zpráva ÚTIA ČSAV č.1463 (1987) 13 7

On robust estimators worth to be applied to real data.

By P. Kovanic and J. Novovičová
Czechoslovak Academy of Sciences

Results of a well-known study comparing classical and robust location estimators applied to real data have been re-evaluated. A significant superiority of robust M- and L- estimators and of an adaptive estimator over the classical ones /sample mean and sample median/ has been shown by a comparison of performances based on three criteria: The set of numerical values of estimates has been obtained by application of twelve estimating methods to 24 data samples originated in classical physical experiments from 18-th and 19-th century. There have been eleven statistical estimators and one entirely new /s.c. gnostical/ method under tests. The results obtained by the latter algorithm have been shown to be the best ones at least when applied to data samples under consideration.

1. Introduction

The performances of a choice of eleven statistical estimators of location parameters were compared by S.M. Stigler in [1] when applied to 24 samples of real data. This comparison included sample mean, median, trimmed means /10%, 15% and 25%/ and outmean, three types of M-estimators /Huber P15, Andrews AMT and Tukey Biweight/, an old estimator of L-type /Edgeworth/ and an adaptive estimator /Hogg T1/. The data were taken from classical physical measurements: determinations of the parallax of the sun /Short 1763/, measurements of the mean density of the Earth /Cavendish 1798/ and measurements of the speed of light /Newcomb 1882, Michelson 1879 and 1882/. Only a part /16/ of the data samples were independent. The size of the samples was 17-100. The evaluation of the estimates in [1] were based on the current knowledge of the "true" values of the physical quantities under consideretion: "The closer the realized value of an estimator to the current "true" value of the estimated quantity, the better the evaluation of the estimator". Such a point of view resulted in a conclusion of [1], that "modern estimators are not worth the time necessary to compute them", "the smallest nonzero trimming percentage included in the study emerged as the recommended estimator" and "the mean itself did rather well". These conclusions were claimed in [1] as only tentative because of small number of data sets employed and of the narrow fields they were selected from. An extension of the classical testing data using a collection of modern analytical--chemistry data were subjected to a similar comparative study in [2]. The main point of evaluation of the quality of the estimators in [2] was that the current "true" value of the classical data is irrelevant to the task of summarizing the measurements because of the possibility of a bias which may be larger then the variations among data: "What is of importance is the variance of the location estimator used since lower variance means that the population location parameter is more precisely determined". The comparison of the variance of the estimators applied to both classical physical data and modern analytical-chemistry data resulted in the suggestion that either severely trimmed means or modern robust estimators are required for optimal performance. There are two main objectives of present paper:

- To employ not only the variance of estimators for evaluation of their efficiency in application to the classical real data but two other measures: the range of estimators /characterizing the ability of the estimator to prevent the worse/ and the mean estimating error based on a "collective" estimation of the actual "true" value of the classical data.
- To compare the statistical estimators tested in [1] with a new estimating method based on the gnostical theory [3]-[5] of real data.

2. A new robust estimator of location

The aim of robust statistics is to develop estimating methods applicable successfully even in the case of behavior of data deviating from the assumptions about the model. Insted of a unique model a set or class of statistical models is used to be considered as a basis for development of robust methods. But practical situations may give rise to doubts of a foundamental nature: is a statistical description of the real process actually adequate? An example: Short's data used in [1] constitute 8 series of sizes 18, 17, 18, 21, 21, 21, 21, 21, altogether 158 data. They form a cluster of 157 data ranged from 6.96 to 10.80.

Only one value /namely 5.76/ appeares to be outside this interval. The mean of all data is 8.61 and the standard deviation 0.67. For the normal distribution the probability of the appearance of the value less or equal to 5.76 would be of the order 10^{-5}. It is an outlier. But what we can say on the statistical model of such an unique real event in terms of distributions? To avoid the necessity of using ~~some~~ any statistical models /"to let data speak for themselves"/ an alternative approach has been developed based on the gnostical theory [3]-[5].

2.1. Main points of the gnostical theory of uncertain data

There are two important problems in modelling the uncertainty of data:

1/ How the amount of uncertainty of an individual datum is to be evaluated

2/ How the data of a sample should be composed to suppress influence of their individual uncertainties on a characteristic of the whole sample.

The former problem is of geometric nature, it is closely connected with the metric. Uncertainty causes an error. To evaluate the uncertainty we measure the error. But how the distance between the true and actual values can be measured? Within the framework of Riemannian geometry the distance is an integral of differentials of the path weighted by a positive weight generally depending on the point coordinates. This weighting function -metric tensor- is a constant only for simplest /uncurved/ spaces- the Euclidean and pseudoeuclidean space. In robust statistics different weights are given to data in dependence on their distance from the true value. It may be also interpreted as using a Riemannian metric. There exists a lot of robust methods differing by their "influence" functions /by metric of the particular Riemannian space/ but what is the proper choice? Have we actually such a free choice? As shown in [3], the metric for measuring the uncertainty of data can be derived rigorously under a simple and plausible assumption on an objective nature of data. If so then such a metric should be prefered before another ones motivated heuristically and subjectively.

The problem of data composition law deserves also attention. The additive composition of statistical moments may be felt as an inheritance from Newtonian mechanics. In gnostical theory the data composition law is subjected to the requirement of consistency with the current physics. Neither linear nor quadratic functions of data but more complicated functions of data are to be composed in the general case. Only in a special case of very weak errors of data the gnostical formulae go over [2] to their statistical equivalents.

It is worth mention that robustness appeares to be an inherent and natural feature of gnostical estimating algorithms and not an "extra" obtained only under some additional assumptions, as a robust super-structure over the nonrobust basic theory.

2.2. Gnostical estimator of location

2.2.1. Problem

Given a sample of ordered data $x_1, \ldots, x_k$ having the model

$$x_i = x_o + s\Omega_i \qquad (x_o \in R_1, \Omega_i \in R_1, s \in R_+) \quad (x_i \geq x_{i-1}) \tag{2.1}$$

where x_o is a location parameter, s is a scale parameter of the sample and Ω_i characterizes an uncertainty of the particular datum x_i. It is required to estimate both parameters x_o and s.

2.2.2. Algorithm

Step I - exponentialization:

$$z_i = \exp((2x_i - x_1 - x_k)/(x_k - x_1)) \qquad (i=1, \ldots, k) \tag{2.2}$$

Step 2 - estimation of the scale parametr:
The estimate $\tilde{s}$ of the scale parametr is obtained as

$$\tilde{s} = \arg\min_s \max_j \max\{|p_{cj} - F(z_j - 0)|, |p_{cj} - F(z_j)|\}$$

where $F(z_j)$ is the value of empirical distribution function

$$j = 1, \ldots, k$$

$$p_{cj} = (1 + h_{cj})/2 \qquad h_{cj} = w_j^{-1} \sum_i^k h_{ij}$$

$$h_{ij} = (q_{ij}^{-2} - q_{ij}^2)/(q_{ij}^{-2} + q_{ij}^2)$$

$$w_j = \left(\left(\sum_i^k f_{ij} \right)^2 + \left(\sum_i^k h_{ij} \right)^2 \right)^{1/2}$$

$$f_{ij} = 2/(q_{ij}^{-2} + q_{ij}^{2})$$

$$q_{ij} = (z_i/z_j)^{1/S}$$

Step 3 - estimation of location parameter z_o of the sample $z_1, \ldots, z_k$:
The variable z_j in /2.2/ is viewed as an independent variable to solve the equation

$$\tilde{z}_o = \arg \max_{z_j} \left\{ \frac{dp_{cj}}{dz_j} \right\}$$

Step 4 - back transformation of the result:

$$\tilde{x}_o = (\log(\tilde{z}_o)(x_k - x_1) + x_1 + x_k)/2$$

2.2.3. Comments

All tested estimators are equivariant with respect to translation. Admissibility of translation is in a correspondence with the model (2.1) of additively distributed data having values $x_i \in R_1$. Data z_i obtained by Step 1 are thus positive and multiplicatively disturbed as required by Axiom 1 of the gnostical theory. The equivariance of location parameter of the additive data will be thus warranted also for the gnostical estimator.

The quantity h_{ij} is called the irrelevance in gnostical theory, it characterizes quantitatively the error resulting from the substitution of z_i instead of z_j. This error is measured in the unique metric derived from the data model. The quantity h_{cj} is the total irrelevancy of all data with respect to z_j. The quantity p_{cj} represents the gnostical estimate of the distribution function at the point z_j. The estimate of the scale parameter is thus sought as the quantity minimizing the maximal absolute difference between the gnostical and empirical distribution functions.

The gnostical estimate of location parameter is determined as the location of a maximum of the data density function which is an estimate of density of probability. This problem has a solution always but there may be more than one solution, say $\tilde{z}_{01}, \ldots, \tilde{z}_{om}$. Then such solution $\tilde{z}_o'$ is to be choosen for which

$$\left(\frac{dp_{cj}}{dz_j}\right)_{z_j=\tilde{Z}_0'} \geq \left(\frac{dp_{cj}}{dz_j}\right)_{z_j=\tilde{Z}_{0k}} \quad (\forall k, k=1,\dots,m)$$

The other maxima of the density characterize locations of some individual outliers or even outlying clusters. An example of this is the second series of Short's measurements where 15 of 17 data form the "main" cluster spread on interval from 7.71 to 9.71 with the maximal density at 8.42. The other values 5.76 and 9.87 appeared to be outliers having "their own" maxima of density located at their measured values. It may be of interest that this was the unique case with a multimodal density from all 24 samples under consideration.

Two remarks are in order here: A/ All formulae given above were derived rigorously from two gnostical axioms, the first one being the multiplicative equivalent of the data model (2.2) and the second one representing the gnostical composition law.

B/ This estimator is reciprocal-equivariant. Thus applied e.g. to Michelson's and Newcomb's data it will give the same estimates of location for velocities as for passage times.

3. Results and the methods of their evaluation

3.1. Standardization of results

The standardization of results of estimating different physical quantities used in [1] will be also accepted here:

Let $\hat{\Theta}_{ij}$ be the value of ith estimator for the jth data set and Θ_j the current "true" value for the jth data set. Let n be the number of data sets and m the number of estimators. Then the quantity

$$s_j = \frac{1}{m}\sum_{i=1}^{m} |\hat{\Theta}_{ij} - \Theta_j|$$

may be used to introduce a new variable

$$e_{ij} = |\hat{\Theta}_{ij} - \Theta_j| / s_j$$

having interesting features:

1/ The set of e_{ij} represents the results of the whole test.

2/ The quantity

$$RE(i) = \frac{1}{n}\sum_{j=1}^{n} e_{ij}$$

/the "index of relative error"/ characterizes the set of results obtained by ith estimator. The mean of RE(i) over all estimators equalling to the mean of e_{ij} over all its realizations equals to 1.

3/ Their distribution is close to normal $N(1, 0.13)$. All values of e_{ij} given in Table 9 of /1/ are summarized by the frequency distribution shown in Fig. 1 for $m = 11$ and $n = 24$.

3.2. Three criteria for evaluation of estimators.

The transformation $\{\hat{\theta}_{ij}\} \rightarrow \{e_{ij}\}$ enabled to unify measurements of different physical quantities in such a way as they would be m series of measurements of a single quantity. A good method applied to these m series of measurements of the same quantity should give estimates spread on a small interval. Thus the first measure of performance of the ith estimator should be

$$SE(i) = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n} (RE(i) - e_{ij})^2}$$

which has been also given in Table 1 in [1] but which seems not to have been used for final conclusions on the performance of individual estimators.

In discussion [1] P. Huber expressed his view that the main purpose of robust procedures is "to prevent the worst", i.e. to prevent a catastrophe due to an occasional bad sample. We shall apply this criterion not in the way as [1] depending critically on the knowledge of the "true" value of the samples but by means of the range

$$r(i) = \max_j(e_{ij}) - \min_j(e_{ij})$$

characterizing the distance between the couple of worst cases and not depending on a "true" value.

But we cannot avoid the discussion on what should be accepted as the "true" value of the unified variable e_{ij} because the mean estimating error must be taken into account. If not then a trivial estimate e_{ij} = constant would be the best one because of its zero variance and zero range. Stigler's and Huber's point of view in [1] is that the "true" values for the old measurements are the same as for today and that a good estimator should find these current true values in the old data inspite of undoubtful /see discussants Eisenhart, Hoaglin, Pratt in [1]/ bias of data. Hence, their criterion is

$$(e_{ij} < e_{kj}) \Longrightarrow \text{the ith estimator is better than the kth one for jth sample}$$

because they accept zero to be the "true" value for jth sample of standardized data set $\{e_{ij}\}$. But we suggest instead to use the unit to be the "true" value of this set and to evaluate the estimators in the following way:

$$|e_{ij} - 1| < |e_{kj} - 1| \Longrightarrow \text{the ith estimator is better than the kth one for jth sample}$$

Our reasoning /motivated by the Table of Eisenhart in [1]/ is based on the idea of an "expert board" including m members: Professor Mean, Professor Median, Professor Edgeworth and others. They all were given the same data sets and asked to give their judgements on the "true" values of these data. What they judged is summarized in Table 9 of [1] and in Figure 1 here. Assume that we did our best to collect the most distinguished, experienced and qualified experts who are unbiased /at least in the sense of "unprejudiced"/ and objective /working only with data/. Then we should believe in their collective wisdom. We are glad to see that this expert board worked "normally" - the distribution of their estimates is close to normal. And with the mean equalling to 1. We therefore take the estimate of mean estimating error of ith estimator equalling to $ME(i) = RE(i) - 1$.

4. Comparison of estimators

The gnostical estimator described in 2.2 was applied to the same 24 data sets as the 11 estimators in [1]. The values e_{ij} and s_j

for 11 estimators taken from Table 9 in [1] were used to get corresponding $|\hat{\Theta}_{ij} - \Theta_j|$ /i = 1, ..., 11/ which together with values $|\hat{\Theta}_{12,j} - \Theta_j|$ /where i = 12 denotes the gnostical estimator/ enabled to get new $\{e_{ij}\}$ and $\{s_j\}$ shown in Table 1. Using this Table and the three criteria discussed above we may obtain a comparison of all 12 estimators for various choices of data sets. In Table 2 there are results for 16 small independent sets, all from N^o1 to N^o16 contained in Table 4,5 and 6 in [1]. The rows of our Table 2 are ordered according to the values of SE(i). For other choices of data samples /all 20 small samples N^o1 - N^o20 or all 24 small and large samples N^o1 - N^o24 , where large samples N^o21 - N^o24 are unions of small ones/ the gnostical estimator would hold the first place with the smallest variance and the order of other estimators would change only slightly.

Following statements seem to be supported by our Table 2:

1/ The three suggested criteria yielded roughly the same order of estimators.

2/ The gnostical estimator worked with these data well

3/ The higher percentage of trimming, the better the result of trimmed means. This is in a ful agreement with the results of /2/.

4/ The role of outmean as a "planned outlier" was confirmed.

5/ Both median and mean worked badly with these data.

6/ Mean and 10% trimmed mean gave approximately the same quality of results.

7/ The robust estimators are worth not only the time necessary to compute them. It is worth to get the standard error of estimates about 4 or 6 % as in the gnostical case and in the case of Hogg's estimator then above 20% as in the case of the mean and 10% trimmed mean.

References

[1] Stigler S.M. and discussants: /1977/ Do Robust Estimators Work with Real Data? Annals of Stat., Vol. 5, No. 6, 1055-1098.

[2] Rocke D.M., Downs G.W., Rocke A.J.: /1982/ Are Robust Estimators Really Necessary? Technometrics 24, No. 2, 95-101.

[3] Kovanic P.: /1984/ Gnostical Theory of Individual Data, Problems of Control and Information Theory /PCIT/, Vol. 13, No. 4, 259-274.

[4] Kovanic P.: /1984/ Gnostical Theory of Small Samples of Real Data, PCIT, Vol. 13, No. 5, 304-319.

[5] Kovanic P.: /1984/ On Relation Between Information and Physics, PCIT, Vol. 13, No. 6, 383-399.

TABLE 1

The realized values of the estimates e_{ij} and their average s_j for twelve estimates and twenty-four data sets

Data Set	Mean	Median	Edgeworth	Outmean	10% Trim	15% Trim	25% Trim	Huber P15	Andrews AMT	Tukey Biweight	Hogg T1	Gnostical	s_j
1	.80	1.43	1.08	.46	.96	1.01	1.13	.96	.98	1.02	1.06	1.06	.208
2	1.05	1.10	.93	1.16	.93	.95	.96	.93	.98	.99	1.02	.95	.397
3	.29	1.78	1.09	1.68	.56	.88	1.09	.64	1.01	1.04	1.01	.92	.094
4	.73	.94	1.16	.32	.97	1.06	1.14	1.04	1.07	1.41	1.15	1.04	.320
5	.92	1.76	.91	2.92	.19	.60	1.09	.18	.45	1.04	.92	.99	.078
6	.77	1.07	1.18	.42	.98	1.06	1.11	.99	1.05	1.28	1.09	1.02	.456
7	.99	1.00	.98	1.00	1.01	1.00	.98	1.01	1.01	1.02	.99	.99	.238
8	.96	.98	.95	.92	1.00	1.01	1.00	1.01	1.11	1.04	1.01	1.00	.222
9	1.35	.90	.90	1.82	.90	.89	.87	.90	.83	.77	.88	.94	8.355
10	.98	1.10	1.03	.90	.98	.99	1.06	.97	1.00	1.00	.98	1.04	4.576
11	.97	1.12	1.01	.89	.99	.99	1.04	.99	1.00	.99	1.00	1.03	5.373
12	.94	1.10	1.02	.81	1.00	1.03	1.06	.99	1.00	1.02	1.04	1.04	187.050
13	1.02	.93	.93	1.09	.99	.98	.95	1.01	1.01	1.00	1.09	.98	118.950
14	.92	1.00	1.03	.82	:98	1.03	1.02	1.03	1.09	1.09	1.00	1.01	120.369
15	1.01	.95	.96	1.03	1.00	1.00	.99	1.01	1.01	1.01	1.03	1.00	85.118
16	1.06	.82	1.01	1.14	1.01	1.00	.97	1.01	1.00	.98	1.00	1.00	91.898
17	.94	1.31	1.03	.75	.90	.99	1.13	.99	.87	.89	1.07	1.08	48.350
18	.87	1.47	1.07	.56	1.00	1.11	1.17	.85	.90	.95	.87	1.19	.039
19	1.09	.90	.95	1.24	.93	.94	.94	.98	.99	.93	1.09	.99	.063
20	.93	1.27	1.10	.83	.99	1.08	1.04	.90	.95	.96	.93	1.02	.037
21	.82	1.33	.83	.50	.86	.95	1.13	.94	1.15	1.17	1.24	1.06	.223
22	.71	1.25	.99	.24	.95	1.04	1.19	1.03	1.10	1.28	1.11	1.11	.238
23	1.15	1.01	.96	1.32	.95	.95	.97	.95	.85	.91	.97	.98	5.917
24	1.01	.99	.99	1.04	1.00	1.00	.98	1.00	1.00	1.00	.99	.99	116.984

TABLE 2

Estimating errors for twelve estimators applied to sixteen independent data sets.Estimators are ordered according to the measure of spread SE(i).

ESTIMATOR		ERRORS		
		Mean square error	Mean error	Range
i	Name	SE(i)	ME(i)	$\max_j e_{ij} - \min_j e_{ij}$
12	Gnostical	0.038	-0.001	0.139
11	Hogg T1	0.061	-0.017	0.261
7	25% Trim	0.070	-0.029	0.261
3	Edgeworth	0.079	-0.011	0.273
6	15% Trim	0.104	0.032	0.447
10	Tukey Biweight	0.131	-0.043	0.631
9	Andrews AMT	0.147	0.025	0.660
8	Huber P15	0.210	0.083	0.856
5	10% Trim	0.211	0.097	0.821
1	Mean	0.212	0.078	1.055
2	Median	0.278	-0.124	0.962
4	Outmean	0.610	-0.086	2.603

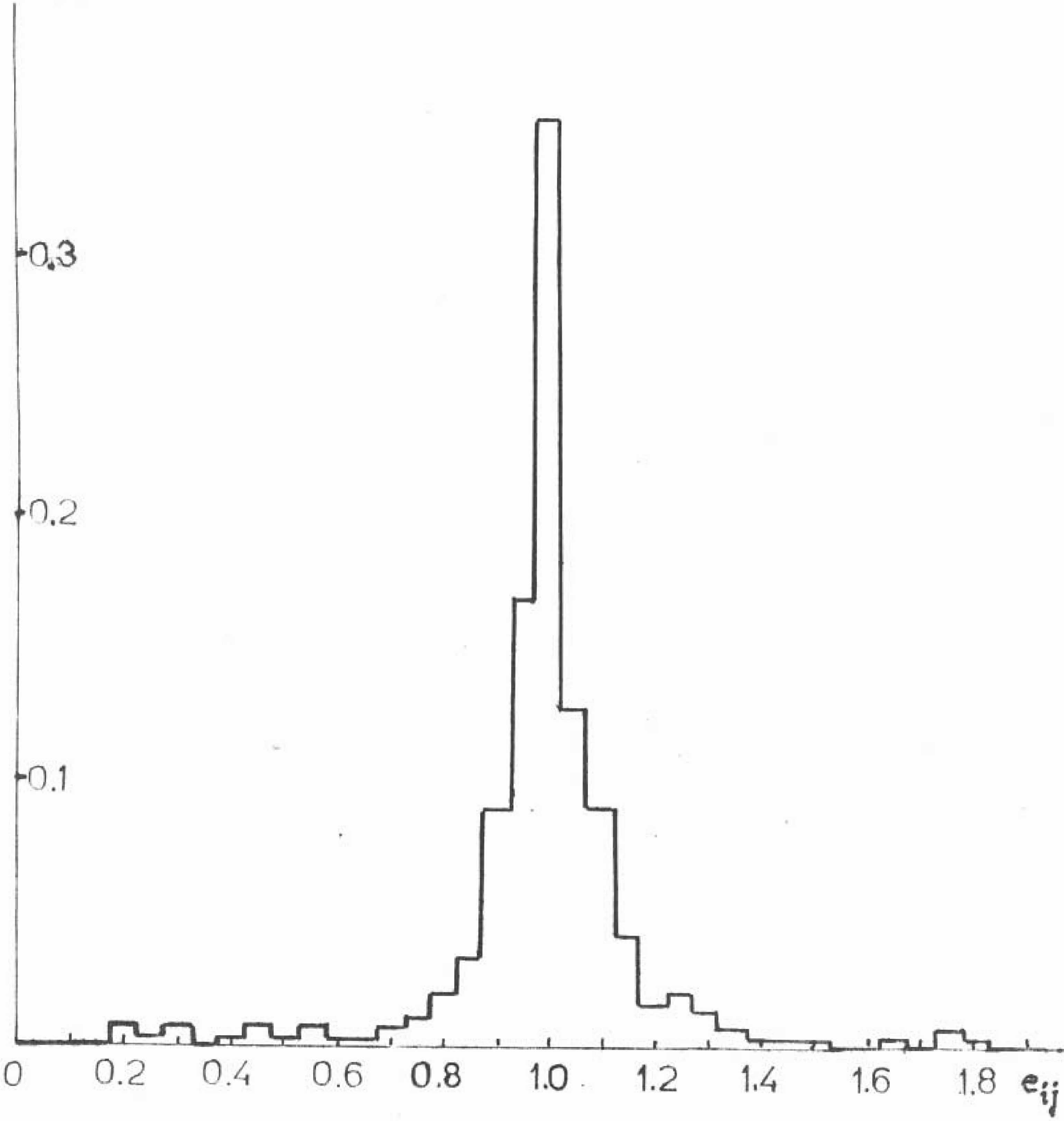

FIG.1