) defines just the maximum likelihood estimate in the classical case of $\varepsilon=0$ and $\rho(t)=t^2/2$.

The estimates obtained from (10) are denoted by $N_1, N_2, \dots$ in fig.3. Studying many different samples simulated by the computer we noticed the cases of just one estimate (10) in the whole area $\sigma>0$ as well as the cases with many simultaneous estimates (9) on several curves $\gamma_1, \gamma_2, \dots$. The multiplicity of solutions (10) corresponds to the existence of clusters in the sample $x_1,\dots,x_n$.

5. <u>Remark</u>. Some authors modify the function $L(a,\sigma)$ in such a way that its minima $\inf\limits_{a,\sigma} L(a,\sigma)$ would be joint M-estimates of a,σ. For example, $L_1(a,\sigma)=\sigma(L(a,\sigma)+\text{const})\to\inf\limits_{a,\sigma}$ provides the joint estimate (3),(8) with $\chi(t)=t^2w(t)-\rho(t)-\text{const}$ (see /^{2}/). Another example is

$$L_2(a,\sigma)=L(a,\sigma)-\frac{\sum w_ix_i^2\sum w_i-(\sum w_ix_i)^2}{2\sigma^2(\sum w_i)^2}+$$

$$+\left(\frac{\sigma^2(\sum w_i)^2}{\sum w_ix_i^2\sum w_i-(\sum w_ix_i)^2}-1\right)^2\to\inf_{a,\sigma}$$

which gives exactly the joint estimate (3),(7).

One more promising way is to generalize the likelihood function in the normal distribution case

$$-\ln\left(\prod_i\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-a)^2}{2\sigma^2}}\right)=\sum\frac{(x_i-a)^2}{2\sigma^2}+(\ln\sigma+\ln\sqrt{2\pi})n$$

as follows:

$$L_3(a,\sigma)=\sum\rho\left(\frac{x_i-a}{\sigma}\right)+(\ln\sigma+\ln\sqrt{2\pi})\sum w_i\to\inf_{a,\sigma}.$$

Unfortunately, this estimate is not scale-invariant and is therefore unsuitable for practical use.

6. <u>Algorithm for computation of M-estimate</u>. The existence of at least one estimate (10) on the curve χ_1 suggests the following algorithm. Starting from some point (a_0,σ_0) on the curve γ_1 with sufficiently large σ_0 and $a_0\approx(x_1+\dots+x_n)/n$ we move along the curve γ_1 decreasing σ and looking through all solutions of (10). The one closest to the axis $\sigma=0$ must be chosen as the M-estimate.

In our model $\sigma\ll 1/h_0$, i.e. σ is much less than the range of the sample. Therefore we can simplify our algorithm by moving along γ_1 without looking for solutions of (10) but stopping when we reach a small threshold $\sigma=\sigma_{min}$.

7. <u>Monte-Carlo results</u>. We studied a linear regression model $y=ax+b$ with a uniform contamination arising in special emulsion data processing in high energy physics /^{5}/. The sample consisted of N_0 "good" points $y_i=ax_i+b+\varepsilon_i$ and N_1 noise points uniformly distributed in the square $(0,1)\times(0,1)$. The factors x_i were uniformly distributed in the intervals $((i-1)/N_0, i/N_0)$, $i=1,2,\dots,N_0$ and $\varepsilon_i\sim N(0,10^{-6})$. The parameters a,b were uniformly distributed in the domain $\{(a,b): 0<b<1,\ 0<a+b<1,\ |a|<\text{tg}\,30°\}$.

The parameters a,b were estimated by the iterational reweighted least-square procedure

$$\begin{cases} a^{(k)}\sum w_ix_i^2+b^{(k)}\sum w_ix_i=\sum w_ix_iy_i \\ a^{(k)}\sum w_ix_i+b^{(k)}\sum w_i=\sum w_iy_i \end{cases}$$

with $w_i=w_T((y_i-a^{(k-1)}x_i-b^{(k-1)})/\sigma^{(k-1)})$ by c=3.5. Starting with $\sigma^{(0)}=1$ we diminished σ very slowly from iteration to iteration: $\sigma^{(k)}=(1-\delta)\sigma^{(k-1)}$ ($\delta=0.05$), in order to retain the point

$(a^{(k)}, b^{(k)})$ in the neighbourhood of γ_1. The procedure stopped at $\sigma^{(k)} < \sigma_{min}=0.001$. The obtained estimates $\hat{a},\hat{b}$ were considered as correct if they differed from the "true" values a,b by less than 0.001.

Fig.4 shows the percentage of the cases of correct estimating to the total number of simulated samples. One can note rather high efficiency of the algorithm even for $N_1=10N_0$ (i.e. ε =0.9 in the model (1)). The efficiency increases when the sample size augments and N_1/N_0 remains fixed, that indicates the possible consistency of the obtained estimate for any $\varepsilon < 0.9$.

The authors are grateful to V.K.Khoromskaya and S.V.Kunyaev for valuable help in preparing the manuscript.

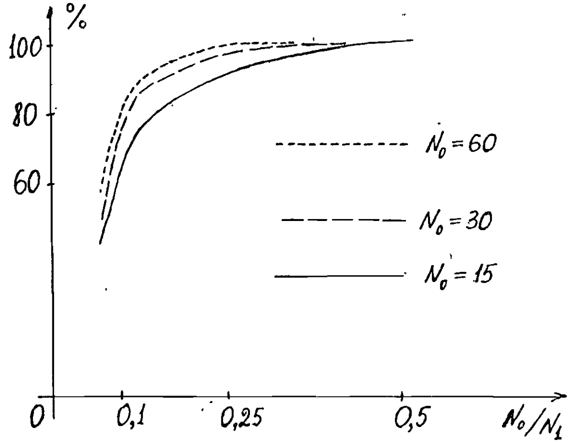

Fig. 4

REFERENCES

1. Aivazyan S.A., Yenyukov I.S., Meshalkin L.D. Applied Statistics. Study of Relationships. Fin. i Statist., Moscow, 1985.
2. Huber P.J. Mathem. Operationsforschung und Statist., Ser. Statist., 1977, 8(1), p.41-53.
3. Huber P.J. Robust Statistics. J.Willey and Sons., New York, 1981.
4. Ramsay J.O. J. Amer. Statist. Assoc., 1977, 72(359), p.608-615.
5. Bencze Gy.L., Soroko L.M. JINR, P13-85-502, Dubna, 1985.
6. Tukey J.W. in: Critical Evaluation of Chemical and Physical Structural Information. Nat. Acad. Science, Wash., 1974, p.3-14.
7. Yohai V.J., Maronna R.A. Annals of Statist., 1979, 7(2), p.258-268.

Received by Publishing Department
on April 29, 1986.

Чернов Н.И., Ососков Г.А. E10-86-282

Совместные робастные оценки параметров положения
и масштаба

В работе исследованы робастные оценки параметров в линейных регрессионных моделях с высоким уровнем равномерно распределенного шума. На основе совместного анализа оценок параметров положения и масштаба методом максимального правдоподобия создан численный алгоритм для вычисления регрессионных параметров. Методом Монте - Карло исследованы свойства алгоритма в модели данных с реального физического эксперимента. Алгоритм продемонстрировал высокую эффективность в случае, когда отношение сигнал/шум не менее 1/10.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна 1986

Chernov N.I., Ososkov G.A. E10-86-282

Joint Robust Estimates of Location
and Scale Parameters

Robust estimates of regression parameters are studied in linear models for heavy contaminated distribution of errors with uniformly distributed noise. Maximum likelihood approach to joint estimating location and scale parameters leads to an algorithm for computation of regression parameters. This algorithm was tested by the Monte - Carlo method in experimental data models of a particle track detector. Its high efficiency was demonstrated for signal to noise ratio greater than 1/10.

The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR.

Preprint of the Joint Institute for Nuclear Research. Dubna 1986