A new method of identifying and locating gross errors
—Quasi-accurate detection

OU Jikun
Institute of Geodesy and Geophysics, Laboratory of Dynamical Geodesy, Chinese Academy of Sciences, Wuhan 430077, China

Abstract A new idea and a distinctive method have been proposed, which concern real errors and their estimators. By using the idea of "Quasi-Stable Adjustment" created by Prof. Zhou Jiangwen for reference, the rank-deficient equations on real errors are resolved by adding the conditions under which the minimum of the norm of the real errors of the quasi-accurate observations is restrained.

The formulae and the scheme of the new method called "quasi-accurate detection of gross errors (QUAD)" are presented. By using this method, not only multiple gross errors can be identified and located exactly with these estimators calculated accurately, but also the precision of the estimators is able to be evaluated strictly. This method may be suitable to dealing with the gross errors existing in various fields of science and engineering.

Keywords: least squares (LS), detection of gross errors (or outliers), hypothesis test, robust estimation, real errors.

In general, there exist gross errors inevitably in the data of surveying or experiments of various fields. If the influence of the gross errors is not freed from, the final results would be biased, even distorted, which would cause waste of labor power, materials and financial resources.

Besides strictly following the standard rules to obtain data, in the past there are two methods normally used to cope with the gross errors in many post-procedures. One is hypothesis test[3], for example, a famous scheme called "Data Snoopig"[2] proposed by Prof. Baarda used widely for about thirty years in surveying and geodesy. The other is robust estimation. Mathematicians have perfected lots of originating work in the past thirty years, such as the theories presented by Drs. Huber[4], Hampel[5] and Rousseeu et al.[6], which laid a foundation of robust estimation. In the field of geodesy, Drs. Kubik[7], Caspary[8], Koch[9] and Zhou et al.[10] have made studies deeply, specially in application.

Although the hypothesis test is used widely in many fields, its effectiveness is not as good as to be expected when there exist multiple outliers in data, particularly in the case that the strength of a surveying network (or a system) is quite weak. Robust M-estimation has been developed for many years and become a very useful method for getting rid of the influence of outliers in data. Its key is how to design a weight function suitable for practical problem. Nevertheless, it is difficult to structure a universal weight function now, and there is no precise method suited to evaluate the precision of estimators under the multiple dimension.

In the present note, a new idea and a distinctive method different from those used previously are proposed. The real errors of observations are taken as the studied objects, and a reasonable scheme of estimating real errors is presented. By using this method, the gross errors are able to be identified and located; furthermore! the values and the precision of them can be calculated and assessed.

1 Mathematical model

The linearized observation equation can be written as

$$AX_0 = L + \Delta,$$

where $A$ is the coefficient matrix of the order $n \times m$, its rank being $m$; $X_0$ is an $m$-dimensional vector of unknown parameters; $L$ is an $n$-dimensional observation vector, and $\Delta$ is a vector of the real errors corresponding to $L$. Note that only equal weight and independent observations are discussed in this note.

Let $J = A(A^TA)^{-1}A^T$ be called projective matrix, or matrix of adjustment factors[10]. $J$ is idempotent (i.e. $J^2 = J$) and $JA = A$. Denote $R = I - J$, where $I$ is an identity matrix of the order $n \times n$; $R$ is a matrix of orthogonal complements of $J$, also idempotent, and there is $A^T R = 0$, and the rank of $R$ is $n - m$. Since $JAX_0 = AX_0 = L + \Delta = J(L + \Delta)$, we get $(I - J)\Delta = -(I - J)L$, or

$$R\Delta = -RL.$$  (1.2)
This is a determinate and analytic representation relationship between real errors $\Delta$ and observation values $L$. It can be taken as the linear equations on $\Delta$. However, eq. (1.2) is rank-deficient. In the light of mathematics, it is not difficult to resolve this kind of equations. But it should be emphasized that the physical significance of the solutions has to be clear and reasonable.

2 Quasi-accurate detection of gross errors

The gross errors are practically considered as the large ones far from the majority of observations or locally deviating from models.

The analysis of numerous data shows that the quantity of observations with outliers is usually less than 1%−10% of the total\textsuperscript{[4]}\textsuperscript{[4]}. Therefore, it can be believed sufficiently that most of the data is normal. The parts of data which are considered essentially normal but still required to verify are referred to as "quasi-accurate observations (QAO)", corresponding to the values of the real errors which are relatively less than those of the others.

By using the idea of "Quasi-Stable Adjustment" created by Prof. Zhou Jiangwen for reference, by selecting the $r$ appropriate observations as QAO, $r > m$, the determinate solutions of eq. (1.2) can be achieved under the restrained condition—the minimum of the norm of real errors of QAO.

It can be also proved that this process is equivalent to resolving the jointed equations which are written as

$$\begin{cases} R\Delta = - RL + \epsilon, \\ G_0\Delta = 0, \end{cases} \quad (2.1)$$

where $\epsilon = R\Delta + RL$ is the vector of fitting residuals, $\Delta$ is partitioned as $[\Delta_1, \Delta_2]$, in which $\Delta$, represents the real errors of the QAO, and $\Delta_1$ represents those of the others. The matrix $G_0$ of the order $m \times n$ can be taken as

$$G_0 = (0 \quad A^T), \quad (2.2)$$

where 0 is the zero matrix, $A^T$ is an $m \times r$ transposed matrix of $A$, corresponding to the QAO.

Assume that the row vectors of $R$ and $G_0$ are not correlated mutually. By means of the technique we get the conditional extremum, and the solution of eq. (2.1) can be derived as

$$\hat{\Delta}_0 = -(R + G_0^T G_0)^{-1} RL, \quad (2.3)$$

where subscript Q denotes the "Quasi-Accurate" solution.

The cofactor matrix (or matrix of weight inverse) of $\hat{\Delta}_0$ is given by

$$Q_{\Delta_0} = (R + G_0^T G_0)^{-1} R (R + G_0^T G_0)^{-1}. \quad (2.4)$$

It can also be proved that the estimator of $\Delta$ is the pseudo-inverse solution of eq. (1.2) if all observations are taken as QAO, i.e. $r = n$. This estimator is signified as $\hat{\Delta}^+$, of which the numerical values are equal to those of the residuals of the classical least-squares (LS), i.e. $V = AX_1S - L$, where $\hat{X}_1S$ is the LS estimator of parameters. This means $V = \hat{\Delta}^+$. It should be noticed that the result is obtained under the condition $||\Delta||^2 = \min$ or $||V||^2 = \min$. This is an unreasonable restrained condition when some observations are contaminated by outliers. In that case the estimators of the parameters would be distorted so that the residuals $V$ are not refered to as the studied objects in the present note.

If there exist gross errors, the estimators of real errors achieved by the above technique as the suitably selecting QAO would appear as a prominent characteristic that their values can be divided into two groups; the ones of the part $\hat{\Delta}$, are obviously less than those of the part $\hat{\Delta}_1$. Thus it is able to provide a reliable basis for identifying and locating the outliers. According to this, if $\hat{\Delta}_i$ is distinctly larger than the given criterion, the $i$th observation can be judged as contaminated by an outlier.

Suppose that the $b$ gross errors have been found, if existing, the $b$ unit vectors of $n$-dimension can be structured, such as $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T (j = 1, \ldots, b)$, in which the element corresponding to the $j$th outlier is 1, the others 0. Denoting the outliers as a vector $\nabla_b$ of attached parameters, eq. (1.1) is rewritten as

$$AX_0 = L + \Delta = L - C_b \nabla_b + N, \quad (2.5)$$

where $C_b$ is a coefficient matrix of the order $n \times b$, $C_b = (e_1 \cdots e_b)$, and $N$ is a vector of real errors sep-