نشریه علمی - پژوهشی علوم و فنون نقشه برداری، دوره پنجم، شماره ۲، آبان ماه ۱۳۹۴

Application of Recursive Least Squares to Efficient Blunder Detection in Linear Models

A. R. Amiri-Simkooei [*1], H. Ansari [2], M. A. Sharifi [3]

[1]Dept. of Geomatics Engineering, Faculty of Engineering, University of Isfahan, Isfahan, Iran
[*]amiri@eng.ui.ac.ir

[2]Remote Sensing Technology Institute (IMF), Earth Observation Center (EOC),
German Aerospace Center (DLR), Wessling, Germany
Technical University of Munich (TUM), Munich, Germany
homa.ansari@dlr.de

[3]School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran
Tehran, Iran
sharifi@ut.ac.ir

(Received: September 2014, Accepted: November 2015)

Key Words: Outlier detection, Baarda data-snooping method, DIA algorithm, w-test statistic, Hypothesis testing, Geodetic network analysis.

Abstract

In many geodetic applications a large number of observations are being measured to estimate the unknown parameters. The unbiasedness property of the estimated parameters is only ensured if there is no bias (e.g. systematic effect) or falsifying observations, which are also known as outliers. One of the most important steps towards obtaining a coherent analysis for the parameter estimation is the detection and elimination of outliers, which may appear to be inconsistent with the remainder of the observations or the model. Outlier detection is thus a primary step in many geodetic applications. There are various methods in handling the outlying observations among which a sequential data snooping procedure, known as Detection, Identification and Adaptation (DIA) algorithm, is employed in the present contribution. An efficient data snooping procedure is based on the Baarda's theory in which blunders are detected element-wise and the model is adopted in an iterative manner. This method may become computationally expensive when there exists a large number of blunders in the observations. An attempt is made to optimize this commonly used method for outlier detection. The optimization is performed to improve the computational time and complexity of the conventional method. An equivalent formulation is thus presented in order to simplify the elimination of outliers from an estimation set-up in a linear model. The method becomes more efficient when there is a large number of model parameters involved in the inversion. In the conventional method this leads to a large normal matrix to be inverted in a consecutive manner. Based on the recursive least squares method, the normal matrix inversion is avoided in the presented algorithm. The accuracy and performance of the proposed formulation is validated based on the results of two real data sets. The application of this formulation has no numerical impact on the final result and it is identical to the conventional outlier elimination. The method is also tested in a simulation case to investigate the accuracy of the outlier detection method in critical cases when large amount of the data is contaminated. In the application considered, it is shown that the proposed algorithm is faster than the conventional method by at least a factor of 3. The method becomes faster when the number of observations and parameters increases.

[*] Corresponding author

1. Introduction

In many data mining tasks in general and geodetic applications in particular a large number of observations are being measured to retrieve information about an underlying physical process. One of the most important steps towards obtaining a coherent analysis of the physical process is the detection of falsifying observations, also known as outliers. Outliers are abnormal observables that appear to be inconsistent with the remainder of the dataset (Johnson, 1992). This inconsistency may be the result of systematic and/or gross errors. The inclusion of such data in information retrieval procedures may adversely lead to model misspecification and/or biased parameter estimation. Therefore consideration of a process which bounds the impact of outliers in the estimation is a necessity in any information retrieval procedure. There are various methods in handling the outliers. These methods are categorized in two broad groups as

- The statistical tests based on least-squares (LS) method, in which the conventional least-squares estimation is performed followed by the post-processing statistical tests on the outcome. The tests aim at revealing the presence of outliers and identifying them among the observations (Baarda 1968; Pope 1976; Pelzer 1983; Teunissen 2000).
- Robust estimation methods in which the impact of the falsifying observations is bound by the estimation process rendering the final estimation results unaffected by the outliers (Huber 1981; Hampel 2001; Rouseeuw and Leroy 2003).

For the LS estimates, a single gross error is adjusted in a way that it affects all residuals, and hence it contaminates the estimated parameters. In contrast to the LS, in robust estimators, the gross error will only affect the corresponding observation residual (not all the others), leaving the estimation result unaltered by the blunder. Thus, robust estimators are less sensitive to blunders in the data. A number of robust methods have been proposed, which include the M-, R-, and L-estimators, the least median of squares (LMS) and the sign-constrained robust least squares (Huber 1981; Hampel et al. 1986; Rousseeuw and Leroy 2003; Koch 1999). One of the commonly used robust estimation method is the L1 norm minimization method. It has been used for outlier detection in geodetic observations by the studies of Marshall and Bethel (1996); Amiri-Simkooei (2003); Khodabandeh and Amiri-Simkooei (2011). Other robust methods have also been applied to geodetic networks. Examples include studies by Fuchs (1982); Gao et al. (1992); Youcai (1995); Awange and Aduol (1999); Wicki (1999); Berber and Hekimoglu (2001); Marshall (2002).

Despite their superior performance in dealing with outliers, robust methods do not provide minimum variance and/or maximum likelihood estimators as the LS method does. Furthermore, the computational cost and complexity of the robust estimators is higher compared to the LS methods. These disadvantages prevent the robust estimators to substitute the LS, specifically in geodetic applications. Therefore, improvement of the statistical outlier detection methods in provision of unbiased estimation in the LS framework is a contribution to geodetic applications.

The LS based statistical method is further divided into two subcategories: the single step and the sequential procedure. The former performs a simple statistical test and detects all observations failing the test as outliers. The latter performs the statistical test in a step-by-step manner by elimination of the extreme observation at each step, adapting the estimation results and iterating the statistical test until acceptance. The sequential procedure is more reliable for outlier identification in comparison with the single-step methods (Ben-Gal, 2005; Davies and Gather, 1993). Schwarz and Kok (1993) show that sequential statistical method implemented based on Baarda's *data-snooping* algorithm (Baarda 1968) is capable of correct detection of multiple outliers. It is additionally shown that this approach may have comparable results to the robust estimators such as L1 norm minimization, or may even outperform them, if implemented by Baarda method and in a sequential manner.

Although an effective method, the sequential data-snooping is computationally expensive. The computational burden arises from repetition of the estimation with the filtered dataset after identification of each outlier. The main purpose of the current work is to propose an algebraic formulation to eliminate the outlier's impact form the estimation result, i.e. to update the available estimated parameters instead of repeating the estimation with a filtered dataset. The proposed method is conceptually similar to the sequential adjustment expressed in Mikhail and Ackerman (1976). On a second level, the aim is to validate the proposed method; i.e. to assure that the employed formulation does not impose any numerical inaccuracies in the results and lead to exact the same result as the conventional approach. Additionally, the performance of the sequential data-snooping in correct identification of outliers is investigated. The investigations include cases in the presence of large number of outliers for which the LS-based outlier detection methods are known to be less effective.

This paper is organized as follows. Initially, the sequential data-snooping method is described in

details, followed by the through explanation of the algebraic formulation which optimizes this method in Section 3. The fourth section provides three cases to investigate the proposed methodology. Finally, we provide some conclusions in Section 5.

2. Sequential data-snooping

The sequential data-snooping, also known as Detection, Identification and Adaptation (DIA) algorithm, is performed in following three steps (Teunissen 2000):

- **Detection**: A statistical test on the overall validity of the assumed model given the observations. Rejection of the test indicates that the data at hand does not support the model. In this case, one alternative hypothesis is the presence of outliers in the data. Thus if the overall validity is rejected, the following two steps are investigated to identify outliers and adopt the model.
- **Identification**: Specification of the error which causes the rejection of the overall validity test. Each of the observations are treated as a potential outlier and statistically tested in an attempt to locate the falsifying observation.
- **Adaptation**: Compensating the identified error in order to reach an unbiased estimation. In this case the identified outlier must be eliminated from the data set and the estimation process must be repeated without the falsifying observation.

In the sequential approach, only one outlier is identified and accounted for at each iteration. The DIA is then repeated until the overall validity test is accepted or the framework runs out of the *redundancy*. Baarda's data-snooping method defines the efficient statistical test for identification of the outliers which is of common practice in geodetic applications. This approach is further elaborated in the following.

The DIA is essentially a hypothesis testing process. There are two hypothesis tests involved in this algorithm; the *overall model test* in the *detection* step as well as the *w-test* in the *identification* step. Introducing a proper test statistic is the first step toward the statistical testing. To introduce the test statistics consider the linear functional model as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is the m-vector of observations, $\boldsymbol{A}$ is the $m \times n$ design matrix, and $\boldsymbol{x}$ is the n-vector of parameters. The least-squares estimate of the parameters is obtained by

$$\hat{\boldsymbol{x}} = \boldsymbol{N}^{-1}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{Q}_{\boldsymbol{y}}^{-1}\boldsymbol{y}, \tag{2}$$

where $\boldsymbol{N} = \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{Q}_{\boldsymbol{y}}^{-1}\boldsymbol{A}$ is the normal matrix, $\boldsymbol{Q}_{\boldsymbol{y}}$ is the $m \times m$ covariance matrix of the observations. The estimated residuals follow from

$$\hat{\boldsymbol{e}} = \boldsymbol{P}_{\boldsymbol{A}}^{\perp}\boldsymbol{y} \tag{3}$$

where

$$\boldsymbol{P}_{\boldsymbol{A}}^{\perp} = \boldsymbol{I} - \boldsymbol{A}\,\boldsymbol{N}^{-1}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{Q}_{\boldsymbol{y}}^{-1} \tag{4}$$

is an orthogonal projector. The overall model test statistic is defined based on the estimated residuals and reads as (Teunissen, 2000)

$$T = \hat{\boldsymbol{e}}^{\boldsymbol{T}}\boldsymbol{Q}_{\boldsymbol{y}}^{-1}\hat{\boldsymbol{e}} \tag{5}$$

The above test statistic follows the Chi-square distribution with $m - n$ degrees of freedom. Thus the null and alternative hypotheses are formulated as:

$$H_0: T \sim \boldsymbol{\chi}^{\boldsymbol{2}}(m-n, 0) \text{ vs. } H_a: T \sim \boldsymbol{\chi}^{\boldsymbol{2}}(m-n, \lambda)$$

This statistical test is in fact a general check on discrepancies between the observed data and the assumed model. The rejection of the null hypothesis indicates the inadequacy of the assumed model or the presence of outliers among the observations. Therefore, in case the overall model test fails one possible approach is to identify and eliminate the outliers which is in the scope of the two further steps of the DIA algorithm.

When the overall test is rejected, the identification step starts. In this step the w-test statistic is proposed based on the estimated LS residuals and its covariance matrix. The covariance matrix of the estimated residuals reads is

$$\boldsymbol{Q}_{\hat{\boldsymbol{e}}} = \boldsymbol{Q}_{\boldsymbol{y}} - \boldsymbol{A}(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{Q}_{\boldsymbol{y}}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\boldsymbol{T}} = \boldsymbol{P}_{\boldsymbol{A}}^{\perp}\boldsymbol{Q}_{\boldsymbol{y}} \tag{6}$$

Finally, the residuals are compared to their standard deviations and form the w-test statistic for each single observation (Teunissen, 2000)

$$w_i = \frac{\boldsymbol{c}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{Q}_{\boldsymbol{y}}^{-1}\hat{\boldsymbol{e}}}{\sqrt{\boldsymbol{c}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{Q}_{\boldsymbol{y}}^{-1}\boldsymbol{Q}_{\hat{\boldsymbol{e}}}\boldsymbol{Q}_{\boldsymbol{y}}^{-1}\boldsymbol{c}_{\boldsymbol{i}}}} \tag{7}$$

where here $\boldsymbol{c}_{\boldsymbol{i}}$ is a *canonical unit vector*, with one single non-zero element in the i^{th} vector position as $\boldsymbol{c}_{\boldsymbol{i}} = [0 \cdots 0\ 1\ 0 \cdots 0]^T$. In a simplified form, if the covariance matrix $\boldsymbol{Q}_{\boldsymbol{y}}$ of the observations is diagonal, the expression for the w-test statistic reduces to the simple form as

$$w_i = \frac{\hat{\boldsymbol{e}}_i}{\sqrt{(\boldsymbol{Q}_{\hat{\boldsymbol{e}}})_{ii}}} = \frac{\hat{\boldsymbol{e}}_i}{\sigma_{\hat{e}_i}} \tag{8}$$

which indicates that this test statistic can be interpreted as the normalized estimated residuals.

In the absence of outliers, the observations as well as the w-test statistic follow the standard normal distribution. Thus the outlier detection is formulated with the following null and the alternative hypothesis (Teunissen et al., 2005):

$$H_0\colon\ w \sim \boldsymbol{N}(\,0\,,1)\ \text{ vs. }\ H_a\colon\ w \sim \boldsymbol{N}(\,\nabla w\,,1) \qquad (9)$$

assuming that the covariance matrix of the observables is known; in geodetic literature this is referred to as σ known. Therefore, the null and alternative hypotheses consider the absence and presence of an outlier, respectively, and ∇w accounts for the anomalies, disturbances or large errors in the observations. The w statistics that are rejected under the null hypothesis can in fact have potential outliers. In the sequential method, only the largest w is considered at each step and is statistically tested. If it is rejected, the corresponding observation is identified to have an outlier.

After the identification step, the *adaptation* step follows and the contribution of the outlier is eliminated from the solution. The estimation process is to be repeated without the outlier. For clarification of the adaptation step, consider a set of m observations. Without loss of generality, assume that the m^{th} observation is identified as an outlier in the first step of the procedure. This observation must be removed from all the vectors and matrices involved, and the LS must be repeated with the filtered set-up. If y_m is the observation # m, σ_m^2 is its variance, and $\boldsymbol{a_m^T}$ is the row vector in the design matrix $\boldsymbol{A}$ related to the outlier, one has

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A_{m-1}} \\ \boldsymbol{a_m^T} \end{bmatrix}, \boldsymbol{Q}_y = \begin{bmatrix} \boldsymbol{Q}_{y_{m-1}} & 0 \\ 0 & \sigma_m^2 \end{bmatrix}, \boldsymbol{y}=\begin{bmatrix} \boldsymbol{y_{m-1}} \\ y_m \end{bmatrix} \qquad (10)$$

Removing the effect of the y_m from the least-squares adjustment yields the new estimates as

$$\hat{\boldsymbol{x}}_{\boldsymbol{m-1}} = \left(\boldsymbol{A_{m-1}^T}\,\boldsymbol{Q_{y_{m-1}}^{-1}}\boldsymbol{A_{m-1}}\right)^{-1}\boldsymbol{A_{m-1}^T}\boldsymbol{Q_{y_{m-1}}^{-1}}\boldsymbol{y_{m-1}} \qquad (11)$$

and

$$\begin{cases} \hat{\boldsymbol{e}}_{\boldsymbol{m-1}} = \boldsymbol{y_{m-1}} - \boldsymbol{A_{m-1}}\hat{\boldsymbol{x}}_{\boldsymbol{m-1}} \\ \boldsymbol{Q}_{\hat{e}_{m-1}} = \boldsymbol{Q}_{y_{m-1}} - \boldsymbol{A_{m-1}}(\boldsymbol{A_{m-1}^T}\boldsymbol{Q_{y_{m-1}}^{-1}}\boldsymbol{A_{m-1}})^{-1}\boldsymbol{A_{m-1}^T} \end{cases} \qquad (12)$$

where $\hat{\boldsymbol{x}}_{\boldsymbol{m-1}}$ is the vector of unknown parameters which is derived after omission of the outlier, $\hat{\boldsymbol{e}}_{\boldsymbol{m-1}}$ is the corresponding residual vector and $\boldsymbol{Q}_{\hat{e}_{m-1}}$ is its covariance matrix. The new residual vector is the input for the next iteration of the DIA algorithm.

The algorithm is iterated until the overall model test is accepted or the set up runs out of *redundancy* (Teunissen, 2000). This approach can be computationally expensive because the estimation process is to be repeated after identification of each single outlier. In the following section an algebraic strategy is employed to deal with this problem and hence optimize the application of the sequential data-snooping algorithm.

3. Sequential data-snooping

Prior to introduction of the strategy we employed in this contribution, the applied algebraic concepts are explained in the following subsection. These concepts are then applied to optimize the adaptation step of the DIA algorithm in a subsequent section.

3.1 Mathematical background

Consider the invertible $n \times n$ matrix $\boldsymbol{N}$ and three arbitrary matrices $\boldsymbol{U}$, $\boldsymbol{B}$ and $\boldsymbol{V}$ of appropriate size of $n \times p$, $p \times p, p \times n$, respectively One can then prove that the matrix $\boldsymbol{N} + \boldsymbol{UBV}$ can be inverted as follows (Henderson and Searle, 1981):

$$(\boldsymbol{N} + \boldsymbol{UBV})^{-1} = \boldsymbol{N^{-1}} - \boldsymbol{N^{-1}}\,\boldsymbol{U}\,(\boldsymbol{I} + \boldsymbol{BVN^{-1}U})^{-1}\,\boldsymbol{BVN^{-1}} \qquad (13)$$

When $p = 1$, the preceding equation can be simplified. In this case, the matrices $\boldsymbol{B}$ and $\boldsymbol{I}$ become scalars β and 1, respectively, $\boldsymbol{U} = \boldsymbol{u}$ becomes a column vector of size n, $\boldsymbol{V} = \boldsymbol{v^T}$ becomes a row vector of size n. Therefore, one has $(\boldsymbol{I} + \boldsymbol{BVN^{-1}U})^{-1} = \boldsymbol{1}/(1 + \beta\boldsymbol{v^T N^{-1}u})$ (a scalar), which with Eq. (13) yields

$$(\boldsymbol{N} + \beta\boldsymbol{uv^T})^{-1} = \boldsymbol{N^{-1}} - \frac{\beta}{1 + \beta\boldsymbol{v^T N^{-1}u}}\boldsymbol{N^{-1}uv^T N^{-1}} \qquad (14)$$

The above identity is useful in indirect inversion of an updated matrix to which some rows or columns are added or subtracted. If $\boldsymbol{N^{-1}}$ is available, the direct inversion of the updated matrix is prevented if the right-hand side of Eq. (14) is used. Because the term $\frac{\beta}{1+\beta\boldsymbol{v^T N^{-1}u}}$ is a scalar, computing the right-hand side of Eq. (14) is much simpler and computationally faster than the direct inversion of $\boldsymbol{N} + \beta\boldsymbol{uv^T}$.

Applying this identity helps in elimination of outliers' impact from an already estimated parameter vector without successive inversion of the normal matrix $\boldsymbol{N}$. This idea will be further elaborated in the following subsection.

3.2 Optimization of outlier elimination

The necessity of repeated estimation process in the adaptation step was explained previously. Equation (10) provides a possibility to update the estimated parameters when one outlier is to be removed from the dataset. The updating is fulfilled

by elimination of the numerical impact of the corresponding outlier form the estimated parameter vector. As a clarification, assume the same set of m observations of which the m^{th} element is identified as an outlier. The aim is to eliminate the contribution of this element from the estimated parameter vector. Let us have

$$\hat{\boldsymbol{x}} = \boldsymbol{N^{-1}u} \tag{15}$$

with $\boldsymbol{N} = \boldsymbol{A^T Q_y^{-1} A}$ and $\boldsymbol{u} = \boldsymbol{A^T Q_y^{-1}\, y}$. The updated parameter vector in which the effect of the outlier has been eliminated is written as

$$\hat{\boldsymbol{x}}_{\boldsymbol{m-1}} = (\boldsymbol{N} - \sigma_m^{-2}\boldsymbol{a_m a_m^T})^{-1}(\boldsymbol{u} - \sigma_m^{-2} y_m \boldsymbol{a_m}) \tag{16}$$

which is identical to the expression in Eq. (11). Denoting the updated normal matrix

$$\boldsymbol{N_{m-1}^{-1}} = (\boldsymbol{N} - \sigma_m^{-2}\boldsymbol{a_m a_m^T})^{-1} \tag{17}$$

As well as the updated $\boldsymbol{u}$ vector

$$\boldsymbol{u_{m-1}} = \boldsymbol{u} - \sigma_m^{-2} y_m \boldsymbol{a_m} \tag{18}$$

where σ_m, y_m and $\boldsymbol{a_m^T}$ are the standard deviation, the observation value, and the row vector (in $\boldsymbol{A}$) corresponding to the outlier, respectively. Indirect inversion of the updated normal matrix is possible through application of Eq. (14). This gives

$$\boldsymbol{N_{m-1}^{-1}} = \boldsymbol{N^{-1}} + \frac{\sigma_m^{-2}}{1 - \sigma_m^{-2}\boldsymbol{a_m^T N^{-1} a_m}} \boldsymbol{N^{-1} a_m a_m^T N^{-1}} \tag{19}$$

Denoting $\boldsymbol{b} = \boldsymbol{N^{-1}a_m}$, $k = \boldsymbol{a_m^T b} = \boldsymbol{a_m^T N^{-1} a_m}$, $p_m = \sigma_m^{-2}$, and $k' = \boldsymbol{b^T u} = \boldsymbol{a_m^T N^{-1} u}$ results in

$$\boldsymbol{N_{m-1}^{-1}} = \boldsymbol{N^{-1}} + p_m(1 - kp_m)^{-1}\boldsymbol{b\, b^T} \tag{20}$$

and therefore

$$\hat{\boldsymbol{x}}_{\boldsymbol{m-1}} = \boldsymbol{N_{m-1}^{-1} u_{m-1}} \tag{21}$$

After multiplication and a few simple mathematical and algebraic operations, the preceding equation simplifies to

$$\hat{\boldsymbol{x}}_{\boldsymbol{m-1}} = \hat{\boldsymbol{x}} - p_m(1 - kp_m)^{-1}(y_m - k')\boldsymbol{b} \tag{22}$$

The second term in the right-hand side of Eq. (19) is a vector which implies the numerical contribution of the m^{th} observation in estimation of parameters. The updated parameter vector is obtained by subtracting the impact of the single observation element from the pre-estimated parameter vector. Here p_m, k' and k are all scalars and $\boldsymbol{b} = \boldsymbol{N^{-1}a_m}$ is a vector of size n. Therefore, computation of the right hand-side of this equation is computationally much more efficient than the usual successive inversion of the normal matrix. Finally the updated parameter vector $\hat{\boldsymbol{x}}_{\boldsymbol{m-1}}$ and the inverted normal matrix $\boldsymbol{N_{m-1}^{-1}}$ are used to update the estimated residuals as well as their covariance matrix. This updated information is further used to construct the test statistics for the next iteration of DIA algorithm.

It is worthwhile mentioning that the above-mentioned idea is conceptually similar to the sequential adjustment expressed in Mikhail and Ackermann (1976). It should be noted that in the sequential adjustment one usually deals with adding new observations in sequential process, while, here the aim is to remove observations (or outliers) in a sequential manner.

The required calculation for eliminating the contribution of a single observation is significantly shortened in this manner. The parameters are estimated once only and are updated by Eq. (22) wherever elimination of outliers' contribution is necessary. Therefore, inverting (huge) matrices in Eq. (7) is replaced by the products of a few scalars and vectors in Eq. (22). Figure 1 summarizes the proposed optimized sequential data-snooping to give an algorithmic overview of this approach.

4. Numerical results and discussions

In this section three different examples are provided to evaluate the proposed method. In the first two examples, real datasets are adopted while the third is a simulated case. In the first example, the validity of the proposed formulation is investigated by comparing it against the conventional approach. The second example reveals the efficiency of the optimized formulation while applying it to a common geodetic application. Finally, the third one investigates the time efficiency of the proposed approach as well as the general performance of the DIA algorithm in correct detection of outliers.

Example 1: Leveling network

A leveling network of six points and nine observations is considered. The data is obtained in 2010 in a precise leveling campaign in the vicinity of the leveling benchmark at University of Isfahan, Iran. The first station of the network is the benchmark, with the known height of H=1708.3933 m, which can resolve the datum defect of the defined network. Having nine observations and five height parameters to be estimated, the network has four degrees of freedom. The observation precision is in the order of one millimeter.

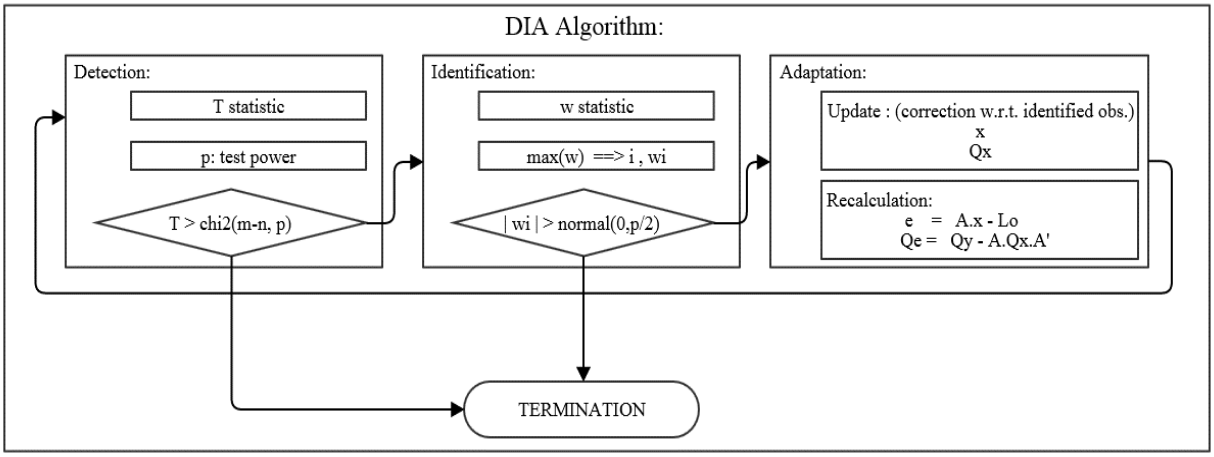

Figure 1: Flowchart of the optimized DIA algorithm; the optimization occurs in the adaptation step by updating the initial estimation result instead of repetition of the estimation procedure

is deliberately contaminated by gross error of one centimeter magnitude. The DIA method is applied in order to account for the imposed outlier. Finally both conventional and proposed methods are employed to perform the adaptation step of the DIA algorithm.

Table 1 summarizes this experiment. The observation scenario is given in the first four columns of the table. The observed values are the height differences between the stations. The result of the experiment is provided in the last four columns of the table. The height estimation for each of the stations is reported in the sixth column. Application of the DIA algorithm reveals the inconsistency between the observations, and finally the data-snooping correctly identifies the third observation as an outlier. Adapting the estimation result with elimination of the outlier gives rise to unbiased estimation of height parameters. This result is obtained by both conventional and optimized adaptation methods and is reported in columns seven and eight, respectively. Comparing the adaptation results reveals that both conventional and optimized methods are completely identical. This proves the numerical accuracy of the proposed formulation. Finally the last column reposts the estimation bias as a result of the contaminated observation.

As a conclusion, the experiment proves that the presented formulation is scientifically valid, correctly developed, and accurately implemented, hence assures the safety of application of the proposed formulation to linear models.

Example 2: GPS time series analysis

To illustrate the efficiency of the proposed method, the DIA algorithm is applied to a larger estimation scenario of GPS position time series analysis. The time series of latitude, longitude and height of permanent GPS station in Wettzzell/ Germany is used in this experiment. The data is available at ftp://sideshow.jpl.nasa.gov/pub/usrs/mbh/point/.
This data is obtained over 14-year period of GPS observations from 1999 to 2013 and is processed in the precise point positioning mode at GPS analysis center, Jet Propulsion Laboratory (Beutler et al., 1999), using the GIPSY software (Zumberge et al., 1977). The time series includes 4500 epochs. The goal of this experiment is to retrieve the deformation signal as well as the relevant periodic signals superimposed on it.

Because the times series are generally known to be noisy and occurrence of outliers are very common in GPS position time series (see Khodabandeh et al., 2012), the data-snooping gains high importance in performing an unbiased estimation of the signal of interest.

The functional model for GPS position time series is $\boldsymbol{y}(t) = \boldsymbol{y_0} + r\boldsymbol{t} + \sum_{k=1}^{q} a_k \cos \omega_k \boldsymbol{t} + b_k \sin \omega_k \boldsymbol{t}$, where $\boldsymbol{y}(t)$ is the vector of time-series observations, i.e. the latitudes, longitudes or height, $\mathbf{t}$ refers to the time instant vector, $\boldsymbol{y_0} + r\boldsymbol{t}$ is the linear trend which describes the deformation behavior of the station. Additional to the deformation signals, the periodic patterns are described by the harmonic functions. The following frequencies are found to be relevant in GNSS time series: annual and semiannual, period of 13.63 and 14.8 days as well as periods of 351.4 days and its higher harmonics 351.4/n, n = 2,..., 10 (Amiri-Simkooei, et. al. 2007, 2009; Amiri-Simkooei, 2013).

The DIA is applied to the estimation of unknown parameters. Both the conventional and the optimized methods are employed in the algorithm to have an assessment of the efficiency of the proposed method. The three coordinate components of latitude, longitude and height are considered, separately. Table 1, gives a summary of the comparison made between the two approaches. As expected, the number of outliers detected by the two

Table 1: Estimation result before and after applying the DIA algorithm to the leveling network

Network Observations				Network Parameters				
					Estimated Heights (m)			
Observation ID.	Back Station	Fore Station	Observation (ΔH)	Point ID.	Prior to Application of DIA	After Application of DIA		Δx Before and after DIA
						Conventional Method	Optimized Method	
1	1	2	1.9161	2	1706.4782	1706.4778	1706.4778	0.0004
2	2	3	2.0664					
3	3	4	2.0647	3	1704.4121	1704.4092	1704.4092	0.0029
4	4	5	-1.9818					
5	5	6	-2.5832	4	1702.4448	1702.4464	1702.4464	0.0016
6	6	1	-1.3826					
7	2	4	4.0291	5	1704.4286	1704.4283	1704.4283	0.0003
8	3	5	-0.0213					
9	4	6	-4.5613	6	1707.0096	1707.0101	1707.0101	0.0004

methods is identical. Out of the 4500 observations of the time series, 826, 1506 and 1543 of the data were masked as outliers for the latitude, longitude and height, respectively. The estimated parameters are also found to be identical. The computational time of the identification and adaptation process decreases for the presented method. This reduction is on average about 20%.

Figure 2 illustrates the outcome of the estimation and DIA algorithm for the three position components of the time series. In this figure, the correct observations and the outliers are marked by the squares and circles, respectively. The estimated signal is shown by a continuous line.

Example 3: Simulated dataset

The optimized method gains more credit in cases where a greater number of parameters are available in the estimation set-up. These cases are of common practice in satellite data processing as well as national geodetic network inversion. To illustrate such a case, a dataset of 2000 observables is simulated. The number of parameters is considered to be 1000 in this case. A random design matrix of size 2000×1000 is multiplied with a random vector of 1000 elements, to simulate the observation vector. A normally distributed random error vector is also added to this product to simulate the random noise superimposed on the observations. The magnitude of the random error is 10^{-3}. 100 elements of the simulated dataset are contaminated by gross error with magnitude of 10^{-1}.

Applying the DIA method reveals that the algorithm successfully identifies the contaminated observations. It is concluded that in this specific case (with the error of two orders of magnitudes higher than the observations' precision and the test power of 95 percent) the correct detection rate is 100 percent. This result degrades with lower error magnitude and is also affected by the test power. Taking the DIA algorithm, none of the correct observations is detected as outlier and no contaminated data is missed by the detection, resulting in the false alarm and missed detection rate of zero percent. The false alarm rate, missed and correct detection rates are under direct impact of the statistical test power. The interested reader is referred to Teunissen et al. (2005) for further information on this effect.

The conventional and optimized approaches are then considered in the adaptation step. The performance of the optimized method is faster by a factor of 3, compared to the conventional approach. The result of the DIA algorithm taking the two mentioned approaches is summarized in Table 3. As a conclusion, the larger the number of observations and parameters are, the more competitive the presented methodology will be. This is directly related to the normal matrix of $\boldsymbol{N = A^T Q_y^{-1} A}$ whose inversion is prevented in the optimized DIA method.

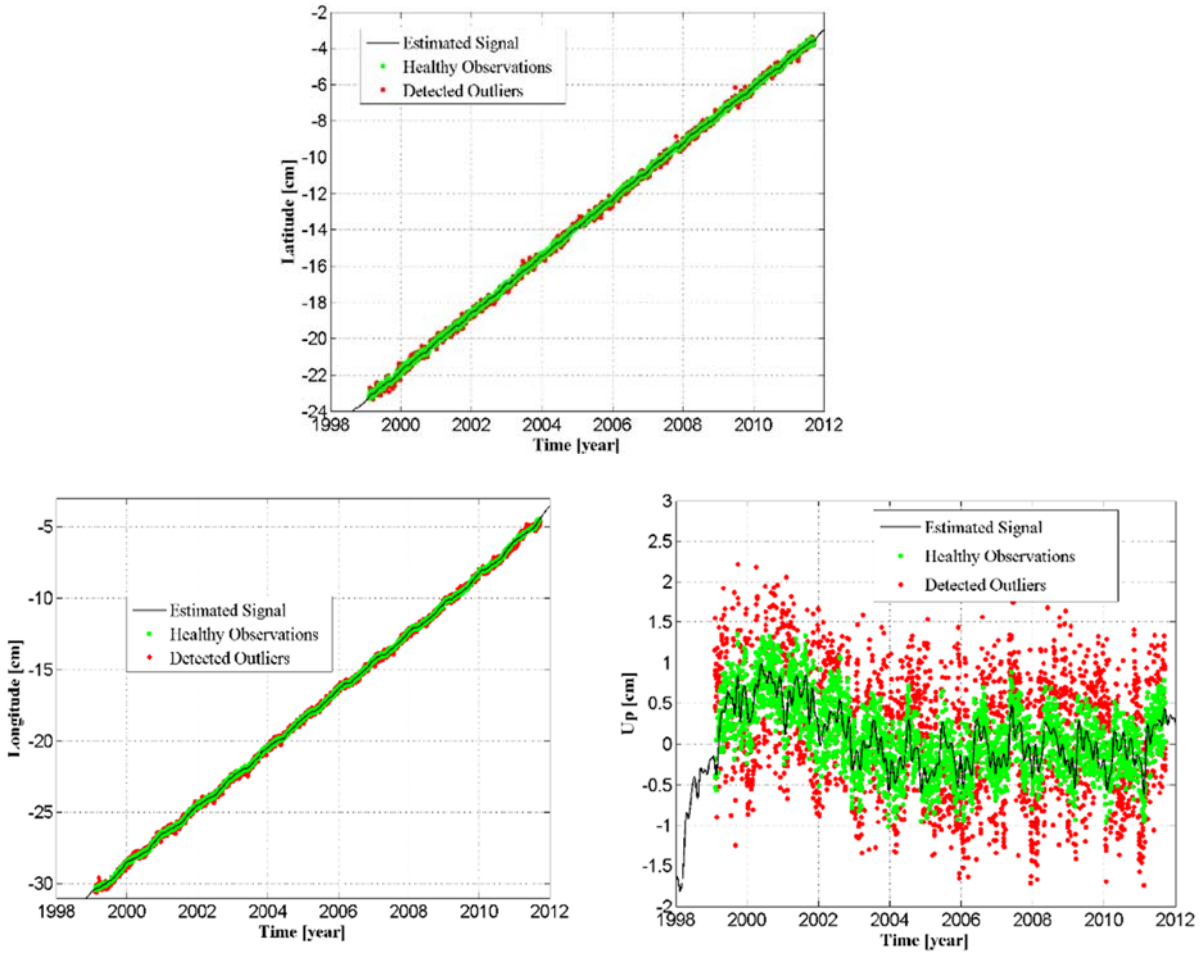

Figure 2: GPS time series analysis results; the latitude time series observations, the estimated signal and detected outliers are depicted for the entire data collection period

Table 2: GPS time series outlier detection results; application of DIA taking the conventional and optimized approach

Observations: 4500 **Parameters : 34**	**Latitude Time Series**	**Longitude Time Series**	**Height Time Series**
Computational time **Conventional Method**	346 sec	538 sec	555 sec
Computational Time **Optimized Method**	285 sec	452 sec	462 sec
Number of Outliers	826	1506	1543
Results of **Parameters Estimation**	Identical for two methods	Identical for two methods	Identical for two methods
Results of Detected Outliers	Identical for the two methods	Identical for the two methods	Identical for the two methods
Reduction in Computational Time	21 %	19%	20 %

Table 3: Results on the simulated example

Parameters : 1000 Observations : 2000 Contaminated Observations: 100	Conventional method	Optimized method
Algorithm run time	1331 sec	832 sec
Number of Detected Outliers	100 Obs.	100 Obs.
Results of Unknown Parameters' Estimation	Identical for the two methods	
Results of Detected Outliers	Identical for the two methods	
Reduction in Computational Time	60 %	

The normal matrix is a quadratic matrix of size $n \times n$, with n being the number of parameters in the estimation set-up. The larger this matrix is the more costly and less accurate the inversion calculation will be. While the successive inversion of this matrix is indispensable in the conventional adaptation method, the optimization employs the update strategy for which no direct inversion of any matrices is necessary.

5. Summary and concluding remarks

An algebraic formulation is proposed in the adaptation step of the sequential DIA algorithm in order to optimize the performance of the conventional method in elimination of outliers after their identification. The optimization is performed to improve the computational time and complexity of the conventional method. The presented method is investigated by three different types of real and simulated geodetic applications to prove its feasibility, efficiency and numerical accuracy. It is concluded that the optimized approach leads to the same results as the conventional method, both in terms of outlier identification and parameter estimation. This equivalency proves the accuracy of the proposed formulation and assures the safety in its application. It is also observed that the optimized method reduces the computational burden of the process. This reduction is dependent on the size of the dataset as well as the number of parameters involved.

The detection rate is related to the statistical test power as well as the magnitude of the contaminating error. The detection rate was shown to be 100% in the investigated test case. As a final remark, the proposed method can be applied to identification and adaptation of outliers in linear models. The generalization of the formulation to linearized models is an open topic for future research.

Acknowledgements

The authors would like to acknowledge the support of S. Montazeri for the technical discussions and comments as well as the comments of anonymous reviewers that improved the quality of this work.

References

[1] Amiri-Simkooei, A. R. (2013). On the nature of GPS draconitic year periodic pattern in multivariate position time series. Journal of Geophysical Research: Solid Earth, 118(5):2500–2511

[2] Amiri-Simkooei, A. R., (2009). Noise in multivariate GPS position time-series, Journal of Geodesy, 83, 175–187, doi: 10.1007/s00190-008-0251-8.

[3] Amiri-Simkooei, A. R. (2003). Formulation of L1 norm minimization in Gauss-Markov models, Journal of Surveying Engineering, 129(1), 37–43.

[4] Amiri-Simkooei, A. R., Tiberius, C. C. J. M. and Teunissen P. J. G., (2007). Assessment of noise in GPS coordinate time series: Methodology and results, Journal of Geophysical Research, Vol. 112, B07413, doi: 10.1029/2006JB004913.

[5] Awange, J.L., Aduol, F. W. O., (1999). An evaluation of some robust estimation techniques in the estimation of geodetic parameters, Survey Review, 35 (273), 146–162

[6] Baarda, W., (1968). A testing procedure for use in geodetic networks, Netherlands Geodetic Commission, Publication on Geodesy, New Series, Vol. 2(5), Delft, The Netherlands.

[7] Ben-Gal, I., Maimon, O., Rockach L., (2005). Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers, ISBN: 0-387-24435-2

[8] Berber, M., Hekimoglu, S. (2001). What is the reliability of robust estimators in networks? First International Symposium on Robust Statistics and Fuzzy Techniques in Geodesy and GIS March 12–16, ETH, Zurich, pp 61–66 Bickel PJ (1978) Using residuals robustly

[9] Beutler, G., Rothacher, M., Schaer, S., Springer, T. A., Kouba, J., Neilan, R. E. (1999). The International GPS Service (IGS): an interdisciplinary service in support of Earth sciences, Advances in Space Research, 23(4), 631–635.

[10] Caspary, W. (1998). Anmerkungen zur balancierten Ausgleichung, Zeitschrift fur Vermessungswesen,8, 271-272.

[11] Davies, L., Gather, U. (1993). The identification of multiple outliers, Journal of the American Statistical Association, 88(423), 782-792.

[12] Fuchs, H., (1982). Contribution to the adjustment by minimizing the sum of absolute residuals, Manuscripta Geodetica, 7, 151–207

[13] Gao, Y., Krakiwsky, E.J., Czompo, J. (1992). Robust testing procedure for detection of multiple blunders, Journal of Surveying Engineering, 118(1), 11-23

[14] Hampel, F. (2001).Robust statistics, Proc., first International Symposium on Robust Statistics and Fuzzy Techniques in Geodesy and GIS, Zurich, Switzerland,13–17.

[15] Hampel, F. R., Ronchetti, E.Z., Rousseeuw, P.J., Stahel, W.A. (1986). Robust Statistics: The Approach Based on Influence Functions, Wiley, New York.

[16] Henderson, H. V., Searle, S. R., (1981). On deriving the inverse of a sum of matrices, Journal of Siam Review, 23 (1), 53-60

[17] Huber, P. J., (1981). Robust statistics, Wiley, New York.

[18] Johnson, R., (1992). Applied Multivariate Statistical Analysis, Prentice Hall.

[19] Khodabandeh, A., Amiri-Simkooei, A.R. (2011). Recursive algorithm for L1 norm estimation in linear models, Journal of Surveying Engineering, 137(1), 1-8

[20] Khodabandeh, A, Amiri-Simkooei, A.R., Sharifi, M.A. (2012). GPS position time-series analysis based on asymptotic normality of M-estimation, Journal of Geodesy, 86 (1), 15-33

[21] Koch, K. R., (1999). Parameter estimation and hypothesis testing in linear models, 2nd Edition, Springer, Berlin, Germany.

[22] Kok, J. J. (1984). On data snooping and multiple outlier testing, NOAA Technical Report NOS NGS 30, National Geodetic Information Center, NOS/NOAA, Rockville, Md.

[23] Marshall, J. (2002). L1-norm pre-analysis measures for geodetic networks,Journal of Geodesy76, 345–352

[24] Marshall, J. and Bethel, J. (1996). Basic concepts of L1 norm minimization for surveying applications, Journal of Surveying Engineering, 122 (4), 168–179

[25] Mikhail, E. and Ackermann, F. (1976). Observations and Least Squares, University Press of America.

[26] Pelzer, H. (1983). Detection of errors in the functional adjustment model, Deutsche Geodetic Commission, Reihe A, Nr. 98, 61–70, Munich, Germany.

[27] Pope, A. (1976). The statistics of residuals and detection of outliers, NOAA Technical Report, No. 66, NGS 1, Rockville, Md.

[28] Rousseeuw, P. J., Leroy, A. M. (2003). Robust Regression and Outlier Detection, Wiley-interscience paperback series, Wiley

[29] Schwarz, C. R., Kok, J. J. (1993). Blunder detection and data snooping in LS and robust Adjustment", Journal of surveying engineering, 119(14), 127-136

[30] Teunissen, P. J. G. (2000). Testing Theory: an introduction. Website: http://www.vssd.nl: Delft University Press. Series on Mathematical Geodesy and Positioning.

[31] Teunissen, P.J.G., Simons, D.G., Tiberius, C.C.J.M., (2005). Probability and Observation Theory, Lecture notes AE2–E01, Faculty of Aerospace Engineering, Delft University of Technology, the Netherlands.

[32] Wicki, F. (1999). Robuste Schätzverfahren für die parameterschätzung in geodätischen Netzen, Institut für Geodäsie und Photogrammetrie an der ETH, Zürich, Mitt. Nr. 67

[33] Xu, P. L. (1989). On the robust estimation with correlated observations, Bull Geod 63, 237–252

[34] Youcai, H. (1995). On the design of estimators with high breakdown points for outlier identification in triangulation networks, Bull Geodesy 69, 292–299.

[35] Zumberge, J.F., Heflin, M.B., Jefferson, D.C., Watkins, M.M., Webb, F.H. (1997).Precise point positioning for the efficient and robust analysis of GPS data from large networks. Journal of Geophysical Research, 102, 5005–5017.