GEOPHYSICAL TRANSACTIONS 1987
Vol. 33. No. 1. pp. 3–9

POSSIBILITIES TO REALIZE HIGHER EFFICIENCY IN GEOPHYSICAL INTERPRETATION

Ferenc STEINER*

If the algorithm of interpretation of geophysical data systems contains, in an explicit or implicit manner, statistical tools, it is deemed advisable to apply such statistical procedures (estimations, fitting techniques, etc.) which have great efficiency for a large range of probability distribution types. The present paper deals, with this special point of view and, in a concise manner, with the statistical procedures based on the generalized most frequent value.

Keywords: probability, statistical distribution, robust estimation, Danish method, least-squares analysis

1. Introduction

Algorithms of geophysical interpretation often contain (in an explicit or implicit manner) statistical components, too. This short paper deals with cases when this constituent part is some sort of fitting.

If there are K equations among the components of the parameter vector $\bar{p} = p_1, p_2, \ldots, p_j, \ldots, p_J$ to be determined, we have to fulfil exactly

$$A_k(\bar{p}) = 0 \qquad (k = 1, \ldots, K)$$

with given analytical expressions $A_k(\ldots)$. (We write $K=0$ if there are no equations to be fulfilled.)

If $\bar{y}_i$ denotes the exactly known vector-variable, its components are, in the case of a fitting with m variables,

$$y_{i1}, y_{i2}, \ldots, y_{im}.$$

The z_i results of the measurements consist of the "exact value" $T(\bar{p}; \bar{y}_i)$ ($T(\ldots)$ is an a priori known analytical expression) and the "error" x_i ($i = 1, \ldots, n$ if n is the number of measured data). It depends on the type of probability distribution of errors x_i what sort of fitting will give optimal interpretation from the statistical point of view.

* Department of Geophysics, Technical University for Heavy Industry, Miskolc, Egyetemváros, H-3515, Hungary

Manuscript received: 26 November, 1986

2. Fitting by most frequent value

For the sake of simplicity let us suppose a symmetrical distribution of errors $f(x)$ with a parameter of scale $S=1$. The symmetry point lies, of course, at zero, $f(x)$ being the probability density function of errors.

Let us suppose that careful investigation of a great number of large samples justifies that the density function has the analytical form:

$$f_G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{1}$$

i.e. the errors x_i have a Gaussian distribution. We have in this case to fulfil the demand

$$\sum_{k=1}^{K} \lambda_k A_k(\bar{p}) + \sum_{i=1}^{n} x_i^2 = \text{minimum} \tag{2}$$

to get the vector of parameters $\bar{p}$ as exact as possible, supposing that the appearance of outliers can be absolutely excluded. (λ_k denotes for $k=1, \ldots, K$ the well known Lagrange multipliers.) However, if the investigation of the type of probability distribution of errors x_i results in a density function

$$f_a(x) = N \cdot (1+x^2)^{-\frac{a}{2}} \tag{3}$$

for some a value $(1<a<\infty)$ as an adequate model for the actual distribution (the explicit formula for the norming factor N is given in [Csernyák and Steiner 1982]), then the solution of the condition

$$\exp\left[\sum_{k=1}^{K} \lambda_k A_k(\bar{p})\right] \cdot \prod_{i=1}^{n} (1+x_i^2) = \text{minimum} \tag{4}$$

gives the optimal $\bar{p}$ (i.e. with minimal asymptotic variance; for $K=0$ [see Steiner 1985]).

The above mentioned statements are condensed in *Fig. 1.*

In most cases of fitting the following form of $T(\bar{p}; \bar{y})$ is accepted:

$$T(\bar{p}; y) = \sum_{j=1}^{J} p_j \cdot T_j(\bar{y}). \tag{5}$$

For simplicity $K=0$ but from now on S is allowed to be arbitrary. After logarithmization of Eq. 4 and derivation according to p_j we shall have formally the same equation system as by fulfilling the least squares condition with a priori weights; the weights, however, must now be calculated according to

$$\varphi(x_i) = \frac{(k\varepsilon)^2}{(k\varepsilon)^2 + x_i^2} \tag{6}$$

where ε denotes the dihesion of the data system x_i [see e.g. Steiner 1985], the

$$\bar{p} = p_1, p_2, \ldots, p_j, \ldots, p_J; \qquad A_k(\bar{p}) = 0 \qquad (k = 1, \ldots, K)$$
$$\bar{y}_i = y_{i1}, y_{i2}, \ldots, y_{im}$$
$$T(\bar{p}, \bar{y})$$
$$z_i = T(\bar{p}, y_i) + x_i \qquad (i = 1, \ldots, n)$$

$$\begin{tikzcd}
 & (S=1) \arrow[dl] \arrow[dr] & \\
f_a(x) = N \cdot (1+x^2)^{-\frac{a}{2}} & & f_G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}
\end{tikzcd}$$

$$\exp\left[\sum_{k=1}^{K} \lambda_k A_k(\bar{p})\right] \cdot \prod_{i=1}^{n} (1 + x_i^2) = \min. \qquad \sum_{k=1}^{K} \lambda_k A_k(\bar{p}) + \sum_{i=1}^{n} x_i^2 = \min.$$

Fig. 1. Schema of optimal fitting techniques for different types of probability distributions of the error *x*

1. ábra. Optimális kiegyenlítési eljárások különböző típusú valószínűségi hibaeloszlásra

Рис. 1. Оптимальные способы выравнивания при различных типах вероятностного распределения ошибок.

value of k corresponds to the actual value of a, with $f_a(x)$ being the density function of the best, i.e. of the most adequate, model. (The function $k(a)$ can be determined, of course, once for all, see [STEINER 1985]. Some values are: $k(2) = 1$; $k(4.4) = 1.9$; $k(6.2) = 2.4$ and $k(8) = 2.8$.) The value x_i, however, as the difference between the measured and the calculated value, depends also upon $\bar{p}$ and, consequently, the solution of the equation system is to be carried out as an iteration—but every step of the iteration can be calculated by means of the standard program for weighted least squares fitting. An additional program gives in a few iteration steps the approximation of the dihesion ε—and we have defined completely the algorithm we call "fitting according to the generalized most frequent value".

3. Examples

This new type of fitting is to be compared, on the one hand, with the conventional least squares fitting, and on the other hand with other robust and resistant methods. For clarity and in order to consider at the same time a high variety of distribution types we make the comparisons on the ground of the $f_a(x)$ distribution family defined in Eq. 3 (for $a > 1.7$). It is well known that $a = 2$ corresponds to the Cauchy distribution, $a = \infty$ to the Gaussian distribution. If the actual distributions are expected between these two types (and nothing more is known about the types), then the most appropriate choice is $k = 1.9$. If more information is known about the most probable a-range, the middle a-value of this shorter interval will guarantee for us an efficiency of 100 per cent or a value very near to this optimum.

The oldest robust procedure minimizes the sum of the absolute differences; in the case of T=const. this corresponds to the calculation of the sample median. I have chosen from the more modern robust procedures, for the sake comparison, the so-called "Danish method" [KRARUP and KUBIK 1983] because Hungarian surveyors seem to show very much interest in this procedure [DETREKŐI 1986]. Although there are some different variants of this method, the most often used weight function by the Danish method is:

$$\varphi(x_i) = \begin{cases} 1, & \text{if} \quad x_i < b \\ \exp(1-(x_i/b)^2), & \text{if} \quad x_i \geq b, \end{cases} \tag{7}$$

where

$$b = \frac{c \cdot \text{med}\,|x_i|}{0.6745}.$$

The most direct way to test the economy of the method is to calculate the efficiency e. Namely, the reciprocal of this value gives a very important ratio: how many times more data are necessary to result in the same reliability using an arbitrary statistical procedure, compared with the optimal one. The efficiency is to be calculated as the quotient of the two asymptotic variances belonging to both procedures in question:

$$e = \frac{A^2_{\text{opt}}}{A^2}. \tag{8}$$

If $S=1$ holds

$$A^2 = \frac{\int\limits_{-\infty}^{\infty} \psi^2(x) f(x)\,\mathrm{d}x}{\left[\int\limits_{-\infty}^{\infty} \psi'(x) f(x)\,\mathrm{d}x\right]^2} \tag{9}$$

(see e.g. [STEINER 1985]) with $\psi(x) = x \cdot \varphi(x)$, and one can easily verify [HAJAGOS 1985] for the distribution family $f_a(x)$ defined by Eq. 3 that

$$A^2_{\text{opt}} = \frac{a+2}{a \cdot (a-1)}. \tag{10}$$

For the sample median we can get the asymptotic scatter simply as the reciprocal of $2f_a(0)$.

Figure 2 shows the curves of efficiencies versus $1/(a-1)$, calculated as discussed above. The Danish method gives two curves for different c-values: according to the literature $c=3$ is the most frequently used value of this parameter; a practical example, however, is shown by KRARUP and KUBIK [1983] to be $c=1.5$, and therefore the efficiency curve was constructed using this c-value, too.

As the asymptotic variance of arithmetical means exists only for $a>3$ and

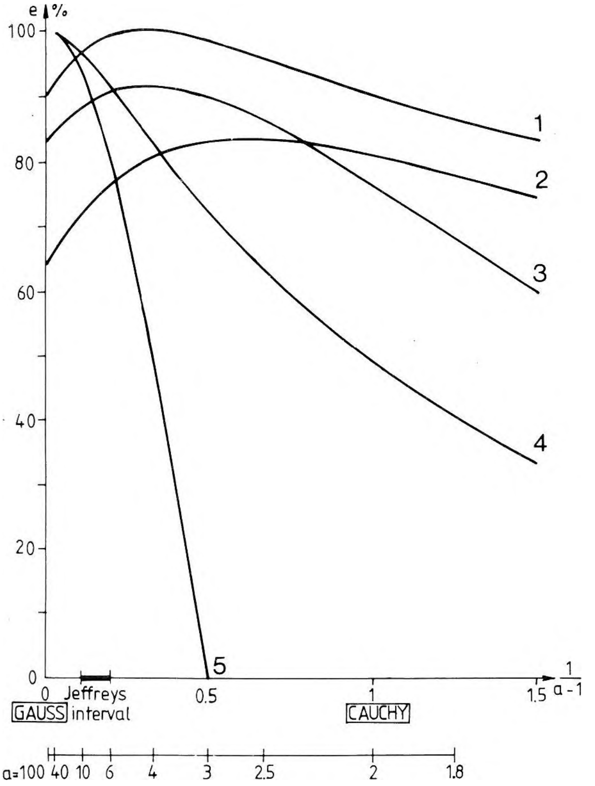

Fig. 2. Efficiences of some statistical procedures for a large range of probability distribution types

e — efficiency (%), 1/(a – 1) — type parameter of the probability distribution of errors; 1 — fitting according to the generalized most frequent value, k = 1.9; 2 — minimization of absolute differences; 3 — Danish method, c = 1.5; 4 — Danish method, c = 3; 5 — least squares fitting

2. ábra. Néhány statisztikai módszer hatékonysága a valószínűségi eloszlások széles skálájára e — hatásfok (%), 1/(a – 1) — az eltérések (hibák) eloszlásának típusparamétere; 1 — általános leggyakoribb érték szerinti kiegyenlítés, k = 1,9; 2 — abszolút eltérések minimalizálása (medián); 3 — dán módszer, c = 1,5; 4 — dán módszer, c = 3; 5 — legkisebb négyzetek módszere

Рис. 2. Эффективность некоторых статистических методов в широком диапазоне вероятностных распределений

е — эффективность (%); 1/(а – 1) — типовой параметр распределения отклонений (ошибок); 1 — выравнивание по найболее частому генеральному значению, k = 1,9; 2 — сведение абсолютных отклонений к минимуму (медиана); 3 — датский способ, с = 1,5; 4 — датский способ, с = 3; 5 — способ наименьших квадратов.

its value is in this case $1/(a-3)$ [Csernyák–Steiner 1982], the efficiency curve starts at 100% (versus $1/(a-1)$)but rapidly decreases to zero. Compared with this behaviour, an enormous gain is offered by the Danish method applying $c=3$. If $c=1.5$ is used instead of $c=3$ the effectiveness increases considerably in the neighbourhood of the Cauchy distribution (from about 50%, to about 80%), the effectiveness in the case of Gaussian distribution, however, decreases to 83% (but this value is still considerably greater than the efficiency value of about 64% of the sample median). The efficiency of the generalized most frequent value is greater than 90% in the whole range from the Gauss type to the Cauchy type (i.e. for $a>2$), and therefore the economic advantages are obvious. If previous investigations concerning the type of actual probability distributions result in a value of $1/(a-1)$ whether near to zero or near to unity, we have only to choose the suitable value of k, i.e. according to the function $k(a)$, to reach an effectiveness very near to 100 per cent (as mentioned yet above). Both cases may occur in practice, not only in geophysics but also, for example, in astronomy: the data system of Short [1963] refers to Cauchy distribution (determined with the know-how belonging to the University of Miskolc); on the other hand, the copybook example of the least squares monograph of Linnik (see Table 6 in [Linnik 1961]) can really be regarded as nearly Gaussian (the adequate value of a is indeed great) if we can disregard a small loss—say, one or two per cent—in the efficiency.

Finally, the economically very important role of the efficiencies should be emphasized. For instance, if we use a procedure with an efficiency of only 50 per cent, it means nothing less than our having thrown out half of our expensively measured data.

Acknowledgements

The author is indebted to J. Somogyi (director of the Geodetical and Geophysical Research Institute Ac. Sci. Hung.), and to professors L. Csernyák (College of Finance and Accountancy, Budapest) and B. Hajagos (Higher Institute of Digital Control for Chemical Engineering, Kazincbarcika) for their valuable help and discussions.

REFERENCES

CSERNYÁK L., STEINER F. 1982: Untersuchungen über das Erfüllungstempo des Gesetzes der großen Zahlen. Publ. of the Techn. Univ. for Heavy Ind., Miskolc Series A, Mining, **37,** 1–2

DETREKŐI Á. 1986: Consideration of robust errors in the processing of survey data (in Hungarian). Geodézia és Kartográfia **38,** 3, pp. 155–160

HAJAGOS B. 1985: Die verallgemeinerten Studentschen t-Verteilungen und die häufigsten Werte. Publ. of the Techn. Univ. for Heavy Ind., Miskolc Series A, Mining, **40,** 1–4, pp. 225–238

KRARUP T., KUBIK K. 1983: The Danish method; experience and philosophy. Seminar Math. models of geodetic (photogrammetric) point determination with regard to outliers and systematic errors (ed. by F. E. Ackermann). Deutsche Geodätische Kommission Reihe A, Nr. 98, München

LINNIK J. W. 1961: Die Methode der kleinsten Quadrate in moderner Darstellung. Deutscher Verlag der Wissenschaften, Berlin, 31 p.

SHORT J. 1763: Second paper concering the parallax of the sun etc. Philos. Trans. Roy. Soc., London **53,** pp. 300–343

STEINER F. 1985: Robust estimations (in Hungarian). Tankönyvkiadó, Budapest, 172 p.

HATÁSFOKNÖVELÉSI KÉRDÉSEK A GEOFIZIKAI ÉRTELMEZÉSBEN

STEINER Ferenc

A terepmérések által szolgáltatott egyre nagyobb geofizikai adatrendszerek kötelességünkké teszik, hogy minél hatásosabban nyerjük ki az azokban levő információkat. Ha értelmezési algoritmusunk (explicit vagy implicit módon) matematikai statisztikai elemet is tartalmaz, akkor olyan becslési módszert célszerű alkalmazni, amely eloszlástípusok széles spektrumára nagy hatásfokú. A tanulmány az általánosított leggyakoribb értékeket mutatja be ebből a szempontból.

ВОПРОСЫ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ГЕОФИЗИЧЕСКОЙ ИНТЕРПРЕТАЦИИ

Ференц ШТЕЙНЕР

Все увеличивающийся объем геофизических данных, поступающих вследствие полевых измерений, заставляет искать пути повышения эффективности извлечения информации, содержащейся в этих данных. Если альгоритм, используемый в интерпретации, содержит элементы математической статистики (в явной или неявной части уравнения), то целесообразно применение такого способа оценки, который обладает высокой эффективностью в широком диапазоне типов распределений. В статье представлены наиболее часто применяемые значения, важные для обсуждаемой проблемы.