XV Congress of the International Society for
Photogrammetry and Remote Sensing
Rio de Janeiro 1984

Commission III

PRESENTED PAPER

AH, ROBUST ESTIMATION!

PER CRÜGER JØRGENSEN

POUL FREDERIKSEN

KURT KUBIK

WILLY WENG

Abstract

This paper gives a review of the different methods of robust estimation and of their suitability for photogrammetric computation, together with an evaluation of their advantages and disadvantages compared to the least squares method.

Per Crüger Jørgensen and Kurt Kubik
Aalborg University, Aalborg, Denmark

Poul Frederiksen and Willy Weng
Technical University of Denmark, Lyngby, Denmark

1. INTRODUCTION

The method of least squares has been used for estimation of parameters from measurements for nearly two centuries. In general, the method has served the purpose well but not without any problems. It is recognized, that the least squares estimate is heavily influenced by outliers among the measurements, which pull the least squares estimate too much towards them, causing a difficult errors-search due to the misleading residuals. This often results in several trials with the method of least squares omitting different measurements (eventually by statistical tests) as outliers; a process which is time consuming and sometimes impossible, when dealing with a huge amount of data.

The method of least squares is based on the assumption that the errors of the measurements are distributed according to the Gaussian error curve and the so-called outliers may just arise from another - more heavily tailed - distribution or are simply due to blunders. To cope with this problem applied statisticians reject observations which seem located too far away from the data bulk before estimating the mean from the remaining observations, but it was not till the beginning of the fifties statisticians (Box, Turkey, and later Huber, Hampel etc.) began to examine the problem, to emphasize the shortcoming of the classical estimate and to establish alternatives - the more robust estimates. A robust estimate may be defined as an estimate which remains unchanged under a variation of the underlying distribution functions or in the presence of observations from contaminating distributions. The sensitivity of the robust procedure to changes in the underlying distribution function is considered as a qualitative aspect of the procedure whereas the quantitative aspect is related to the amount of observations from a contaminating distribution or the amount of outliers among the observations which are still tolerable in estimation.

2. MAXIMUM - LIKELIHOOD ESTIMATORS.

If $Z_1, Z_2 \cdots Z_n$ is a sample for the random variable z with density $f(z-\Theta_0)$ where Θ_0 is a location parameter, the logarithm of the likelihood function is

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(Z_i-\theta) = -\sum_{i=1}^{n} \varphi(Z_i-\theta) \qquad (1)$$

with Θ denoting the estimator for Θ_0.

In maximum likelihood estimation or M-estimation the goal is to maximize $\ln L(\Theta)$ corresponding to minimizing the object function $\Sigma\, \varphi(Z_i-\Theta)$, which can be solved by differentiating the object function with respect to Θ

$$\frac{\partial}{\partial\Theta} (\Sigma\, \varphi(Z_i-\Theta)) = \Sigma\, \omega(Z_i-\Theta) = 0 \quad \text{or} \quad \frac{\partial\varphi}{\partial\theta} = \omega \qquad (2)$$

As seen from eq. 1 the object function must be convex in order to obtain a unique solution.

In order to be able to judge the influence of an additional observation on the estimator obtained by a given method, Hampel (1974) has developed a very efficient tool, namely the influence curve (IC). For M-estimators this influence curve is simply proportional to ω.

The IC enables us to construct a variety of robust estimators satisfying pre-determined conditions to the estimate, eg qualitative robustness, rejection point and gross error sensitivity. Rather than specifying the φ function, M-estimators are usually described by specifying the function ω. The qualitative robustness means that a "small change" in the set of observations should result in a small change in the estimate. This small change may either be many small changes in the observations (rounding off or grouping effects) or a few large changes (outliers).

This means that the IC should be continuous and bounded. The first property is necessary and causes that a small change in many observations induces only a small change in the estimate.

The maximum value of the IC is a measure of the gross error sensitivity. This value measures the worst influence of contaminating observations. Therefore a bound on the IC is the most important step in robustifying an estimator. However, the lower the bound the larger the smallest variance that can be achieved.
The rejection point is the value of the error where the IC is descending towards zero influence of the observation.

3. ROBUST M-ESTIMATORS.

The first alternative to the least squares method, the least sum method was proposed already in 1793 by Laplace and 1887 reintroduced by Edgeworth.

The minimum function is given by

$$\Sigma\ |Z-\theta| \rightarrow \min \qquad (3)$$

giving a ω function

$$\Sigma\ \omega(|Z-\theta|) = \Sigma\ \mathrm{sign}(Z-\theta) = 0 \qquad (4)$$

yielding the sample median as the estimate of location. The median estimator can also be categorized as an L-estimator (linear combinations of ordered statistics) since the median is obtained by taking the "mean" after having deleted or trimmed 50% of the data from each end of an ordered dataset. This 50% trimmed mean (median) is the one extreme of the family of L-estimators indexed by α, the proportion of the sample size removed from each end.

The distribution function producing this median estimator is the double exponential distribution

$$f(Z) = c_1 \cdot \exp\left(-\frac{|z-\theta_0|}{c_2}\right); \qquad c_i \text{ constants}$$

which is longer tailed than the normal distribution and therefore permits more outliers. The method never found large application due to computational difficulties since a solution of a linear programming problem is required. However, the recently developed simplex algoritm provides a solution of the least sum principle without requiring more time than least squares. Furthermore the solution can be obtained easily by a reweighted least squares algorithm.

Due to the computational advances the method is again taken up for treatment and especially Barrodale (1973) recommends the method after several tests showing much better performance of the least sum method than the least squares method when outliers are present in the measurements. See section 7 for examples.

The least squares and least sum estimators are also named norm estimators L_β where

$$\Sigma\ |Z-\theta|^{\beta} \rightarrow \min; \qquad \beta \in [1;2] \tag{5}$$

The least sum method yields less accurate (efficient) estimators than the least squares method in cases where the data actually are normally distributed. Therefore Huber (1964) has derived the following robust minimum function

$$\varphi(Z-\theta) = \begin{cases} \dfrac{(Z-\theta)^2}{2} & , \quad |Z-\theta| \leq a \\ a|Z-\theta| - \dfrac{a^2}{2} & , \quad |Z-\theta| > a \end{cases} \tag{6}$$

and ω function

$$\omega(Z-\theta) = \begin{cases} Z-\theta & , \quad |Z-\theta| \leq a \\ a \cdot \operatorname{sign}(Z-\theta) & , \quad |Z-\theta| > a \end{cases} \tag{7}$$

The Huber estimator is based on an error model being normal in the middle and double exponential in the tails.

The turning constant, a, may be changed according to assumptions about the error model. However, if the estimator should have an efficiency of 95% when the error model is purely normal*), a should be set to 1.5 since most of the scaled (about scaling see section 4) residuals would enjoy the property $\frac{|Z-\theta|}{s} \leq 1.5$, s being the standard deviation.

The above-mentioned M-estimators are the so-called non-descending estimators according to the form of the ω function. Common for these estimators is that all observations, even the most extreme outlier do influence the estimate.

In order to be able to reject outliers Hampel (1973) has proposed a three-part descending estimator including as well the robustness properties mentioned as a rejection point. Hampel's proposal is

$$\omega(Z-\theta) = \operatorname{sign}(Z-\theta) \cdot \begin{cases} |Z-\theta| & 0 \leq |Z-\theta| \leq a \\ a & a \leq |Z-\theta| < b \\ \dfrac{a(c-|Z-\theta|)}{c-b} & b \leq |Z-\theta| < c \\ 0 & c \leq |Z-\theta| \end{cases} \tag{8}$$

*) efficiency of 95%: the variance of the estimator is 5% larger than the variance of the least squares estimator.

Since the ω function is descending the associated object function is not convex and there could be certain convergence problems in the solution procedure. The influence functions of these three robust estimators are shown in Figure 1.

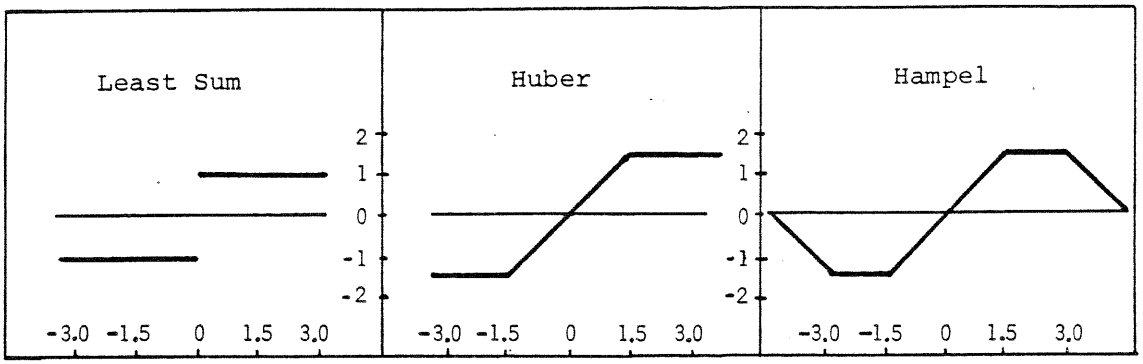

Fig. 1. Influence functions for a = 1.5, b = 3.0 and c = 4.5.

4. SCALE INVARIANCE

The estimator θ based on the observation $Z_1, Z_2, \ldots Z_n$ is scale invariant if

$$\theta(\lambda Z_1, \lambda Z_2 \ldots, \lambda Z_n) = \lambda \cdot \theta(Z_1, Z_2 \ldots Z_n) \tag{9}$$

As it easily can be shown only the norm estimators are scale invariant. Therefore it is necessary to enforce scale invariance by estimating a scale parameter simultaneously with θ.

The usual estimator for the standard deviation is a very non-robust measure of scale, and normally the median of the absolute deviations (the MAD estimate) is used as scaling parameter. The MAD estimate is defined by

$$s = \frac{\text{med}|e_i|}{0.6745} \quad ; \qquad e_i = Z_i - \theta \tag{10}$$

where 0.6745 is the expectation of med $|e_i|$ under strict normal conditions.

5. NUMERICAL SOLUTION

The numerical solution can be obtained by using the method of "reweighted least squares". Eq.(2) can be written as

$$\Sigma \frac{e_i}{s} \cdot p\left(\frac{e_i}{s}\right) = 0 \tag{11}$$

where the weight function p is given by

$$p\left(\frac{e}{s}\right) = \frac{\omega\left(\frac{e}{s}\right)}{\left(\frac{e}{s}\right)} \tag{12}$$

The procedure is iterative, starting with the least squares solution (p=1).

In the second iteration the individual weights are calculated from eq. (12) using the residuals from the least squares solution. In the third and following iterations, the residuals from the previous iteration are used for calculating new weights. This procedure is continued until convergence is obtained. In practice exact convergence is never achieved and a stopping rule is needed.

To achieve scale invariance, a scale has to be estimated simultaneously with the location parameter which may cause certain convergence problems, especially if the least squares solution is heavily influenced by outliers. To avoid part of this problem a more robust solution than the least squares solution can be used as the starting solution in the iteration procedure, eg. the least absolute sum estimator being scale invariant. The object function related to the descending ω-functions is not convex meaning that no unique definition of minimum is possible and the iteration procedure may converge to a local minimum. To cope with this problem it is always a good idea to apply different estimators on the same data for evaluation of the reliability of the result.

6. THE DANISH METHOD

The sensibility of the least squares method to gross errors has also been recognized by the Geodetic Institute of Denmark, where since the late sixties an automatic error search routine has been used in the computation of all larger geodetic problems. This method was developed after the ideas of Krarup (1967) and is used as an automatic mean for singling out gross error candidates for further inspections. It is based on the iterative numerical solution, described in section 5.

The starting point of the method is a conventional least squares adjustment. From the residuals of this first adjustment, new weights are computed for the individual measurements, according to the rule

$$p = \begin{cases} .1 & \text{for } |Z-\theta| < a \cdot s \\ \exp\left(-\left(\frac{|Z-\theta|}{a \cdot s}\right)\right) & \text{else} \end{cases}$$

The constant a usually is set to 3. The convergence grade of the method is usually superior to other robust methods; it depends on the stability (conditioning) of the adjustment problem. For photogrammetric bundle adjustment a variant of the original Danish Method proved to be most efficient,

$$p = \begin{cases} \exp\left(-0.05\left(\frac{|Z-\theta|}{s}\right)^{4.4}\right) & \text{for first 3 iterations} \\ \exp\left(-0.005\left(\frac{|Z-\theta|}{s}\right)^{3.0}\right) & \text{for following iterations} \end{cases}$$

Other exponential weight functions are used for other categories of problems (cf. Crüger and Kubik 1983). A possible further improvement of the Danish Method is to use the variance of the residuals instead of s^2 in computing the weight function. This technique is being implemented at the Geodetic Institute of Denmark.

The Danish Method cannot be classified within the existing maximum likelihood theory of robust estimation; an adjoint error distribution function cannot be derived due to the rapid decrease of the weight function. Another possible interpretation of the Danish Method is given by non-linear programming:

Find the largest number of measurements, which is mutually consistent, and use only these measurements in the least squares adjustment to determine the unknowns. An alternative formulation is to find those observations which are not consistent with the majority and exclude them from adjustment.(cf. Krarup et al, 1980). The iterative method may be interpreted as a penalty method for solution of the above programming problem. In this formulation the problem resembles the clustering problems in statistics.

7. NUMERICAL EXAMPLES

This section presents results from application of three alternatives to the least squares method for estimation of parameters in a first and third order polynomial. These examples are representative for strip- and block adjustments with erroneous ground control points.

The method tested are the least absolute sum, Huber's method and the Danish Method. The estimates are all obtained by the method of reweighted least squares as sketched in section 7.2.

The following weight functions are used in the computations, with $e = (Z-\theta)$:

Absolute Sum:

$$p(e) = \frac{1}{|e|}$$

Huber's Method:

$$p(\frac{e}{s}) = \begin{cases} 1 & \text{for} \quad |\frac{e}{s}| \leq a \; ; \; a = 1.5 \\ \frac{a}{|\frac{e}{s}|} & \text{else} \end{cases}$$

Danish Method:

$$p(\frac{e}{s}) = \begin{cases} 1 & \text{for} \quad |\frac{e}{s}| \leq a \; ; \; a = 1.5 \\ \exp(1-(\frac{e}{a \cdot s})^2) & \text{else} \end{cases}$$

7.1. Linear Regression

Let us first consider the fitting of a straight line to 20 data points, including 5 outliers (cf. figure 2). In the first example (case 2.1) the outliers are concentrated in the center of the interval, next the outliers are situated on both ends (case 2.2), and finally all outliers are concentrated at one end of the interval (case 2.3). The regression lines, as estimated by the different methods, are shown in the figures.

All methods enable the identification of the erroneous observations for the first case (2.1). The robust estimates are obviously better than the least squares estimates. The Danish Method gives the best result, because here the wrong observations have the lowest weights. For the second case (2.2) only the Danish Method gives a clear identification and elimination of the gross errors. The least sum method behaves second best, while Huber´s method and the least squares method cannot locate the blunders. For case (2.3) none of the methods originally gives satisfactory results. By reducing the constant a from 1.5 to 0.4, the Danish Method rejects the gross errors, but also 4 correct observations (point 9, 12, 13 and 16).

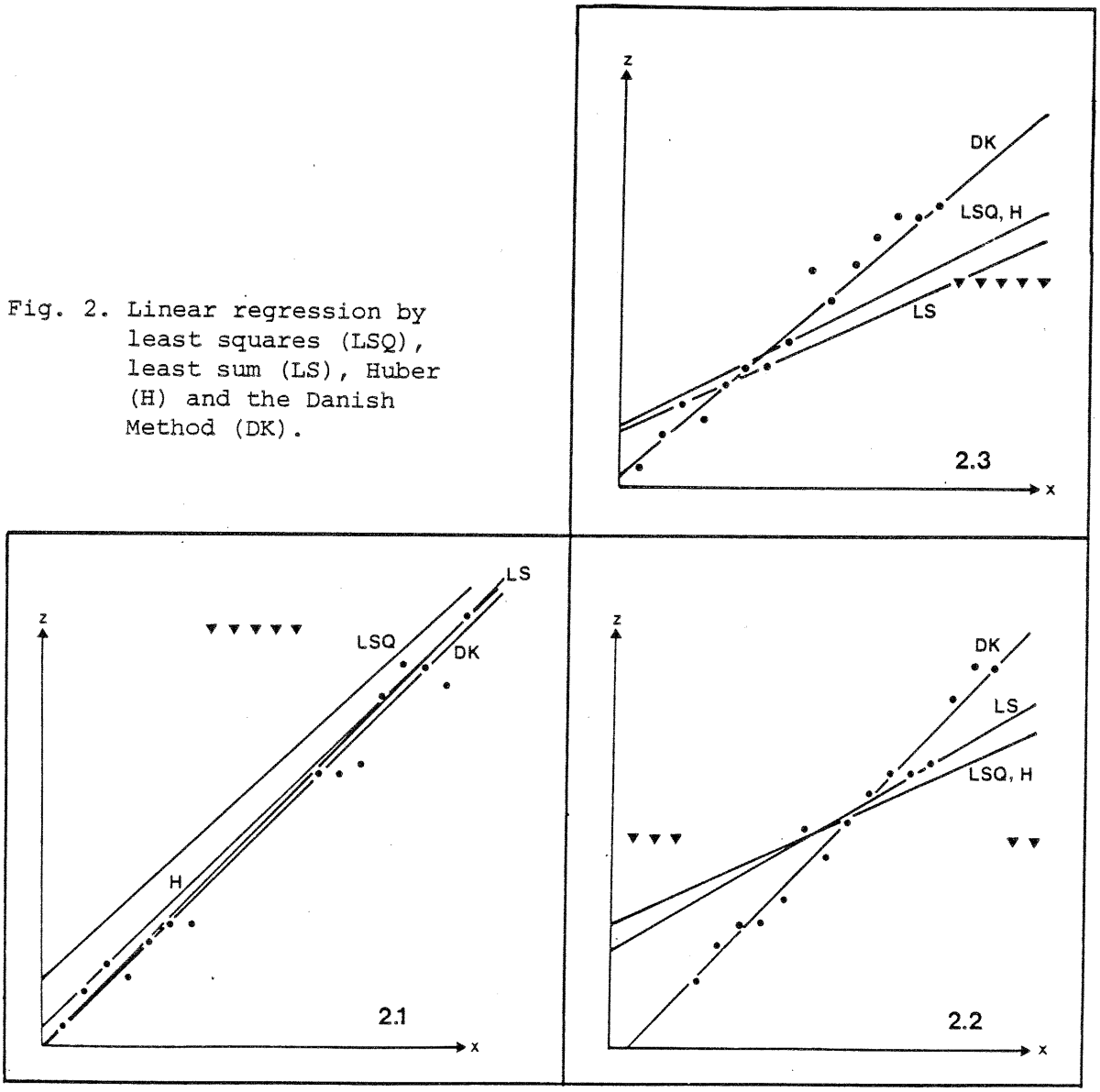

Fig. 2. Linear regression by least squares (LSQ), least sum (LS), Huber (H) and the Danish Method (DK).

7.2. Third order polynomial fitting

In the second example, a third order polynomial is fitted to 10 data points, including one outlier. This case is in its character rather characteristic for photogrammetric and geodetic adjustments, where the degrees of freedom are appr. equal to the number of unknowns.

The 3rd order polynomial considered for the numerical example is

$$Z_i = 21 \cdot X_i - 10 \cdot X_i^2 + X_i^3 + e_i \quad ; \; i = 1,.....,10$$
$$X_i = 0.....,9$$

The errors e_i have been generated with a variance of 1 unit, a gross error of 20 units has been introduced in the first data point. The least sum method and the Danish Method is compared to least squares. In the Danish Method, the median estimate s for the standard deviation is applied in the weight function, when it is smaller than the a priori value σ = 1 unit; otherwise the a priori value is used.

Table 1 and figure 3 show the results of the computations. The Least Sum Method requires 26 iterations to converge, the Danish Method 6. The Danish Method detects and eliminates the gross error in the first data point, while the other two methods are not succesful. For more test results cf. Crüger and Kubik (1983).

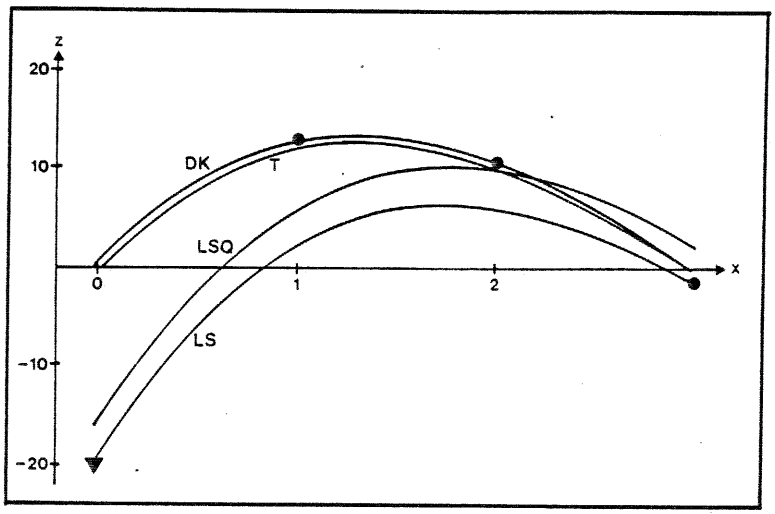

Fig. 3. 3^{o} Polynomial fitting - three estimates and true function (T). Only the first parts of the curves are plotted.

point	1	2	3	4	5	6	7	8	9	10
true value θ	0	12	10	0	-12	-20	-18	0	40	108
true error e	-20	0.6	1.0	-0.7	0	-0.2	-1.0	-0.9	1.5	-0.9
estimated errors										
least squares	-3.8	6.5	1.2	-2.9	-2.1	-0.9	0	1.2	3.2	-2.4
least sum	-0.4	10.5	4.8	0	-0.3	-0.1	0	1.1	3.7	-0.2
Danish Method	-20.1	0	0.2	-1.2	-0.2	0.2	-0.3	0.2	2.7	0

Table 1. 3^{o} Polynomial fitting - residuals after adjustment.

8. FINAL REMARKS

The application of robust estimation methods is still in its beginning. Much work still has to be done, both of theoretical and practical nature. However, one thing is quite clear: The method of least squares should be used with care.

9. REFERENCES

Barrodale, I. and F.D.K.Roberts, 1973. An improved Algorithm for Descrete L1 Linear Approximation. SIAM J. Numer. Anal. Vol. 10, No.5.

Crûger Jørgensen, P. and K.Kubik, 1983. Elimination of Gross and Systematic Errors in Geodetic Adjustment Problems. Aalborg University Center, 1983, 150 pp.

Edgeworth, F.Y., 1887. A New Method for Reducing Observations Relating to Several Quantities. Phil. Mag. (5th Series), 24, pp 222-224.

Hampel, F.R., 1973. Robust estimation: A condensed Partial Survey z. Wahrscheinlichkeitstheorie verw. Geb. 27.

Hampel, F.R., 1974. The Influence Curve and its Role in Robust Estimation. J. of the American Stat. Ass.

Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Statist. 35 (1964) pp 73-101.

Krarup, T., 1967. Personal Communications.

Krarup, T., K.Kubik, J.Juhl, 1980. Götterdammerung over Least Squares Adjustment. Proc. 14th Congress of ISP, Hamburg, 1980.

Laplace, P.S., 1793. Sur quelques points du systeme du monde. Memoires de l'Academie royale des Sciences de Paris, 1789, 1-87 (Reprinted in Oevres Completes de Laplace, Vol. 11, pp 477-558. Gautier-Villars, Paris 1895).