A Robust Method for Multiple Outliers Detection in Multi-Parametric Models

Albert K. Chong
San Diego, CA 92116

ABSTRACT: An analytical method consisting of concepts from the Danish robustified least squares, Pope's TAU statistic, and stepwise analysis is discussed. Simulated data were used to test the efficiency and reliability of the method. The collinearity condition in photogrammetry was the basic mathematical model. A flow diagram of the method and a summary of test results of test samples are included to provide more information on the method. The method can detect up to \((n - (n + u)/2)\) outliers in small samples (where \(n\) is the sample size and \(u\) is the number of unknowns in the mathematical model). In large samples, more than \(n - (n + u)/2\) outliers may be detectable. Geometric checkpoints were also used to reduce the effect of poor geometry of selected observations (especially in photogrammetry).

INTRODUCTION

Many methods are now available for the detection of multiple outliers in multi-parametric estimation of photogrammetric and geodetic analytical models (e.g., Baarda (1968), Pope (1976), Hawkins (1981), and Cross (1985)). None of those, tested statistically, is suitable for samples having a large number of outliers. Five samples are given in Table 1 to show the largest normalized residual of each sample and the corresponding critical test values of Baarda's data snooping and Pope's TAU statistic. A large number of outliers may occur in observations (Chong, 1986); therefore, research was carried out to find a method which would be effective and reliable for analyzing observations with multiple outliers.

A brief description of each basic concept is presented, followed by a detailed description of the method. Results of a typical test sample are given to show the flow of data from observations to final solution.

ROBUSTIFIED LEAST SQUARES

Robustified least squares is a modification of the Danish method. The algorithm is shown below.

By making the weights, a function of \(\psi\) where \(\psi = V/\sigma_V\), then

\[
P = P_i(l(V_i/\sigma_{V_i}, \sigma_v, a))
\]

i.e.,

\[
P = \begin{cases} 
1 & \text{for } |\psi| < a; \ a = \text{assigned constant} \\
0 & \text{else}
\end{cases}
\]

where \(V_i = \text{estimated residual of observation } i\), \(P_i = \text{a priori weight of observation } i\), \(\sigma_v = \text{estimated sigma of residual } V_i\), \(\sigma_r = \text{reference sigma, and} \ a = \text{a predetermined numerical constant of Pope's TAU statistic.}\)

Initially, a constant weight \((P)\) is introduced in the traditional least-squares adjustment. New observation weights are computed from the \(\psi\) of the previous adjustment according to the requirement in Equation 1. A detailed description of the procedure is given in "methodology."

TAU STATISTIC

Pope (1976) discussed the TAU statistic in detail (the selection of TAU statistic was based on research findings in Chong (1986), p. 119). It is based on a distribution which he called the TAU distribution.

TAU is given as

\[
\tau = \frac{\vert \psi \vert}{(\gamma - 1 + t_{\gamma - 1}^{-2})^{1/2}}
\]

where \(t_\gamma = \text{Student's-t distribution,} \ \gamma = (n - u) \text{ or degrees of freedom,} \ \ n = \text{sample size, and} \ u = \text{number of unknown variables.}\)

REJECTION TEST OF TAU

A critical value \(\tau\), is selected in advance which is based on \(\alpha\), the probability of rejecting a true hypothesis (the probability of type I error). The probability of accepting a true hypothesis is the significance of the test, \(1 - \alpha\), while the probabilities of

<table>
<thead>
<tr>
<th>Sample #1</th>
<th>Sample #2</th>
<th>Sample #3</th>
<th>Sample #4</th>
<th>Sample #5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Sample</td>
<td>Number of Outliers</td>
<td>30</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>Information</td>
<td>Size of Outliers</td>
<td>13</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Critical Test</td>
<td>Baarda's Data Snooping ((\omega_o))</td>
<td>67</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Value at (\alpha = 0.05) ((\tau_o))</td>
<td>Pope's TAU</td>
<td>2.91</td>
<td>2.84</td>
<td>2.69</td>
</tr>
<tr>
<td>Largest Normalized Residuals ((V_o/\sigma_v))</td>
<td>1.88</td>
<td>1.88</td>
<td>1.61</td>
<td>1.86</td>
</tr>
</tbody>
</table>

*The critical test values of Baarda's Data Snooping (\(\omega_o\)) and Pope's TAU (\(\tau_o\)) were larger than the largest normalized residuals for all the five samples. The developed method was able to isolate all the outliers from the five samples successfully.*

0099-1112/87/5306-617$02.25/0 ©1987 American Society for Photogrammetry and Remote Sensing

accepting the hypothesis when false, $\beta$, and rejecting a false hypothesis, $1 - \beta$, are the probability of type II error and the power of the test, respectively. According to Pope (1976), the rejection procedure is thought of as a test of the hypothesis that $V_i \sim n(0, \sigma_{\nu_i})$ for all $i$,

where $\sigma_{\nu_i} = L(\hat{\sigma}, \nu_i) = L(\Omega_i / P_i)$ for some unspecified $L$, by use of the test

$$H_0$$ if $\max (|V_i / \hat{\sigma}|) > \tau_i$

$$H_1$$ if $\max (|V_i / \hat{\sigma}|) < \tau_i$

where $Q = (I - A(A'PA)^{-1}A'P)$ and $A$ is the design matrix (for further information, refer Pope to (1976)).

$\tau_i$ is difficult to compute because it requires the distribution of $\max \tau$ under the null hypothesis. Pope (1976) proposed a simplified derivation that evades many of the subtleties involved in a more thorough approach, i.e., $\alpha = P(\max \tau > \tau_i) = P(\text{one or more of } \tau_i > \tau_i) = 1 - P(\text{all } \tau_i \leq \tau_i)$. As such, $\alpha = 1 - (1 - \alpha)^n$ ignoring the dependence of $\tau_i$, where $\alpha = P(\tau_i > \tau_i) = 1 - (1 - \alpha)^n \equiv \alpha/n;

n = \text{total number of observations.}$

**COMPUTATION OF CRITICAL VALUE**

The critical values of TAU ($\tau_i$) could be calculated by a subroutine TAURE, which was given by Pope (1976).

**SIGNIFICANCE LEVEL**

The significance level ($\alpha$) was set at 0.001 (determined by simulation test). Other popular significance levels (e.g., 0.05 and 0.01) may be used. The number of iterations would probably increase with the increase in significance level. A lower significance level is not encouraged because the subroutine on TAU may not generate accurate critical values.

**STEPWISE ANALYSIS**

The stepwise analysis consists of two major steps. The first step (backward selection) involves the selection of a subsample of observations for outlier testing. The second step (forward selection) involves the selection of all tested inliers and one suspected outlier to test whether the suspected outlier is actually an outlier.

**BACKWARD SELECTION**

From test results in Chong (1986), $(u + 2)$ is the most effective, reliable, efficient, and unbiased value for the initial subsample size. Therefore, $(u + 2)$ is adopted in the method.

**FORWARD SELECTION**

In this step, all outlying observations are added to the test inlier one at a time. The process continues until all outlying observations are tested statistically. It must be noted here that if the "smallest" (least $|V_i / \sigma_i|$, $\nu_i$) outlying observation is rejected, then normally all larger ones are candidates for rejection.

**TEST MODEL**

The collinearity condition in photogrammetry was used to test the developed method. The condition may be expressed in the following form, which is known as the collinearity equation:

\[
x_u = -f \left( m_{11}(X_u - X_l) + m_{12}(Y_u - Y_l) + m_{13}(Z_u - Z_l) \right)
\]

\[
y_u = -f \left( m_{21}(X_u - X_l) + m_{22}(Y_u - Y_l) + m_{23}(Z_u - Z_l) \right)
\]

where $X_u, Y_u, Z_u =$ photo coordinates of image point $N$,

$f =$ focal length of camera,

$m_{ij} =$ elements of rotational matrix,

$X_l, Y_l, Z_l =$ object space coordinates of exposure station of photo $L$, and

$X_u, Y_u, Z_u =$ object space coordinates of point $N$.

The above notations are expressed as the following terms in the analysis, i.e.,

$x_i, y_i =$ observations,

$m_{ij}, X_l, Y_l, Z_l =$ unknown variables, and

$f, X_u, Y_u, Z_u =$ known constants.

**TEST DATA**

Three hundred test samples of simulated data were used to test the efficiency and reliability of the method (refer to Chong (1986), pp. 75–87). Two hundred sixty of these test samples are mathematically simulated and the rest are obtained by laboratory simulation.

Mathematically simulated data are generated by computing the variables which would satisfy a photogrammetric model under certain conditions (e.g., to compute the photo coordinates (or observations) of point $P_i$ given the coordinates of exposure stations, values of rotational elements, ground coordinates of $P_i$, and focal length of camera which would satisfy the collinearity condition). Simulated random errors and outliers (range from 1 to 100r) are added on to the computed photo coordinates ($\sigma = \text{standard deviation}$).

Laboratory simulated data are obtained by means of a control frame (three-dimensional model) and two Hasselblad MK-70 close-range cameras. A glass tank filled with water is placed between the cameras and the control frame. The "tank" is used to generate a local disturbance, i.e., objects which are obstructed by the "tank" are displaced from their actual positions on the exposures. Observed photo coordinates of displaced images become outliers.

**METHODOLOGY OF THE ROBUST METHOD**

The method was perfected after extensive testing. For further information relating to this method, one may refer to the dissertation by Chong (1986). Flow diagrams are presented in Figure 1 and Figure 2.

1. Observations are initially analyzed by the method of least squares in the usual way.
2. For each observation, the normalized residual $|V_i / \sigma_i|$ is computed.
3. $(u + 2)$ observations having the least $|V_i / \sigma_i|$ values are selected and are assigned weights of 1. All other observations are assigned weights of 0.
4. For analytical models where geometry is important (e.g., collinearity condition), initial approximate values of unknowns (e.g., Z coordinate of exposure station of collinearity equations), $\sigma_{\nu_i}$ (estimated standard error of observations), or $\sigma_\nu$ (design standard error of observations) may be used as the preselected test values for the geometry check. An example of the geometry check is given in Figure 2. If any of the checks fail, an outlying observation may be selected to replace the weak observation (one having the largest $|V_i / \sigma_i|$).
5. All $|V_i / \sigma_i|$ are computed (1st iteration) with the assigned weights. All observations with weights of 1 are tested against $\tau_i$ at a sample size of $(u + 2)$. If one of these observations is rejected, it is replaced by the observation having the next lower $|V_i / \sigma_i|$ in the initial least-squares solution. If all replacements are rejected, then one may use the acceptable observations to compute the final solution. If all the initial $(u + 2)$ observations are acceptable, one may assign weights of 0 to any observation in the whole sample whose $|V_i / \sigma_i|$ is greater than or equal to $\tau_i$. Except one rejected observation having the least $|V_i / \sigma_i|$ (this observation is called "the selected outlying observation"). This observation together with other accepted observations are assigned weights of 1.
6. Step 4 is repeated and all $|V_i / \sigma_i|$ are computed (2nd iteration) with the assigned weights. If the selected outlying observation is rejected, one may select from the previous iteration an out-
MULTIPLE OUTLIERS DETECTION

lying observation having the next lowest \( |V/\sigma_0| \) (normally not necessary because once an outlying observation is rejected, others are also candidates for rejection). Again, if all replacements (one at a time) are rejected, then one can use the accepted observations to compute the final solution. If a selected outlying observation is accepted, then all the acceptable observations \( (|V/\sigma_0| < \tau) \) in that iteration are assigned weights of 1. Here again, one outlying observation having the lowest \( |V/\sigma_0| \) is also assigned a weight of 1, and all other rejected observations are assigned weights of 0.

(7) All \( |V/\sigma_0| \) are computed (3rd iteration and up) with the assigned weights and step 6 is repeated.

TEST RESULTS

Table 2 provides a summary of the test results. Each test sample carries \( n - (n + u)/2 \) outliers. Test samples are grouped according to the size of outliers (e.g., a sample in group \( u = 5 \) has outliers whose sizes range from 0 to 25). The percentage of samples tested successfully was based on the number of test samples for each group. For further information, one may refer to Chong (1986).

CONCLUDING REMARKS

The developed robust test method has many unique qualities. Some of the qualities are as follows:

- It makes use of robustified least squares which is resistant to outliers;
- It makes use of all observations that are considered good observations with reference to a statistical distribution in the final solution;
- It uses the most advanced and very popular Pope's TAU statistic;
- It uses checkpoints to detect weak geometry in observations for geometry sensitive analytical models;

**FIG. 1. Generalized flow diagram of robust test method.**

**FIG. 2. A flow diagram for analyzing test data by the developed method.**

\[ \text{STOP} \]
### Table 2. Summary of Test Results

<table>
<thead>
<tr>
<th>Type of test data</th>
<th>Size of outliers (in (u))</th>
<th>Number of outliers in each sample</th>
<th>Number of successful sample</th>
<th>Number of successful case</th>
<th>Percent of successful test*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematically</td>
<td>0 to 25</td>
<td>(n - (n+u)/2)</td>
<td>65</td>
<td>55</td>
<td>84.6</td>
</tr>
<tr>
<td>simulated</td>
<td>25 to 50</td>
<td>for all the</td>
<td>65</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>data</td>
<td>50 to 75</td>
<td>test samples</td>
<td>65</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>Laboratory</td>
<td>75 to 100</td>
<td></td>
<td>65</td>
<td>65</td>
<td>100</td>
</tr>
<tr>
<td>simulated</td>
<td>0 to 25</td>
<td>(n = ) sample size</td>
<td>10</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>data</td>
<td>25 to 50</td>
<td>(u = ) unknown</td>
<td>15</td>
<td>15</td>
<td>100</td>
</tr>
<tr>
<td>data</td>
<td>75 to 100</td>
<td>variables</td>
<td>15</td>
<td>15</td>
<td>100</td>
</tr>
</tbody>
</table>

*The percentage of the samples that all outliers in each sample were successfully isolated.

- It can detect up to \((n - (n+u)/2)\) outliers in small samples. In large samples, more than \((n - (n+u)/2)\) outliers may be detectable.
- It can be implemented in any existing adjustment program which uses a conventional least-squares approach, by means of one or two subroutines.

### REFERENCES


(Received 23 August 1986; revised and accepted 4 March 1987)

### Seminar on Photogrammetric Mapping from SPOT Imagery

**Hannover, Federal Republic of Germany**

**21-23 September 1987**

This Seminar — organized by the Institute for Photogrammetry and Engineering Surveys of the University of Hannover — is intended to introduce the participant to the photogrammetric evaluation of SPOT imagery. Subject matter will include

- The bundle adjustment program BINGO for control extension with SPOT imagery
- Stereo plotting of SPOT imagery on the Carl Zeiss C100 Planicomp
- Orthophotomapping with SPOT imagery on the Carl Zeiss Z2 Orthocomp
- Image processing of SPOT imagery on the ContextVision GOP 300 Image Processing System

The registration fee is DM 250 and includes seminar material, lunches, and coffee. An inauguration dinner sponsored by Carl Zeiss, Oberkochen will be offered on Monday evening.

For further information regarding registration and the program of the Seminar, please contact

Seminar Secretariat

% Institute for Photogrammetry and Engineering Surveys

University of Hannover

Nienburger Strasse 1

D-3000 Hannover 1

Federal Republic of Germany

Tel. 0049-511-762-2724