Multiple Outlier Detection: Hypothesis Tests versus Model Selection by Information Criteria

Authors

Rüdiger Lehmann
University of Applied Sciences Dresden
Faculty of Spatial Information
Friedrich-List-Platz 1
D-01069 Dresden, Germany
Tel +49 351 462-3146
Fax +49 351 462-2191
mailto:ruediger.lehmann@htw-dresden.de

Michael Lösler
Frankfurt University of Applied Sciences
Faculty of Architecture, Civil Engineering and Geomatics
Nibelungenplatz 1
D-60318 Frankfurt am Main, Germany
Tel +49 69 1533-2784
Fax +49 69 1533-2058
mailto:michael.loesler@fb1.fra-uas.de

Abstract

The detection of multiple outliers can be interpreted as a model selection problem. Models that can be selected are the null model, which indicates an outlier free set of observations, or a class of alternative models, which contain a set of additional bias parameters. A common way to select the right model is by using a statistical hypothesis test. In geodesy data snooping is most popular. Another approach arises from information theory. Here, the Akaike information criterion (AIC) is used to select an appropriate model for a given set of observations. The AIC is based on the Kullback-Leibler divergence, which describes the discrepancy between the model candidates. Both approaches are discussed and applied to test problems: the fitting of a straight line and a geodetic network. Some relationships between data snooping and information criteria are discussed. When compared, it turns out that the information criteria approach is more simple and elegant. Along with AIC there are many alternative information criteria for selecting different outliers, and it is not clear which one is optimal.

Keywords

Least squares adjustment; Outlier detection; Hypothesis test; Information criterion; Akaike information criterion (AIC); Data snooping; Model selection

Introduction

Geodetic outlier detection is firmly based on the fundament of statistical hypotheses testing in Gauss-Markov and Gauss-Helmert models. There are well-established and workable methods for outlier detection and they are also implemented in present-time geodetic standard software. The most important toolbox for geodetic outlier detection is data snooping, which is based on the pioneering work of *Baarda (1968)*. Later this work was continued by *Pope (1976), Heck (1981)* and others. Today, data snooping is the recommended outlier detection method in most geodetic textbooks *(Teunissen 2000)*. Nevertheless, there is a continued research on the subject *(Lehmann 2013a)*.

In modern geodesy, in which there are typically very large data sets, it is clear that it is not possible to rule out that a set of observations contains multiple outliers. However, data snooping was initially developed only for one outlier. It is common practice to apply data snooping consecutively, detecting one outlier after the other. It is also possible to set up statistical tests for the case of multiple outliers *(Kok 1984; Ding and Coleman 1996)*. The subject is covered at full length by *Teunissen (2000)*.

Multiple outlier detection is hampered by two phenomena:

- Swamping: If the maximum number of outliers to be detected is large then the statistical test "tends to declare more outliers than there are in the sample" *(Beckman and Cook 1983)*.
- Masking: Multiple outliers can mask each other, such that they are hardly detectable *(Rousseeuw and Leroy 1987, p. 222).*

Baselga (2011) showed impressively that in a geodetic adjustment different numbers and patterns of multiple outliers yield the same residuals, which means that the information that observations are outliers is not fully contained in the residuals. However, any outlier detection method is based on the residuals. Therefore, any decision on multiple outliers is partly based on assumptions, which cannot be checked without extra information.

In recent years the theory of reliability has been extended in the direction of multiple outlier detection *(Knight et al. 2010; Teunissen and de Bakker 2013)*. In this theory the minimum detectable bias (MDB) is of fundamental importance. Although the MDB is a scalar for one outlier, it becomes a vector with a dimension equal to the number of suspected outliers. The full MDB vector cannot be computed without some knowledge of the outliers, but bounds for the maximum MDB were computed by solving eigenvalue problems, thus, obtaining measures of reliability also for multiple outlier tests *(Knight et al. 2010)*.

Yang et al. (2013) treated the outlier separability problem when multiple outliers needed to be detected. A serious problem in outlier detection by hypotheses testing is the choice of decision error rates, i.e., significance levels. This is even worse for multiple outlier detection, as will be shown by this contribution. Practical applications sometimes try to avoid this choice by choosing critical values instead. This approach is critically discussed by *Lehmann (2013b)*. It is shown that a critical value cannot be chosen irrespective of the number of observations but must be increasing with this number.

Outlier detection can be viewed as model selection: The observations can be modeled as being free of outliers or accounting for various specific outlier patterns with respect to number, affected observations and stochastic properties. The question arises: Which is the appropriate model to be selected? The hypothesis tests provide criteria to answer this question.

A different approach that is less popular in geodesy arises from information theory. This field tries to formulate the discrepancy between the true and the candidate model in terms of the Kullback-Leibler divergence *(Hurvich and Tsai 1989)* and to estimate this discrepancy on the basis of the observations. The oldest and best established estimate is the Akaike information criterion (AIC). Later it was corrected for a bias, which arises, when the observational redundancy is small, resulting in a corrected version of AIC (AICc). Today, many different alternative information criteria have been proposed, the most important are Bayesian information criterion (BIC) and Mallows' Cp. A comprehensive textbook on the subject of model selection by information criteria is *Burnham and Anderson (2002)*.

Although information criteria were proposed very early by *Blais (1991)* as a tool for model identification in geodesy, there are only a small number of applications available. One application is transformation model selection *(Felus and Felus 2009; Lehmann 2014)*.

Another scope of application is the auto regressive moving-average process *(Klees et al. 2002)* especially in the framework of GNSS time series analysis *(Luo et al. 2011, 2012)*. *Lehmann (2015)* had proposed using information criteria for observation error model selection in geodetic adjustment. Recently, *Lösler et al.*

(2016) applied the AIC technique to avoid an overparametrization of the functional model of a network adjustment.

Thus far, there have been no investigations aimed at detecting outliers in geodetic observations. However, in other branches of science there are applications of information criteria also for outlier detection:

The suggestion to use information criteria for outlier detection came from *Kitagawa (1979)*. At this time, AIC was the only information criterion considered. Nonetheless, the conclusions of this work are quite universal: (1) the problems of determining the number of outliers and of identifying the outlying observations inherent in all classical outlier detection procedures are elegantly unified, (2) various situations (single and multiple and one-sided and two-sided outliers) can be treated consistently, and (3) no choice of a significance level is required.

This contribution inspired many other researchers: *Pynnönen (1992)* applied AIC and BIC to linear and quadratic regression problems. The elegance and simplicity of the approach is underlined. *Fung (1993)* investigated Bayesian analysis of outliers using a non-informative prior distribution for the parameters. It is well known that the problem with this approach is that the analysis is not invariant to the change of scale of the data. This is circumvented by using Akaike's log predictive likelihood for penalizing outlier models with extra parameters in the quasi-Bayesian method. This approach is shown to have a good performance for detecting outliers in a benchmark data set.

Atkinson and Riani (2008) effectively used AIC and Mallows' Cp for the analysis of outliers in ozone concentration data. *Ueda (2009)* presented a simple and efficient method to detect multiple outliers using a modification of the AIC, and it has been successfully applied to sample observations. *Kornacki (2014)* applied the AIC to the detection of outliers for the analysis of ash content in barley straw. The author also comes to the conclusion that the method has two advantages: (1) it "is an objective procedure independent of the assumed significance level, quantity of outliers and of whether the suspicious observations are the lowest or the highest" and (2) it avoids the masking effect of outlier detection.

In this paper the detection of multiple outliers by classical geodetic data snooping is compared with this information criterion approach. The paper is organized as follows. A short overview is given on the Gauss-Markov model for outlier detection. The subject of hypothesis tests for multiple outliers in the framework of geodetic data snooping is discussed in detail. The number of suspected outliers is not known in the case presented in this paper; it may as well be zero. Here, a particular problem arises when portioning the error rate of the multiple tests to the different individual tests. When the number of suspected outliers is not equal in all these tests, the portions are not necessarily the same size. For the selection of the alternative hypothesis, the equivalent of the decision of which observations are outliers, the p-value approach is newly proposed and elaborated.

As a counterpart, the information theoretic approach for multiple outlier detection is presented. AICc is used as a model selection, and in this way it is also used as an outlier selection criterion. Relationships between both the approaches are established, and they are both applied to the problem of fitting a straight line and to a geodetic network. Nonetheless, both can be applied to any problem that can be formulated as a Gauss-Markov or Gauss-Helmert model. Because a Gauss-Helmert model can be expressed as a Gauss-Markov model, the discussion is restricted to the common Gauss-Markov model without loss of generality.

The paper concludes with a comparison of the pros and cons of the hypothesis test and AICc criterion for multiple outliers.

Gauss-Markov Model for Outlier Detection

Starting from a linear or linearized functional adjustment model (observation equations)

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

with the known n vector of observations $\boldsymbol{l}$, the unknown n vector of observation errors $\boldsymbol{e}$, the unknown u vector of adjustment parameters $\boldsymbol{x}$ and the known $n \times u$ matrix $\boldsymbol{A}$ (matrix of observation equations), and from a stochastic adjustment model for normal distributed observation errors

$$e \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{P}^{-1}) \tag{2}$$

with a known positive definite symmetric weight matrix $\boldsymbol{P}$ and the a priori variance factor σ^2, which may be known or unknown, for the least squares solution of the vector of residuals

$$\boldsymbol{v} = -\boldsymbol{Q}_{vv}\boldsymbol{P}\boldsymbol{l} \tag{3}$$

the multivariate normal distribution

$$\boldsymbol{v} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}_{vv}) \tag{4}$$

with cofactor matrix of the residuals is found

$$\boldsymbol{Q}_{vv} = \boldsymbol{P}^{-1} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{P}\boldsymbol{A})^{-}\boldsymbol{A}^T. \tag{5}$$

The superscript minus sign symbolizes a generalized matrix inverse. It will be requested for rank-deficient adjustment models (e.g., free geodetic networks). If $\boldsymbol{A}$ and $\boldsymbol{P}$ have full column rank, then the generalized matrix inverse is unique and coincides with the classical matrix inverse. This model is also known as *Gauss-Markov model of geodetic adjustment*. This contribution is confined to regular models, in which no singular matrices occur. Singular cases can be treated in an analogous manner.

Note this was started from a linear or linearized model, while the observation equations of many geodetic problems like 2D or 3D networks are inherently nonlinear. The consequences for outlier detection will be analyzed later.

Alternatively, one may suspect a number of n_g gross errors affecting the observations. The common procedure is to extend the model by a n_g vector of bias parameters $\boldsymbol{\nabla}$, accounting for those gross errors as

$$\boldsymbol{l} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{C}\boldsymbol{\nabla} + \boldsymbol{e} = (\boldsymbol{A} \quad \boldsymbol{C})\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{\nabla}\end{pmatrix} + \boldsymbol{e} \tag{6}$$

where $\boldsymbol{C}$ is the $n \times n_g$ matrix relating gross errors to observations. Typically, $\boldsymbol{C}$ consists exclusively of elements with values 0 and 1, whereas 1 in row i and column k means that the k th gross error affects the ith observation by its full magnitude, such that it becomes an outlier; 0 implies that this gross error does not affect this observation at all *(Teunissen 2000, p. 37)*. For the sake of simplicity, it is subsequently assumed that $\boldsymbol{C}$ is of such a simple type. This type of alternative model, where $\boldsymbol{\nabla}$ are nonrandom bias parameters, is known as the *mean shift model* [see (*Lehmann 2013a*) for a synopsis of possible alternative models].

The alternative model [Eq. (6)] is now opposed to the null model [Eq (1)]. Note that setting up the alternative model (6) requires knowledge of the number of suspected outliers n_g and of the subset of affected outlying observations, coded in matrix $\boldsymbol{C}$. If this knowledge is not available then many alternative models in parallel can be set up as follows:

$$\boldsymbol{l} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{C}_j\boldsymbol{\nabla}_j + \boldsymbol{e} = (\boldsymbol{A} \quad \boldsymbol{C}_j)\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{\nabla}_j\end{pmatrix} + \boldsymbol{e}, \;\; j = 1, \dots, m \tag{7}$$

where each $\boldsymbol{C}_j$ is a $n \times n_{g,j}$ matrix; and each $\boldsymbol{\nabla}_j$ is a $n_{g,j}$ vector. The theoretical maximum number of alternative models depends on the total number of observations n and the assumed maximum number of suspected outliers $n_{g,max}$, and is given by

$$m = \sum_{n_g=1}^{n_{g,max}} \binom{n}{n_g} \tag{8}$$

For example, one may get $m = n$ alternative models with one outlier ($n_{g,\max} = 1$). This approach is theoretically only limited by the total redundancy $r = n - u$: The total number of parameters $u + n_{g,\max}$ of the alternative model must not exceed the number of observations n. Therefore, $n_{g,\max} < r$ is required. From practical considerations of reliability, the maximum number of suspected outliers should be much smaller. There may be other practical reasons to strongly restrict the set of alternative models; e.g., there may be pairs of observations being either both outliers or not.

It appears to be sufficient to specify only alternative models [Eq. (7)] with $n_{g,j} = n_{g,\max}$ because the case of $n_{g,j} < n_{g,\max}$ is somehow contained in Eq. (7) by ∇_j having some zero components. But this is not possible because practically one would never get an estimate of ∇_j with zero components. Consequently, one would have to discard $n_{g,\max}$ observations, i.e., the worst case scenario would always apply.

The problem is now to decide, which model applies to the observations: the null model [Eq. (1)], requiring the observations to be free of outliers, or the alternative model [Eq. (6)] or one of the alternative models [Eq. (7)]. A problem of such a type is often called *model selection problem*.

Outlier Model Selection by Multiple Hypothesis Test

Individual Test

The common approach to model selection is performing a statistical hypothesis test. The standard procedure is listed as follows:

1. The null hypothesis H_0 supposes that the observations correspond to the null model [Eq. (1)].
2. The alternative hypothesis H_A supposes that the observations correspond to the alternative model [Eq. (6) or Eq. (7)].
3. A test statistic T is invoked, that is a function of the observations l, which is able to separate between the two hypotheses. Hopefully, it is even best able to do this, but such T is often not possible to derive rigorously. Typically, if H_0 is false then T tends to assume extreme values.
4. A probability of Type I decision error α (probability that a true H_0 is rejected) is defined, typically α=0.10, 0.05 or 0.01.
5. The probability distribution of T is derived under the condition that H_0 is true, and the critical value c is obtained as $(1 - \alpha)$ quantile of this distribution.
6. The critical value c is compared to the value of $T(l)$ computed from the actual observations l. If $T(l)$ exceeds c, then H_0 is rejected. In this case one must assume the observations to contain outliers; otherwise, the null model is used to process the observations.

For Eq. (6) the optimal test statistics were derived by *Baarda (1968)* and *Pope (1972)* as

$$T_{\text{prio}} = \frac{\widehat{\boldsymbol{\nabla}}^T \boldsymbol{Q}_{\widehat{\nabla}\widehat{\nabla}}^{-1} \widehat{\boldsymbol{\nabla}}}{n_g \sigma^2} = \frac{\boldsymbol{v}^T \boldsymbol{P}\boldsymbol{C}(\boldsymbol{C}^T \boldsymbol{P}\boldsymbol{Q}_{vv}\boldsymbol{P}\boldsymbol{C})^{-1}\boldsymbol{C}^T \boldsymbol{P}\boldsymbol{v}}{n_g \sigma^2} \tag{9a}$$

$$T_{\text{post}} = \frac{\widehat{\boldsymbol{\nabla}}^T \boldsymbol{Q}_{\widehat{\nabla}\widehat{\nabla}}^{-1} \widehat{\boldsymbol{\nabla}}}{n_g \hat{\sigma}'^2} = \frac{\boldsymbol{v}^T \boldsymbol{P}\boldsymbol{C}(\boldsymbol{C}^T \boldsymbol{P}\boldsymbol{Q}_{vv}\boldsymbol{P}\boldsymbol{C})^{-1}\boldsymbol{C}^T \boldsymbol{P}\boldsymbol{v}}{n_g \hat{\sigma}'^2} \tag{9b}$$

where Eqs. (9a) and (9b) correspond to the cases that the a priori variance factor σ^2 is known and unknown, respectively; $\boldsymbol{C}$ is supposed to guarantee regularity of $\boldsymbol{C}^T \boldsymbol{P}\boldsymbol{Q}_{vv}\boldsymbol{P}\boldsymbol{C}$; $\widehat{\boldsymbol{\nabla}}$ denotes the least squares estimate of $\boldsymbol{\nabla}$ in Model Eq. (6); $\boldsymbol{Q}_{\widehat{\nabla}\widehat{\nabla}}$ is the corresponding cofactor matrix; $\boldsymbol{v}$ and $\boldsymbol{Q}_{vv}$ are computed from Eqs. (3) and (5) as before; and $\hat{\sigma}'^2$ denotes the *external* estimate of the variance factor σ^2, i.e., excluding the outlier-suspected observations *(Heck 1981)*:

$$\hat{\sigma}'^2 = \frac{\overline{\boldsymbol{v}}^T \boldsymbol{P} \overline{\boldsymbol{v}}}{n-u-n_g} = \frac{\boldsymbol{v}^T \boldsymbol{P} \boldsymbol{v} - \widehat{\boldsymbol{V}}^T \boldsymbol{Q}_{\widehat{V}\widehat{V}}^{-1} \widehat{\boldsymbol{V}}}{n-u-n_g} \tag{10}$$

The advantage of this estimate is that under the condition that H_0 is true, T_{post} is known to follow an F distribution, in the same way as T_{prio}:

$$T_{prio}|H_0 \sim F_{n_g,\infty} \tag{11a}$$

$$T_{\text{post}}|H_0 \sim F_{n_g,n-u-n_g} \tag{11b}$$

Here, $F_{n,m}$ denotes the central F distribution with n and m degrees of freedom. Using the common *internal* estimate of σ^2 in Eq. (9b) one would get the more cumbersome τ-distribution *(Pope 1972)*. Note that if one uses the second expressions for T_{prio} in Eq. (9a) or T_{post} in Eq. (9b), it is not necessary to solve Eq. (6) explicitly. Under the alternative hypothesis the test statistics follow a non-central F distribution, but the non-centrality parameter is not known.

In the case $n_g = 1$ (single outlier) one must set

$$\boldsymbol{C} = (0\ 0\ 0\ \cdots 0\ 1\ 0\ \cdots 0\ 0)^T \tag{12}$$

and if additionally $\boldsymbol{P}$ is a diagonal matrix, T_{prio} in Eq. (9a) and T_{post} in Eq. (9b) equal the squares of the well-known normalized and externally studentized residuals of the j th observation, respectively, when j is the index of the element 1 of C *(Teunissen 2000, p. 37)*:

$$T_{\text{prio}} = \frac{v_j^2}{\sigma^2 q_{vv,j}} \tag{13a}$$

$$T_{\text{post}} = \frac{v_j^2}{\hat{\sigma}'^2 q_{vv,j}} \tag{13b}$$

where $q_{vv,j}$ denotes the j th diagonal element of Q_{vv} in Eq. (5).

Generally, T_{prio} in Eq. (9a) and T_{post} in Eq. (9b) are compared to the computed critical values of the tests

$$c_{\text{prio}} = F_F^{-1}(1-\alpha, n_g, \infty) \tag{14a}$$

$$c_{\text{post}} = F_F^{-1}(1-\alpha, n_g, n-u-n_g) \tag{14b}$$

where $F_F^{-1}(\cdot, n, m)$ denotes the inverse probability function (quantile function) of the F distribution with n and m degrees of freedom.

Multiple Test

Practically, there is often no knowledge of the number of suspected outliers n_g and of the affected outlying observations. Consequently, one has to deal with multiple alternative hypotheses $H_{A,j}, j = 1, \ldots, m$ in parallel. This requires a multiple hypothesis test, which, in principle, is a set of m standard hypotheses tests H_0 vs. $H_{A,j}$ with test statistics T_j and critical values $c_j, j = 1, \ldots, m$. If in any of the m tests c_j is exceeded by T_j, then H_0 is rejected:

$$\text{Reject } H_0 \text{ if } T_1 > c_1 \text{ or } T_2 > c_2 \text{ or } \ldots \text{ or } T_m > c_m \tag{15}$$

In this case one must assume the observations to contain outliers and reject them. Otherwise, the null model [Eq. (1)] is used to compute estimates of the desired quantities.

The test statistics for the alternative models [Eq. (7)] are derived from Eqs. (9a) and (9b) as

$$T_{\text{prio},j} = \frac{\widehat{\boldsymbol{\nabla}}_j^T \boldsymbol{Q}_{\widehat{\nu}\widehat{\nu},j}^{-1} \widehat{\boldsymbol{\nabla}}_j}{n_{g,j}\sigma^2} = \frac{\boldsymbol{v}^T \boldsymbol{P}\boldsymbol{C}_j (\boldsymbol{C}_j{}^T \boldsymbol{P}\boldsymbol{Q}_{vv}\boldsymbol{P}\boldsymbol{C}_j)^{-1} \boldsymbol{C}_j{}^T \boldsymbol{P}\boldsymbol{v}}{n_{g,j}\sigma^2} \tag{16a}$$

$$T_{\text{post},j} = \frac{\widehat{\boldsymbol{\nabla}}_j^T \boldsymbol{Q}_{\widehat{\nu}\widehat{\nu},j}^{-1} \widehat{\boldsymbol{\nabla}}_j}{n_{g,j}\hat{\sigma}_j'^2} = \frac{\boldsymbol{v}^T \boldsymbol{P}\boldsymbol{C}_j (\boldsymbol{C}_j{}^T \boldsymbol{P}\boldsymbol{Q}_{vv}\boldsymbol{P}\boldsymbol{C}_j)^{-1} \boldsymbol{C}_j{}^T \boldsymbol{P}\boldsymbol{v}}{n_{g,j}\hat{\sigma}_j'^2} \tag{16b}$$

$$j = 1, \ldots, m$$

The external estimate $\hat{\sigma}_j'^2$ of σ^2 depends on j via

$$\hat{\sigma}_j'^2 = \frac{\boldsymbol{v}^T \boldsymbol{P}\boldsymbol{v} - \widehat{\boldsymbol{\nabla}}_j^T \boldsymbol{Q}_{\widehat{\nu}\widehat{\nu},j}^{-1} \widehat{\boldsymbol{\nabla}}_j}{n - u - n_{g,j}} \tag{17}$$

Roughly speaking, the risk of rejecting a true H_0 in the multiple hypothesis test is now m-fold: The undesired random event "reject a true H_0" can occur in any of the m tests. Let the probability of rejecting a true H_0 in test j be α_j (the so-called *experimentwise error rate*) and let $\alpha_j \ll 1$. Furthermore, assume the random events "reject a true H_0 in test j" to be approximately statistically independent (see discussion below). Then the total probability of rejecting a true H_0 in the multiple hypothesis test (the so-called *familywise error rate*) is

$$\alpha \approx 1 - \prod_{j=1}^{m} (1 - \alpha_j) \approx \sum_{j=1}^{m} \alpha_j \tag{18}$$

The common approach is to simply choose

$$\alpha_j := \alpha / m \tag{19}$$

which is called the Bonferroni equation *(Abdi 2007)*.

Unfortunately, the test statistics [Eqs. (16a) and (16b)] and consequently the random events "reject a true H_0 in test j" are statistically dependent. In the case $n_g = 1$ (single outlier) the dependency is caused by $\boldsymbol{Q}_{vv}$ being a nondiagonal matrix. *Lehmann (2012)* has shown how to improve this situation by using a Monte-Carlo type approach. In the other cases the dependency is even stronger because in contrast to Eqs. (13a) and (13b) the test statistics Eqs. (16a) and (16b) involve many or all residuals v_k simultaneously. As a remedy one can propose to weaken Eq. (19) by choosing a tuning parameter α' such that

$$\alpha_j := \alpha' \tag{20}$$

which makes sense even without the required independence. By Eq. (20) the total risk of rejecting a true H_0 is portioned equally to the individual tests, although they do not exactly add up to α. Nonetheless, α' should be chosen smaller for m getting larger. Next, a method to avoid the explicit choice of α' is proposed.

Some discussions on correlation issues among outlier test statistics can be found in *Wang et al. (2012)* and *Wang and Knight (2012)*.

The critical values are now computed as

$$c_{\text{prio},j} = F_F^{-1}(1 - \alpha', n_{g,j}, \infty) \tag{21a}$$

$$c_{\text{post},j} = F_F^{-1}(1 - \alpha', n_{g,j}, n - u - n_{g,j}) \tag{21b}$$

$$j = 1, \ldots, m$$

Note that $c_{\text{prio},j}, c_{\text{post},j}$ depend on j only via $n_{g,j}$. This shows that for the subset of individual tests with identical $n_{g,j}$ one can also get identical critical values $c_j = c$ in Eqs. (21a) and (21b). Thus, it is sufficient to compare them only with the maximum of the related test statistics within this subset

$$\text{Reject } H_0 \text{ if } \max_j T_j > c \tag{22}$$

This procedure is generally recommended in geodetic outlier detection, e.g., by *Knight et al. (2010)*.

In the simple case $n_g = 1$, one gets from Eqs. (13a) and (13b) and Eq. (22) the well-known tests of the (square of the) normalized and studentized residuals, respectively:

$$\text{Reject } H_0 \text{ if } \max_{j=1,\ldots,n} \frac{v_j^2}{\sigma^2 q_{vv,j}} > c_{prio} \tag{23a}$$

$$\text{Reject } H_0 \text{ if } \max_{j=1,\ldots,n} \frac{v_j^2}{\hat{\sigma}_j'^2 q_{vv,j}} > c_{post} \tag{23b}$$

These are the standard individual tests for outliers in geodetic adjustment.

However, if in a multiple test there are individual tests with different $n_{g,j}$ then Eq. (20) is not mandatory and Eq. (22) does not apply anyway. One could argue that the risk of rejecting a true H_0 should be particularly small when $n_{g,j}$ is large because practically one should especially avoid running the risk of discarding a large number of good observations. On the other hand, if there are indeed $n_{g,j}$ gross errors then they must be rather extreme to be rejected. The problem of finding a best tradeoff can be solved by Monte-Carlo based data snooping as suggested by *Lehmann and Scheffler (2011)*, but this issue will not be brought up here. For the sake of simplicity Eq. (20) is used here instead.

Selection of the Alternative Model: p-Value Approach

If n_g is fixed and known then Eq. (22) applies, and it is intuitively clear, which alternative model must be used: It is Eq. (7) with the index $j = j_{\max}$, for which the maximum in Eq. (22) is attained. This makes sense from the following point of view. If one would decrease α' in (20), whose value is always to some degree debatable, and in this way increase $c_j = c$ in Eqs. (21a) and (21b) beyond the second largest T_j in Eq. (22) then one would end up with H_0 rejected only in test $j_{\max}$, e.g., if $n_g = 1$ is known then the $j_{\max}$ th observation is rejected as the single outlier in Eqs. (23a) and (23b). (Nonetheless, this apparently simple truism does not strictly follow from the theory of statistical hypothesis testing: one rejects H_0 in favor of H_A, but this time there are multiple alternative hypotheses, and it is immediately unknown which is the favorable one.)

Unfortunately, this reasonable argument applied to Eq. (22) does not carry over to the general case Eq. (15), in which no maximum is taken. Rather, it is an exception that in Eq. (15) H_0 is rejected in exactly one test j, such that the observations selected by $\boldsymbol{C}_j$ are clearly identified as outliers. If H_0 is rejected in many tests then one faces the problem of which observations should be rejected as outliers. This shows that the detection of multiple outliers with n_g unknown by a hypothesis test is not straightforward.

It is suggested to follow the earlier line of reasoning also in Eq. (15): if required, one decreases α' up to that point, where only one H_0 in Eq. (15) is still rejected. This point is found as follows. To each T_j one assigns the so-called p-value. Although not so popular in geodesy, this is a well-known statistical quantity. It denotes the imaginary error rate $\alpha' = p_j$, at which $T_j = c_j$ would hold, i.e., the decision of the individual test is balancing on a knife's edge.

Hence, Eqs. (21a) and (21b) have to be solved for $\alpha' = p_j$ getting

$$p_{\text{prio},j} = 1 - F_F\left(T_{\text{prio},j}, n_{g,j}, \infty\right) \tag{24a}$$

$$p_{post,j} = 1 - F_F\left(T_{post,j}, n_{g,j}, n - u - n_{g,j}\right) \tag{24b}$$

The significance level α', at which exactly one H_0 in Eq. (15) is rejected, is between the smallest and the second smallest p-value [Eqs. (24a) and (24b)], and the rejection is triggered by $T_j > c_j$, with j being the index of the smallest p-value.

As a result one obtains the rule

$$\text{Reject } H_0 \text{ if } \min_{j=1,\ldots,m} p_j < \alpha' \tag{25a}$$

and accept $H_{A,j}$ with j denoting the index, at which the minimum in Eq. (25a) is attained. Hence, the observations selected by C_j are rejected as outliers. In practice, it is recommended to exchange α' and p_j by their logarithmic representations to avoid numerical underflow in case of large values of $T_{\text{prio},j}$ or $T_{\text{post},j}$, and Eq. (25a) reads

$$\text{Reject } H_0 \text{ if } \min_{j=1,\ldots,m} \log p_j < \log \alpha' \tag{25b}$$

Using a Global Model Test

The value of the experimentwise error rate α' in Eq. (20) decides, if H_0 is rejected or not via Eq. (25a) or Eq. (25b). It does not decide which $H_{A,j}$ is accepted. Since α' should become smaller as m gets larger, it might be difficult to choose a suitable value. In this respect a so-called *global model test*, also known as *overall model test*, is welcome. Unfortunately, this test is only possible, if σ^2 is known.

In data snooping it is often recommended to start the outlier detection procedure with such a global model test. This step is often called *detection*. The test statistic of this test is

$$T_{\text{global}} = \frac{\boldsymbol{v}^T \boldsymbol{P} \boldsymbol{v}}{(n-u)\sigma^2} \tag{26}$$

Its distribution is

$$T_{\text{global}} | H_0 \sim F_{n-u,\infty} \tag{27}$$

(equivalent to a χ^2 distribution with $n - u$ degrees of freedom), which together with an error rate α gives rise to a critical value

$$c_{\text{global}} = F_F^{-1}(1 - \alpha, n - u, \infty) \tag{28}$$

The rationale of this test is that its test statistic is optimal for a large number of alternative hypotheses $H_{A,j}$, namely those with $n_{g,j} = n - u$ *(Teunissen 2000)*:

$$T_{\text{prio},j} = \frac{\widehat{\boldsymbol{\nabla}}_j^T \boldsymbol{Q}_{\widehat{\boldsymbol{\nabla}}\widehat{\boldsymbol{\nabla}},j}^{-1} \widehat{\boldsymbol{\nabla}}_j}{n_{g,j}\sigma^2} = \frac{\boldsymbol{v}^T \boldsymbol{P} \boldsymbol{v}}{(n-u)\sigma^2} = T_{\text{global}} = \text{const for all } j \tag{29}$$

If $T_{\text{global}} > c_{\text{global}}$, then H_0 is rejected and it remains to identify the outliers. This second step of data snooping is, consequently, called *identification*. It is a multiple test with all possible alternative hypotheses as proposed earlier.

It can be difficult to suitably adapt the error rates of the detection and the identification step such that any detected outliers are really identified *(Hahn et al. 1989, 1991)*. Here it is proposed to use the p-value approach [Eqs. (25a) and (25b)] in such a way that one accepts $H_{A,j}$ with j denoting the index, at which the minimum in Eqs. (25a) and (25b) is attained, regardless of α'. The advantage is that there is no need to specify α', but only α in Eq. (28). The latter has a much clearer meaning: it is the familywise error rate of the multiple test, because if H_0 holds true then it will be rejected exactly with probability α. Also, α must not depend on m, like α' does.

Once it has been decided by the global test that H_0 must be rejected, the minimum p-value in Eqs. (25a) and (25b) identifies the outliers. Therefore, the number of identified outliers does not depend on α! If α is chosen very large then outliers are always detected, but no larger number of them.

Example: Fit of a Straight Line

It is useful to illustrate the theoretical considerations with a simple practical example. The straight line fit with n equidistant data points was chosen. This model is used in various fields of engineering sciences by

- Extracting a linear trend from a time series;
- Fitting a linear calibration function for calibration of measuring devices; and
- Surveying points on a spatial straight line, which deviate from a straight line caused by observation errors.

Nonetheless, it is used here merely for illustration of the theory. A truly practical application is given later.

With error-free abscissae $1, \ldots, n$ the observations in Eq. (1) read

$$l_i = x_1 + i \cdot x_2 + e_i, \qquad i = 1, \ldots, n$$

One can see that $u = 2$. Let $n = 10$, $\boldsymbol{P} = \boldsymbol{I}$. Furthermore, let $\alpha = 0.01$ and σ^2 be known. From practical considerations let $n_{g,\max} = 3$.

Start with a global test. If

$$T_{\text{global}} = \frac{v_1^2 + \cdots + v_{10}^2}{(10-2)\sigma^2} > F_F^{-1}(1-\alpha, 10-2, \infty) = 2.51 \tag{30}$$

holds then outliers are detected und must be identified by the multiple test as follows:

1. Test of $n_g = 1$ (single outlier in unknown place): Here one can simply identify the outlier by Eq. (23a), i.e., by the index, where the maximum in

$$T_1 = \frac{1}{\sigma^2}\max\left(\frac{v_1^2}{0.655}, \frac{v_2^2}{0.752}, \frac{v_3^2}{0.824}, \frac{v_4^2}{0.873}, \frac{v_5^2}{0.897}, \frac{v_6^2}{0.897}, \ldots, \frac{v_{10}^2}{0.655}\right) \tag{31}$$

is attained. In the denominators of Eq. (31) the diagonal elements of Q_{vv} are found in Eq. (5).

Outer residuals v_1, v_{10} do not need to be so large in magnitude to be identified as an outlier. The rational of this is that a gross error in the outer observation would result in a residual of only a smaller size caused by the leverage effect. An observation of such a kind is called *leverage observation* *(Rousseeuw and Leroy 1987)*.

2. Test of $n_g = 2$ (pair of outliers in unknown place): $\binom{10}{2} = 45$ pairs of observations can be built. The more general form [Eq. (22)] of the multiple test is used: The index, where

$$T_2 = \frac{1}{2\sigma^2}\max\left(\begin{pmatrix}v_1\\v_2\end{pmatrix}^T \begin{pmatrix}0.655 & -0.291\\-0.291 & 0.752\end{pmatrix}^{-1} \begin{pmatrix}v_1\\v_2\end{pmatrix}, \ldots, \begin{pmatrix}v_9\\v_{10}\end{pmatrix}^T \begin{pmatrix}0.752 & -0.291\\-0.291 & 0.655\end{pmatrix}^{-1} \begin{pmatrix}v_9\\v_{10}\end{pmatrix}\right) \tag{32}$$

is attained, identifies the pair of outliers. The 2×2 matrices in Eq. (32) are submatrices of $\boldsymbol{Q}_{vv}$ in Eq. (5).

3. Test of $n_g = 3$ (triplet of outliers in unknown place): $\binom{10}{3} = 120$ triplets of observations can be built. The identification is by

$$T_3 = \frac{1}{3\sigma^2}\max\left(\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}^T \begin{pmatrix}0.655 & -0.291 & -0.236\\-0.291 & 0.752 & -0.206\\-0.236 & -0.206 & 0.824\end{pmatrix}^{-1} \begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}, \ldots\right) \tag{33}$$

4. Test of $n_{g,\max} = 3$ (single outlier or pair or triplet of outliers in unknown place): $m = 10 + 45 + 120 = 175$. The p-value approach from Eqs. (25a) and (25b) is used

$$p_1 = 1 - F_F(T_1, 1, \infty) \qquad (34)$$
$$p_2 = 1 - F_F(T_2, 2, \infty)$$
$$p_3 = 1 - F_F(T_3, 3, \infty)$$

The index, where

$$\min(p_1, p_2, p_3) \qquad (35)$$

is attained, identifies the number of outliers.

For a numerical example, let $\boldsymbol{l} = (-5,0,0,0,0,0,0,0,3,5)^T\sigma$. The residuals become $\boldsymbol{v} = (2.27, -2.05, -1.38, -0.71, -0.04, 0.64, 1.31, 1.98, -0.35, -1.67)^T\sigma$.

The global test [Eq. (30)] detects outliers with $T_{\text{global}} = 2.60$.

1. The first observation is identified as the detected outlier by Eq. (31) with $T_1 = 7.89$.

2. The first and the tenth observation are identified as a pair of outliers by Eq. (32) with $T_2 = 7.76$.

3. The first, ninth and tenth observation are identified as a triplet of outliers by Eq. (33) with $T_3 = 6.92$.

4. The p-values

$$p_1 = 1 - F_F(7.89,1,\infty) = 0.00497$$
$$p_2 = 1 - F_F(7.76,2,\infty) = 0.00043$$
$$p_3 = 1 - F_F(6.92,3,\infty) = 0.00012$$

attain their minimum at $n_{g,j} = 3$. Therefore, the first, ninth, and tenth observation are finally identified as a triplet of outliers by Eq. (35).

Outlier Model Selection by Consecutive Hypothesis Tests

The easiest and also most common way of applying data snooping to the detection of multiple outliers is consecutive detection, identification and rejection of single outliers. Here an iterative procedure is executed, assuming $n_g = 1$ in each iteration step and continuing until no further outlier is detected. If in Eq. (23a) or Eq. (23b) the maximum is attained at index j and exceeds c_{prio} or c_{post} then the j th observation is rejected as a single outlier. Then one would go on with a test of the rest of the observations for a further single outlier. From a rigorous point of view this approach is invalid because in the previous test it has been assumed that $n_g = 1$! One should not both discard this assumption and retain the result of the test based on it.

However, as shown below, also the proposed multiple test has weaknesses, and it could turn out that they are practically more severe.

Nonetheless, the test in each iteration also is a multiple test with $m = n$ alternative hypotheses. But they are stochastically less dependent, such that in each iteration Eq. (18) is often a good approximation.

Unlike in the p-value approach, the number of outliers finally detected here depends on the choice of the error rate α: A large α means that many outliers are detected, perhaps spuriously, and vice versa. Fortunately, there are some reports of long experiences with the choice of α, e.g., in *Mierlo (1983)*. A global test is not required, but often used if σ^2 is known.

Example: Fit of a Straight Line, Consecutive Test

After discarding the observation, where the maximum in Eq. (31) is attained, the procedure is repeated with $n := n - 1$, starting with a global test. In the previous numerical example the first observation has been identified as the detected single outlier by Eq. (31). Repeating the global test with the remaining nine observations, again with $\alpha = 0.01$, does not reject H_0. Therefore, no further outlier is detected. (If the individual tests are done, nonetheless, the tenth observation would be identified as the next outlier.)

This disagreement with the earlier result shows the dilemma of the multiple hypothesis tests.

Outlier Model Selection by Information Criteria

AICc

Among all $m + 1$ models [Eqs. (1) and (7)] under consideration the multiple hypotheses test privileges the null model [Eq. (1)] in such a way that it is tested against each alternative model [Eq. (7)]. This is done because for each alternative model a different test statistic [Eq. (16a) or Eq. (16b)] is optimal. The distribution of the test statistic must be known, and this is only fulfilled for the null model. (For the alternative models the test statistic follows a non-central F distribution, but the non-centrality parameter is not known.)

From information theory there are different approaches of model selection based on information criteria. The oldest and best known is the AIC *(Akaike 1974)*:

$$AIC = 2k - 2\log L(\widehat{\boldsymbol{\theta}};\, \boldsymbol{l}) \tag{36}$$

where L denotes the likelihood function of the model, which is maximized by the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ of the k vector of parameters $\boldsymbol{\theta}$ with respect to the n vector of observations $\boldsymbol{l}$. The AIC states: among all models under consideration the one with the least AIC is to be selected. It has high likelihood and at the same time few k parameters. If different models give AIC values very close to the minimum, it is generally recommended to avoid the selection, if possible.

A corrected version of AIC is

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{37}$$

which is supposed to work better for small sample sizes. If n is small or k is large then AICc is strongly recommended rather than AIC. Because AICc converges to AIC as n gets large, AICc generally should be used regardless *(Burnham and Anderson 2004)*. Note that $\boldsymbol{\theta}$ should comprise all parameters, i.e., not only those in x, ∇, but also σ^2, if it is unknown.

For Eq. (1) in combination with Eq. (2) the AICc assumes the form

$$AICc_{\text{prio},0} = 2u + \frac{2u(u+1)}{n-u-1} + \frac{\boldsymbol{v}^T\boldsymbol{P}\boldsymbol{v}}{\sigma^2} + C_{\text{prio}} \tag{38a}$$

$$AICc_{\text{post},0} = 2(u+1) + \frac{2(u+1)(u+2)}{n-u-2} + n \cdot \log\left(\frac{\boldsymbol{v}^T\boldsymbol{P}\boldsymbol{v}}{n}\right) + C_{\text{post}} \tag{38b}$$

for known and unknown variance factor σ^2, respectively. C_{prio} and C_{post} denote constant terms, neither depending on u, nor on v, such that they are of no interest when selecting the minimum. A derivation of Eq. (38b) is given in the Appendix. Here, $AICc$ measures fit of the data via the third term and penalizes models with too many parameters via the first and second terms. Via the third term in Eq. (38a) there is a relationship between $AICc_{\text{prio},0}$ and T_{global} in Eq. (26)

$$AICc_{\text{prio},0} = 2u + \frac{2u(u+1)}{n-u-1} + (n-u)T_{\text{global}} + C_{\text{prio}} \tag{38c}$$

From this equation it can be concluded that, if the global test is positive then AICc (and also AIC) for the null model is small. This makes it likely that this model is selected, such that the observations are declared free of outliers.

Analogously, for Eq. (7)

$$AICc_{\text{prio},j} = 2(u+n_{g,j}) + \frac{2(u+n_{g,j})(u+n_{g,j}+1)}{n-u-n_{g,j}-1} + \frac{\boldsymbol{v}^T\boldsymbol{P}\boldsymbol{v} - \widehat{\boldsymbol{\nabla}}_j^T\boldsymbol{Q}_{\widehat{\nabla}\widehat{\nabla},j}^{-1}\widehat{\boldsymbol{\nabla}}_j}{\sigma^2} + C_{\text{prio}} \tag{39a}$$

$$AICc_{\text{post},j} = 2(u+n_{g,j}+1) + \frac{2(u+n_{g,j}+1)(u+n_{g,j}+2)}{n-u-n_{g,j}-2} + n\cdot\log\left(\frac{\boldsymbol{v}^T\boldsymbol{P}\boldsymbol{v} - \widehat{\boldsymbol{\nabla}}_j^T\boldsymbol{Q}_{\widehat{\nabla}\widehat{\nabla},j}^{-1}\widehat{\boldsymbol{\nabla}}_j}{n}\right) + C_{\text{post}} \tag{39b}$$

Comparing Eqs. (38a) and (39a)

$$AICc_{\text{prio},j} = AICc_{\text{prio},0} + \text{terms}(n_{g,j}) - n_{g,j}T_{\text{prio},j} \tag{40}$$

Therefore, in the set of alternative models [Eq. (7)] having the same $n_{g,j}$, the one with maximum $T_{prio,j}$ has minimum $AICc_{\text{prio},j}$. This proves that for fixed $n_{g,j}$ the model selection by $T_{\text{prio},j}$ and $AICc_{\text{prio},j}$ yields identical results. Only the preference of more or less outliers is different. In contrast, the model selection by $T_{\text{post},j}$ and $AICc_{\text{post},j}$ can be different even for fixed $n_{g,j}$.

Example: Fit of a Straight Line Continued

The computation of the straight line fit is resumed. As before, it is restricted to the case of known variance factor σ^2. From Eq. (38a)

$$AICc_0 - C_{\text{prio}} = 4 + \frac{12}{7} + \frac{\boldsymbol{v}^T\boldsymbol{v}}{\sigma^2} = 5.7 + (n-u)T_{\text{global}} = 26.5$$

When alternative models are considered, Eq. (39a) needs to be evaluated

1. Test of $n_g = 1$ (single outlier in unknown place): The candidate model is the one that identifies the first observation as an outlier

$$AICc_1 - C_{\text{prio}} = 6 + \frac{24}{6} + \frac{\boldsymbol{v}^T\boldsymbol{v} - \frac{v_1^2}{0.655}}{\sigma^2} = AICc_0 - C_{\text{prio}} + 4.3 - T_1 = 22.9$$

2. Test of $n_g = 2$ (pair of outliers in unknown place): The candidate model is the one that identifies the first and the tenth observation as outliers

$$AICc_2 - C_{\text{prio}} = 8 + \frac{40}{5} + \frac{1}{\sigma^2}\left(\boldsymbol{v}^T\boldsymbol{v} - \begin{pmatrix} v_1 \\ v_{10} \end{pmatrix}^T \begin{pmatrix} 0.655 & 0.145 \\ 0.145 & 0.655 \end{pmatrix}^{-1} \begin{pmatrix} v_1 \\ v_{10} \end{pmatrix}\right)$$

$$= AICc_0 - C_{\text{prio}} + 10.3 - 2T_2 = 21.2$$

3. Test of $n_g = 3$ (triplet of outliers in unknown place): The candidate model is the one that identifies the first, ninth, and tenth observations as outliers

$$AICc_3 - C_{\text{prio}} = 10 + \frac{60}{4} + \frac{1}{\sigma^2}\left(\boldsymbol{v}^T\boldsymbol{v} - \begin{pmatrix} v_1 \\ v_9 \\ v_{10} \end{pmatrix}^T \begin{pmatrix} 0.655 & 0.091 & 0.145 \\ 0.091 & 0.752 & -0.291 \\ 0.145 & -0.291 & 0.655 \end{pmatrix}^{-1} \begin{pmatrix} v_1 \\ v_9 \\ v_{10} \end{pmatrix}\right)$$

$$= AICc_0 - C_{\text{prio}} + 19.3 - 3T_3 = 25.0$$

4. Test of $n_{g,\max} = 3$ (single outlier or pair or triplet of outliers in unknown place): The minimum of AICc is attained at $n_g = 2$. Consequently, the first and the tenth observation are identified as a pair of outliers.

For the sake of completeness one mentions that $AICc_4 - C_{\text{prio}} = 40$ and the minimum of AIC without correction is attained at $n_g = 3$, which would indicate that the ninth observation is also an outlier.

Using the Alternative Models with Suspected Outliers Discarded

It is well known that one can benefit from the special structure of C by computing the solution of Eq. (7) in such a way that the outliers and the extra bias parameters ∇ are discarded and standard least squares estimation with the remaining observations is done. Neither the redundancy of the Eq. (7), nor the test statistics [Eqs. (16a), (16b) and (26)] or their distributions would change in any way. Therefore, when doing outlier detection by hypothesis tests in the framework of data snooping, this approach is practically equivalent.

Surprisingly, this is not so with AIC and AICc. In Eqs. (39a) and (39b) it is obvious that AICc depends on the number of observations and parameters not only via the redundancy, as the test statistics [Eqs. (16a), (16b) and (26)] do. This is a peculiar phenomenon not yet investigated in the literature. As shown in Table 1, in the example of the straight line fit, there may even be a different optimal model: with extra bias parameters ∇ the first and the tenth observation were identified as a pair of outliers, but with suspected outliers discarded, the first, ninth, and tenth observation are identified as a triplet of outliers (Table 1).

Table 1. Model Selection by AICc for the Example of the Straight Line Fit

	With extra bias parameters ∇			With suspected outliers discarded		
n_g	k	n	$AICc_{n_g} - C_{\text{prio}}$	k	n	$AICc_{n_g} - C_{\text{prio}}$
0	2	10	26.5	2	10	26.5
1	3	10	22.9	2	9	18.9
2	4	10	**Min = 21.3**	2	8	11.6
3	5	10	25.0	2	7	**Min = 7.0**
4	6	10	40	2	6	8.0

Note: Bold numbers indicate minima of AICc.

Advanced Aspects

Linearization Errors

This paper started from a linear or linearized Gauss-Markov model, whereas the observation equations of many geodetic problems like 2D or 3D geodetic networks are inherently nonlinear. This has two consequences:

1. The test statistics [Eqs. (9a) and (9b)] and all derived forms like Eqs. (16a) and (16b) are so-called likelihood ratios and as such take advantage of being the uniformly most powerful invariant (UMPI) test

statistics, but only in linear models *(Kargoll 2007)*. In a nonlinear model there could be more powerful test statistics, but it is not possible to derive them. It is even very complicated to prove this property because the statistical power of such a test statistic would involve a multidimensional integral, which can generally only be solved numerically. Lacking results of such a computation, one can only hope that the linearization error does not spoil the UMPI property, such that Eqs. (9a) and (9b) are also reasonable test statistics in a nonlinear model. A different perspective of the same problem shows that only for linear models the matrix $\boldsymbol{A}$ in Eq. (1) is constant. Otherwise it is computed from the observations, also from those that are, perhaps, outliers.

2. The distributional results like Eqs. (11a) and (11b) are exactly valid only in linear models. Otherwise, the distribution would differ slightly from an F-distribution. The critical values cannot exactly be taken from a statistical lookup table or computed by a standard statistical library function. A rigorous computation would require a multidimensional numerical integration. The necessary procedure has been worked out by *Lehmann (2012)*, but is used to study a different aspect: the neglected statistical dependencies between the test statistics in the multiple test. If an F-distribution is used, nonetheless, the outlier detection is performed with an incorrect significance level. The extend, to which this lack of rigor spoils the result, depends on the degree of nonlinearity of the problem. In some cases it can be drastic. Consider *Lehmann (2014)*, in which it is tried to detect a sinusoidal oscillation of unknown frequency in a time series, which is a seriously nonlinear problem. A likelihood ratio test also is applied there, but to the original nonlinear problem. Then the distribution of the test statistic is investigated by Monte Carlo integration. It was shown that the test statistic is not even approximately F-distributed. For outlier tests in geodetic networks no similar investigations exist, but from investigations of the parameter estimation it is known that the degree of nonlinearity increases with the size of the network. Now look at the information criteria. Here AICc in Eq. (37) is used as a pure definition. If the likelihood function $L(\widehat{\boldsymbol{\theta}};\ \boldsymbol{l})$ in Eq. (36), which is referred to the nonlinear model, is maximized by iteration, and the iteration converges to the global maximum, information criteria in Eq. (38a) or Eq. (38b) are computed without any linearization errors (see also the Appendix). The question would be whether Eq. (36) is a sufficient definition for both the linear and the nonlinear models. In the literature the series critics yet go in a different direction: The question is whether the balance between model complexity and goodness of fit is optimal both for small and large sets of observations. In summary, linearization is a subject insufficiently investigated for both multiple and single outlier detection.

Computational Costs

It became obvious that for multiple outlier detection both the multiple hypothesis tests and the information criteria approach are computationally expensive. The hypothesis tests involve the computation of a large number of test statistics [Eq. (16a) or Eq. (16b)]. The expensive step is the inversion of $\boldsymbol{C}_j^T\boldsymbol{P}\boldsymbol{Q}_{vv}\boldsymbol{P}\boldsymbol{C}_j$, a symmetric $n_g \times n_g$ matrix, which is sometimes not sparse. The computational complexity of such an operation is of the order $O(n_g^3)$. Compared with the $u \times u$ matrix to be inverted in Eq. (5), such a matrix is of relatively small size. However, if $n_{g,\max}$ is increasing, there is a rapidly increasing number of such matrices of increasing maximal dimension. For example, for $n = 100$ observations, which may contain up to five outliers, there are 100 matrices of dimension 1×1; 4,950 matrices of dimension 2×2; 161,700 matrices of dimension 3×3; $3.9 \cdot 10^6$ matrices of dimension 4×4 and $75 \cdot 10^6$ matrices of dimension 5×5. If $\boldsymbol{Q}_{vv}$ is not sparse, this would require about 10^{10} floating point operations, but there are great opportunities for improvement: Note that all $n_g \times n_g$ matrices are partially identical to a previous $(n_g - 1) \times (n_g - 1)$ matrix, except an added row and an added column, with one being the transposition of the other. This allows one to profit by fast block matrix inversion *(Koch 1999, p. 33)*. The essential operation is now only a matrix vector multiplication of the existing inverse with the extending column vector. This reduces the computational costs immensely, because the computational complexity of such an operation is only of the order $O(n_g^2)$.

From Eq. (40) the computation of the $AICc_{\text{prio},j}$ of an alternative model essentially requires the computation of the $AICc_{\text{prio},0}$ of the null model, which must be done only once, and the computation of $T_{\text{prio},j}$. However, the latter is the same operation required for the multiple hypothesis tests. For $AICc_{\text{post},j}$ only an additional logarithm is needed, which is not costly. This shows that the computational costs of multiple hypothesis tests and the information criteria are basically the same. A further opportunity of cost reduction is that not all possible hypotheses are tested; e.g., if two observations have very small normalized residuals [Eq. (13a)] or externally studentized residuals [Eq. (13b)], it is not likely that they form a pair of outliers with $n_g = 2$. This allows one to skip the computation of the related test statistic. Following this line of argument allows the computation of only a small subset of test statistics [Eq. (16a) or Eq. (16b)]. The same argumentation is valid for information criteria. It would be interesting to investigate, how much the computational costs can be reduced, without the risk to miss the true outliers. This investigation is planned for the future. Finally, computer technology is recently making breathtaking progress, which will enable the detection of multiple outliers even more extensively.

Geodetic Network for Monitoring of the Reference Point of a Radio Telescope

As a truly practical application, a geodetic network is chosen. The network under consideration was measured in 2012 for monitoring the reference point of a radio telescope at the Geodetic Observatory Wettzell, Germany *(Lösler et al. 2013)*. The instrument used was a total station TS30 from Leica Geosystems AG (St. Gallen, Switzerland). The observations are not simulated, but the real distances and angles measured are used in this monitoring campaign; therefore, it is not known for sure, which are the outliers.

To estimate the 2D coordinates of the 10 survey pillars and the three additional tripod positions, $n = 118$ observations were selected from the campaign. The adjustment is performed as a free network adjustment. Three additional restrictions are introduced to compensate the resulting rank deficiency of the matrix of observation equations $\boldsymbol{A}$ *(Kotsakis 2012)*. The redundancy of the network is 77. All computations are performed with the variance factor known from long-standing experiences with this measurement technology.

The global test [Eq. (26)] is not rejected at the $\alpha = 0.01$ level

$$T_{\text{global}} = \frac{82.735}{77} = 1.07 < F_F^{-1}(1-\alpha, 77, \infty) = 1.41 \tag{41}$$

This is not surprising in the light of Eq. (29), because this kind of test gets blunt, if the redundancy becomes large and the number of outliers and their absolute values are small. To ensure that no outliers exist, it is indispensable to specify an appropriate alternative model. The number of outliers n_g is generally unknown. Furthermore, the number of alternative models [Eq. (8)] strongly depends on n and n_g, e.g., $\binom{118}{6} \approx 3.3 \cdot 10^9$. To restrict the number of permutations, the p-value approach is used as a simple indicator, which selects the alternative model with $\min \log p_{n_g}$. Moreover, $AICc_{n_g}$ is derived for comparison as given in Eq. (38a). Table 2 summarizes the results of the p-value approach and $AICc$, respectively, listing the maximum and minimum values of all models with the same n_g in agreement with the notation introduced previously.

The strategy of Eq. (15) yields ambiguous results, because in most cases T_{n_g} exceeds the critical value c_{n_g}. On the other hand, the p-value approach as well as the $AICc$ become minimal for $n_g = 3$ and select the same alternative model; the p-value approach and the $AICc$ reject the same observations as outliers. Remember that they are not simulated, such that the ground truth is not known. As pointed out, the number of alternative models [Eq. (8)] increases dramatically. Thus, the computational costs become large due to the number of suspected outliers n_g. Moreover, the results confirm the failure of the global test in Eq. (41). Already for $n_g = 1$, T_1 significantly exceeds the critical value for any reasonable value for α'. This may motivate consecutive hypothesis tests assuming $n_g = 1$ in each iteration i as described earlier. The critical values of the test is given by Eq. (14a) and reads

$$c_{\text{prio}} = F_F^{-1}(1 - 0.01, n_g, \infty) = 6.635 \tag{42}$$

The results of the consecutive tests are presented in Table 3. The test statistic T_1 does not exceed the critical value [Eq. (42)] after the third iteration and corroborates the same three dubious observations. The computational costs are comparatively low.

Table 2. Comparison of p-Value Approach and AICc

n_g	$\binom{n}{n_g}$	c_{n_g}	T_{n_g}	$\log p_{n_g}$	$AICc_{n_g} - C_{\text{prio}}$
0	0	-	-	-	210.050
1	118	6.635	15.013	-9.145	199.882
2	6,903	4.605	14.698	-14.698	190.474
3	266,916	3.782	14.178	**-19.594**	**182.448**
4	7,673,835	3.319	11.272	-19.385	185.146
5	174,963,438	3.017	9.5147	-19.255	188.062
6	3,295,144,749	2.802	8.372	-19.283	190.958
77	$9.6 \cdot 10^{31}$	1.413	1.074	-1.1811	-

Note: bold numbers indicate minima of p value and AICc.

Table 3. Results of the i *Consecutive Hypothesis Tests Assuming* $n_g = 1$

i	n	$\max T_{1,i}$
1	118	15.013
2	117	14.382
3	116	13.139
4	115	2.555

Conclusions

The authors have discussed and applied three different approaches to multiple outlier detection: (1) the multiple test with p-value approach, (2) the consecutive test and (3) the information criteria approach. Based on the numerical examples it is not justified to conclude, which approach detects the outliers best, but it is demonstrated that they behave differently and sometimes even produce different results. To find the best approach in some practical sense, more experiences must be gained.

The multiple test with p-value approach suffers from the presence of statistical dependencies between test statistics. These dependencies are amplified as the maximum number of suspected outliers increases. Moreover, there are implications from linearization errors also known from single outlier detection. A rigorous computation of critical values in multiple tests would require a Monte Carlo method following the line of *Lehmann (2012)*. If such dependencies and nonlinearities are disregarded then critical values are only

coarse approximations and the test decisions do not have the desired low error rates. From the model selection point of view it is not justified that the null model plays a special role. It is tested against any alternative model, because only under the null hypothesis does the test statistic have a known probability distribution. (Under the alternative hypothesis the test statistics follow a noncentral F distribution, but the noncentrality parameter is not known.) Finally, the computational costs can be extremely high, as the maximum number of suspected outliers increases. In the paper how to exploit numerical advantages has been shown such that the increase of those costs is manageable.

On the other hand, if a global model test can be put in front, then the choice of a Type I error rate α is straightforward and transparent because of the p-value approach proposed here. A tuning of error rates between global and individual tests can be sidestepped.

The consecutive test is practically approved and implemented in geodetic standard software. A major advantage is the comparatively low computational costs. However, it is theoretically disputable, because the assumption of the first test (that only one outlier is present in the set of observations) is later dropped, but the result of the test (discarded observation) is retained. Thus, the approach has some heuristic property.

There is the problem of computation of critical values in the presence of statistical dependencies, but not as severe as in the multiple test with p-value approach. Furthermore, the number of detected outliers depends on the choice of the Type I error rate α: a larger α means more detected outliers, perhaps spuriously. Masking may cause that no single outlier is detected, although there are multiple outliers present, masking each other. Again, the null model plays a special role without good reason.

In the information criteria approach there is no problem with any statistical dependencies or nonlinearities or choices of any error rate. The null model does not play any special role in the set of selectable models. The approach is easily extendable to cases not yet considered, e.g., other types of alternative models like variance inflation models or nonstandard adjustment models with more unknown variance components. However, there is a diversity of information criteria giving different results in terms of detected outliers. It is not always clear, which criterion suits best for a particular purpose, e.g., when dealing with GNSS time series analysis, *Luo et al. (2011)* are in favor of the combined information criterion (CIC). It was discovered that different equivalent formulations of the outlier detection model lead to different values of the information criterion and possibly also to different decisions in model selection. This phenomenon should be further studied. Moreover, several almost identical least AIC values leave the outlier detection undecided or hardly decidable. And finally, the computational costs are about as high as for the multiple test.

Via Eqs. (38c) and (40) it has been established that a relationship between test statistics and information criteria exists: If the number of outliers is fixed and the variance factor is known, then data snooping and AIC as well as AICc identify the same outliers. Generally, the results do not always coincide.

In summary, the authors can recommend using the information criteria approach to geodetic outlier detection, not least because of its great simplicity and flexibility.

Appendix. Derivation of the Formula for AICc in the Case of an Unknown Variance Factor

Eq. (38b) is derived assuming normal distributed observations, such as Eq. (2), the likelihood function reads

$$L(\hat{\boldsymbol{\theta}};\, \boldsymbol{l}) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} \exp\left(-\frac{\boldsymbol{v}^T\boldsymbol{P}\boldsymbol{v}}{2\hat{\sigma}^2}\right) \tag{43}$$

or equivalently

$$\log L(\hat{\boldsymbol{\theta}};\, \boldsymbol{l}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}^2 - \frac{\boldsymbol{v}^T\boldsymbol{P}\boldsymbol{v}}{2\hat{\sigma}^2} \tag{44}$$

with the maximum likelihood estimator $\hat{\sigma}^2$ of the unknown variance factor σ^2 *(Koch 1999, p. 162f)*

$$\hat{\sigma}^2 = \frac{\boldsymbol{v}^T \boldsymbol{P} \boldsymbol{v}}{n}. \tag{45}$$

By substituting Eq. (45) into Eq. (44), the likelihood function becomes

$$\log L(\hat{\theta};\ l) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\frac{\boldsymbol{v}^T \boldsymbol{P} \boldsymbol{v}}{n} - \frac{n}{2}. \tag{46}$$

Taking the unknown variance factor σ^2 as additional (unknown) parameter $k = u + 1$ in Eq. (37) into account, the AICc is given by

$$AICc_{\text{post},0} = 2(u+1) - 2\log L\left(\hat{\boldsymbol{\theta}};\ \boldsymbol{l}\right) + \frac{2(u+1)(u+2)}{n-u-2} \tag{47}$$

which completes the derivation of Eq. (38b).

References

Abdi, H. (2007). "The Bonferroni and Šidák corrections for multiple comparisons." Encyclopedia of measurement and statistics, N. Salkind, ed., Sage, Thousand Oaks, CA.

Akaike, H. (1974). "A new look at the statistical model identification." IEEE Trans. Autom. Control, 19(6), 716–723.

Atkinson, A. C., and Riani, M. (2008). "A robust and diagnostic information criterion for selecting regression models." J. Japan Stat. Soc., 38(1), 3–14.

Baarda, W. (1968). A testing procedure for use in geodetic networks, Vol. 2, Number 5, Netherlands Geodetic Commission, Publication on Geodesy, Delft, Netherlands.

Baselga, S. (2011). "Nonexistence of rigorous tests for multiple outlier detection in least-squares adjustment." J Surv Eng, 10.1061/(ASCE)SU .1943-5428.0000048, 109–112.

Beckman, R. J., and Cook, R. D. (1983). "Outlier….s." Technometrics, 25(2), 119–149.

Blais, J. A. R. (1991). "On some model identification strategies using information theory." Manuscr. Geodaet., 16(5), 326–332.

Burnham, K. P., and Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach, Springer, Berlin.

Burnham, K. P., and Anderson, D. R. (2004). "Multimodel inference: Understanding AIC and BIC in model selection." Sociol. Methods Res., 33(2), 261–304.

Ding, X., and Coleman, R. (1996). "Multiple outlier detection by evaluating redundancy contributions of observations." J. Geod., 70(8), 489–498.

Felus, Y. A., and Felus, M. (2009). "On choosing the right coordinate transformation method." Proc., FIG Working Week 2009, Eilat, Israel.

Fung, W. K. (1993). "A quasi-Bayesian analysis of regression outliers using Akaike's predictive likelihood." Stat. Pap., 34(1), 133–141.

Hahn, M., Heck, B., Jäger, R., and Scheuring, R. (1989). “Ein Verfahren zur Abstimmung der Signifikanzniveaus für allgemeine Fm,n -verteilte Teststatistiken.” Teil I: Theorie, ZfV, 114, 234–248.

Hahn, M., Heck, B., Jäger, R., and Scheuring, R. (1991). “Ein Verfahren zur Abstimmung der Signifikanzniveaus für allgemeine Fm,n-verteilte Teststatistiken.” Teil II: Anwendungen, ZfV, 116, 15–26.

Heck, B. (1981). “Der Einfluß einzelner Beobachtungen auf das Ergebnis einer Ausgleichung und die Suche nach Ausreißern in den Beobachtungen.” AVN, 88(1), 17–34.

Hurvich, C. M., and Tsai, C. L. (1989). “Regression and time series model selection in small samples.” Biometrika, 76(2), 297–307. Kitagawa, G. (1979). “On the use of AIC for the detection of outliers.” Technometrics, 21(2), 193–199.

Klees, R., Ditmar, P., and Broersen, P. (2002). “How to handle colored observation noise in large least-squares problems.” J. Geod., 76(11), 629– 640.

Knight, N. L., Jinling, W. J., and Rizos, C. (2010). “Generalised measures of reliability for multiple outliers.” J. Geod., 84(10), 625–635.

Koch, K. R. (1999). Parameter estimation and hypothesis testing in linear models, 2nd Ed., Springer, Heidelberg.

Kok, J. J. (1984). “On data snooping and multiple outlier testing.” NOAA Tech. Rep. NOS NGS 30, National Geodetic Information Center, NOS/ NOAA, Rockville, MD.

Kornacki, A. (2014). “Application of the Akaike criterion to detect outliers for the analysis of ash content in barley straw.” Int. Agrophys., 28(2), 257–260.

Kotsakis, C. (2012). “Reference frame stability and nonlinear distortion in minimum-constrained network adjustment.” J. Geod., 86(9), 755–774.

Lehmann, R. (2012). “Improved critical values for extreme normalized and studentized residuals in Gauss–Markov models.” J. Geod., 86(16), 1137–1146.

Lehmann, R. (2013a). “On the formulation of the alternative hypothesis for geodetic outlier detection.” J. Geod., 87(4), 373–386.

Lehmann, R. (2013b). “The 3σ-rule for outlier detection fromthe viewpoint of geodetic adjustment.” J. Surv. Eng., 10.1061/(ASCE)SU.1943-5428.0000112, 157–165.

Lehmann, R. (2014). “Transformation model selection by multiple hypotheses testing.” J. Geod., 88(12), 1117–1130.

Lehmann, R. (2015). “Observation error model selection by information criteria vs. normality testing.” Stud. Geophys. Geod., 59(4), 489–504.

Lehmann, R., and Scheffler, T. (2011). “Monte Carlo based data snooping with application to a geodetic network.” J. Appl. Geod., 5(3–4), 123–134.

Lösler, M., Haas, R., and Eschelbach, C. (2016). “Terrestrial monitoring of a radio telescope reference point using comprehensive uncertainty budgeting– Investigations during CONT14 at the Onsala Space Observatory.” J. Geod., 90(5).

Lösler, M., Neidhardt, A., Mähler, S. (2013). “Impact of different observation strategies on reference point determination–Evaluations from a campaign at the Geodetic Observatory Wettzell.” Proc., 21th European VLBI for Geodesy and Astrometry (EVGA) Working Meeting, Espoo, Finland, N. Zubko and M. Poutanen, eds. Finnish Geodetic Institute, Helsinki, Finland.

Luo, X., Mayer, M., and Heck, B. (2011). “Verification of ARMA identification for modelling temporal correlations of GNSS observations using the ARMASA toolbox.” Stud. Geophys. Geod., 55, 537–556.

Luo, X., Mayer, M., and Heck, B. (2012). “Analysing time series of GNSS residuals by means of AR(I)MA processes.” VII Hotine-Marussi Symp. on Mathematical Geodesy, Int. Association of Geodesy Symposia 137, N. Sneeuw et al., eds., Springer-Verlag, Berlin.

Mierlo, J. (1983). “Problems of computing costs in decision problems.” Mathematical models of geodetic/photogrammetric point determination with regard to outliers and systematic errors, F. E. Ackermann, ed., DGK Reihe A, München, Germany.

Pope, A. J. (1976). “The statistics of residuals and the detection of outliers.” NOAA Tech. Rep. NOS65 NGS1, U.S. Dept. of Commerce, National Geodetic Survey, Rockville, MD.

Pynnönen, S. (1992). “Detection of outliers in regression analysis by information criteria.” Proc., Univ. of Vaasa/Discussion Papers, Vol. 146, Vaasa, Finland.

Rousseeuw, P. J., and Leroy, A. M. (1987). Robust regression and outlier detection, John Wiley& Sons, New York.

Teunissen, P. J. G. (2000). Testing theory: An introduction, 2nd Ed., Series on Mathematical Geodesy and Positioning, Delft Univ. of Technology, Delft, Netherlands.

Teunissen, P. J. G., and de Bakker, P. F. (2013). “Single-receiver singlechannel multi-frequency GNSS integrity: Outliers, slips, and ionospheric disturbances.” J. Geod., 87(2), 161–177.

Ueda, T. (2009). “A simple method for the detection of outliers. EJASA, Electron.” J. App. Stat. Anal., 2(1), 67–76.

Wang, J., Almagbile, A., Wu, Y., and Tsujii, T. (2012). “Correlation analysis for fault detection statistics in integrated GNSS/INS systems.” J. Global Positioning Syst., 11(2), 89–99.

Wang, J., and Knight, N. (2012). “New outlier separability test and its application in GNSS positioning.” J. Global Positioning Syst., 11(1), 46–57.

Yang, L., Wang, J., Knight, N. L., and Shen, Y. (2013). “Outlier separability analysis with a multiple alternative hypotheses test.” J. Geod., 87(6), 591–604.