On the formulation of the alternative hypothesis for geodetic outlier detection

Author

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

Abstract

The concept of outlier detection by statistical hypothesis testing in geodesy is briefly reviewed. The performance of such tests can only be measured or optimized with respect to a proper alternative hypothesis. Firstly, we discuss the important question whether gross errors should be treated as non-random quantities or as random variables. In the first case, the alternative hypothesis must be based on the common mean shift model, while in the second case, the variance inflation model is appropriate. Secondly, we review possible formulations of alternative hypotheses (inherent, deterministic, slippage, mixture) and discuss their implications. As measures of optimality of an outlier detection, we propose the premium and protection, which are briefly reviewed. Finally, we work out a practical example: the fit of a straight line. It demonstrates the impact of the choice of an alternative hypothesis for outlier detection.

Keywords

Geodetic adjustment, Outlier detection, Observation errors, Gross errors, Hypothesis testing, Power of a test, Premium, Protection, Mean shift model, Variance inflation model, Monte Carlo method

1. Introduction

Outlier detection belongs to the daily business activities of modern geodesists. In every good textbook on geodetic adjustment and on estimation in linear models there is a chapter on this subject (e.g. Koch 1999). 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 outlier detection is data snooping, which is based on the pioneering work of Baarda (1968).

We list a number of reasons why there is a continued research on the subject:

(1) Today, we obtain very large sets of observations. It is nearly impossible that such a set is free of outliers.

(2) In former times the standard preprocessing step of geodetic adjustment used to be visual data screening. Here outliers were often detected intuitively. Today, we often use automated and real-time processing algorithms. Here the classical workflow is often not applicable.

(3) Our geodetic ancestors used to be very meticulous people. But in modern geodetic business time is often money, such that we can often no longer afford working like them. A certain amount of outliers in a set of raw observations must be allowed for.

(4) Today, new mathematical tools become available, e.g. tools provided by the fuzzy set theory. Their potential for outlier detection is not yet fully exploited (Neumann et al. 2006).

(5) Our present-day computers are powerful enough for numerical methods very valuable for outlier detection, e.g. Monte Carlo methods (also applied here). We have not yet taken full advantage of them. One advantage is that we can try to optimize outlier tests.

The most often quoted **definition of outliers** is that of Hawkins (1980):
"An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism."
Finding a **definition of gross errors** (also called blunders) is harder than of outliers. The following is taken from (Fan 2010):
"Gross errors are errors due to human mistakes, malfunctioning instruments or wrong measurement methods. Gross errors do not follow certain rules and normally cannot be treated by statistical methods. In principle, gross errors are not permitted and should be avoided by surveyor's carefulness and control routines."
In geodesy, outliers are most often caused by gross errors and gross errors most often cause outliers. This is why they are so often confused. (In the literature one can even find statements that they are the same.) But on the one hand outliers may rarely be the result of fully correct measurements and on the other hand mistakes or malfunctions may not always lead to large deviations, e.g. a small correction wrongly applied. Since Hawkins' and most of the other definitions of outliers restrict themselves to samples (repeated observations) we propose a modified definition:
"An outlier is an observation that is so probably caused by a gross error that it is better not used or not used as it is."
In the following, we will try to discriminate correctly between gross errors and outliers. According to Hawkins (1980) we distinguish between two **mechanisms,** how outliers are supposed to be generated:
(A) **All** standard and gross observation errors come from the **same** non-normal, usually leptokurtic (i.e. thick-tailed) distribution. The outliers are mere realizations of observations coming from the tails of this distribution.
(B) **Some** observation errors come from the normal distribution, but the outliers are "generated by a different mechanism", see Hawkin's definition of outliers above, and therefore follow a different distribution.

If we want to apply our standard geodetic adjustment procedure then those outliers need to be discarded or down-weighted because for leptokurtic distributions this procedure is not optimal. Alternatively we can of course accommodate the outliers by application of robust estimation procedures, see (Rousseeuw and Leroy 2003, Yang 1991, Yang 1999). Robust estimation is outside the scope of this paper.
The paper is organized as follows: After a discussion of the important question whether gross errors should be treated as non-random quantities or as random variables we review the derivation of the elements of the hypothesis tests for outlier detection: Null and alternative hypotheses, test statistics, probabilities of decision errors and critical values. We review various formulations of alternative hypotheses for outlier detection found in the statistical literature (Hawkins 1980, Barnett and Lewis 1994): Although there is a great wealth of forms of such formulations, we have in geodesy restricted ourselves to those formulations, for which we find the critical values in statistical look-up tables. If a present-time computer is available, this restriction is no longer necessary. For the first time it will be made clear that an outlier test performing well for one alternative hypothesis may not be suited for another. This is proofed for a practical example: fitting a straight line. Thus, it is important to formulate the alternative hypotheses in such a way that it best describes the stochastic behavior of the outliers.

2. The statistical modeling of gross errors

The classical separation of geodetic observation errors into

- random errors e_r (noise),
- systematic errors e_s (biases and drifts) and
- gross errors e_g

is motivated by the different stochastic properties of the three components. While random errors are usually treated as random variables in geodesy and systematic errors show by definition a fully predictable non-random behavior, the situation with gross errors is intricate. Randomness is a mathematical abstraction. It is used in engineering sciences to describe phenomena, whose degree of complexity does

not allow us to describe them deterministically. Thus, whether gross errors should be treated as random variables in geodesy depends on the degree of complexity of their generating process.

The **frequentist inference** introduces the concept of probability as a limit of relative frequency. We have to consider the behavior of gross errors if we repeat the gross error generating process in the same way as we do it with random and systematic errors. As a result we realize that it often depends on our definition of repetition whether we get the same gross errors or not (see examples 1 and 2 below).

The **Bayesian inference** uses probabilities to represent the degree of belief that a quantity is close to its true value. Here we can always attribute a probability density function (PDF) to gross errors, even to biases (Koch 2007).

Example 1: Consider a mistake in handwritten recording of a GNSS antenna height on a tripod. The observed value in meter is usually in the interval [1.00,1.99] and is given with two positions after decimal point. Instead of $1.xy$ we spuriously write down $1.yx$. The gross error is $e_g = (y - x)/10 + (x - y)/100 = 9(y - x)/100$.

Frequentist inference: If we repeat the reading and again make the same mistake then we get the same e_g. This indicates non-randomness of this gross error. However, if we repeat the whole setup of the tripod **and** the reading then we will probably end up with different values: Instead of $1.XY$ we write down $1.YX$ and get $e_g = 9(Y - X)/100$. e_g assumes values in the range of -0.81 … +0.81 with different discrete probabilities given in Fig. 1. This indicates randomness of this gross error. Only in the latter case it would be a gain in accuracy to average the observation values. Setting up the tripod is a process of so high complexity that any deterministic treatment is out of the question (like tossing the dice).

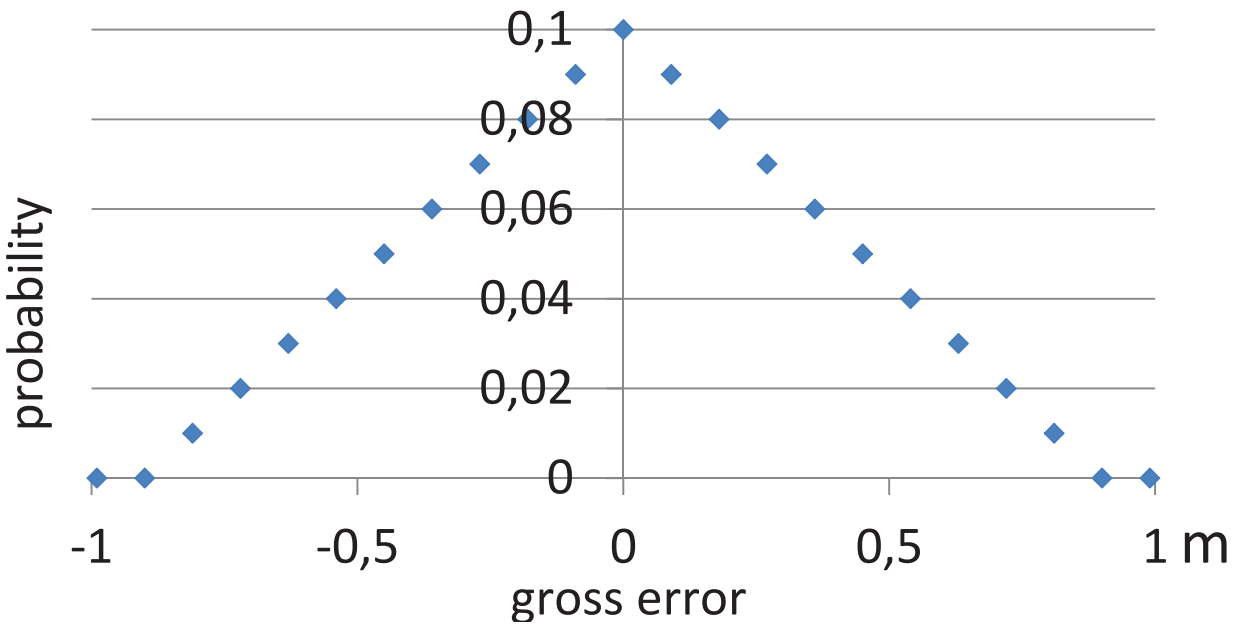

Fig. 1 Discrete probabilities of gross errors in example 1

Example 2: Consider a terrestrial laser scanner erected at a free (more or less randomly selected) station. It scans a wall (Fig. 2) with some target points on a specular surface and some target points hidden by an obstacle blocking the laser path. They produce gross errors in the observed distance, the former positive, the latter negative.

Frequentist inference: If we repeat the scanning from the same station then we get the same gross errors. This indicates non-randomness of these gross errors. However, if we repeat the station setup **and** the scanning from a different station then the mirage point would move and the obstacle effect would change. A random selection of the station would result in a random distribution of the gross errors caused by the specular surface and the obstacle. This indicates randomness of these gross errors. Here the distribution is not as easily specified as in example 1.

Bayesian inference: If we do not know if the erroneous reflection of a terrestrial laser scanner comes from an obstacle blocking the path or if the beam is diffracted on a specular surface then we have to attribute a PDF to this gross error, which has its probability dispersed over the range of its possible values.

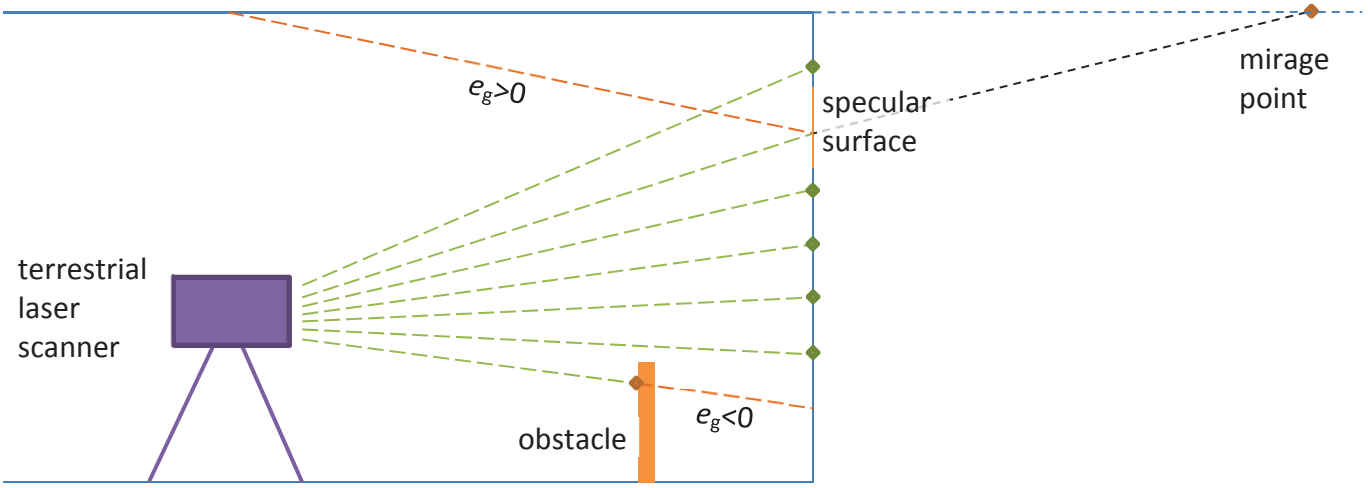

Fig. 2 Gross errors in terrestrial laser scanning, see example 2

If gross errors are treated as biases then they act like systematic errors by shifting the random error distribution by their own value. If we additively combine random and gross errors as $e_r + e_g$ then we get a shifted PDF f_{r+g} for this value of the form

$$f_{r+g}(e_r + e_g) = f_r(e_r) \tag{1}$$

where f_r denotes the PDF associated with the random errors. This assumption is known as the **mean shift model**, see Fig. 3.

If gross errors are treated as random variables then they act like random errors by increasing the variance of the total errors. If we additively combine random and gross errors then we get a convoluted PDF of the form (cf. Mood et al. 1974)

$$f_{r+g}(e_r + e_g) = \int_{-\infty}^{\infty} f_r(e_r + \varepsilon)\, f_g(e_g - \varepsilon) d\varepsilon \tag{2}$$

where f_g denotes the PDF associated with the gross errors. This assumption is known as the **variance inflation model**, see Fig. 3. From this line of reasoning the definition of gross errors quoted in the introduction claiming that those errors "cannot be treated by statistical methods" seems questionable.

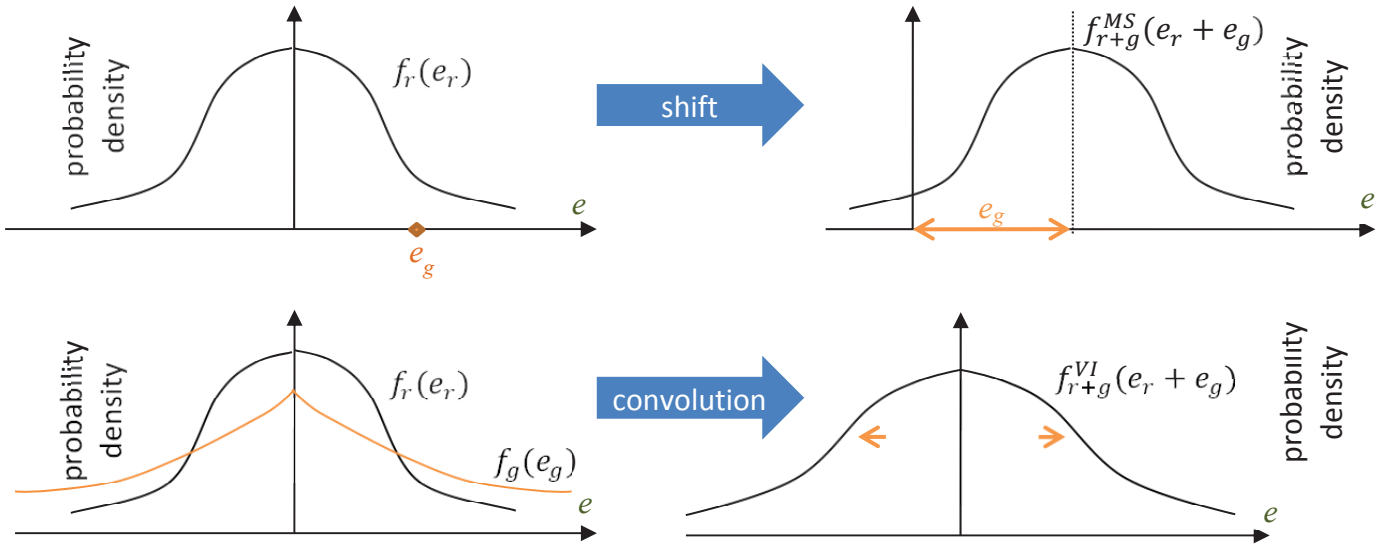

Fig. 3 Probability density functions f of the mean shift (MS) model (top) versus the variance inflation (VI) model (bottom)

In the following, underlined symbols denote random variables.

Example 3: The central normal distribution is a well-established model for random errors: $\underline{e_r} \sim N(0, \sigma_r^2)$. If $\underline{e_g}$ is treated as random variable and no other information on its stochastic behavior is available except for its variance σ_g^2 then it may serve as a model also for gross errors: $\underline{e_g} \sim N(0, \sigma_g^2)$. This choice is justified by the principle of maximum entropy, see (Koch 2007). In view of Fig. 1, this model is not fully correct in example 1, but is also not too far apart. A welcome result is that $\underline{e_r} + \underline{e_g}$ also follows a normal distribution both in case of the mean shift model, where the variance is preserved and the mean is shifted:

$$\underline{e_r} + e_g \sim N(e_g, \sigma_r^2)$$

and in case of the variance inflation model, where the mean is preserved and the variance is inflated:

$$\underline{e_r} + \underline{e_g} \sim N(0, \sigma_r^2 + \sigma_g^2)$$

In geodesy, the mean shift model is by far more popular and widespread for three reasons:

(1) It is analytically convenient to handle.
(2) It is not necessary to introduce a PDF for the gross errors, which is often more arbitrary than for random errors (Example 1 is really an exception).
(3) It is well-known to geodesists from displacement analysis, where it is well justified because deformations clearly exhibit a non-random behavior in the frequentist's sense that immediately repeating the measurement does not change the deformations.

The mean shift model is often tacitly assumed to be the only possible model.

3. Null hypotheses and decision errors

The theoretical framework of outlier detection by hypothesis testing is mathematical statistics, either in the form of the frequentist or the Bayesian inference. In mathematical statistics a hypothesis H is a proposed explanation that the probability distribution of the random vector of n observations $\underline{\boldsymbol{y}}$ belongs to a certain parametric family of distributions W with parameter vector $\boldsymbol{\theta}$:

$$H: \underline{\boldsymbol{y}} \sim W(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \tag{3}$$

The parameter vector $\boldsymbol{\theta}$ may assume values from a set Θ of admissible parameter vectors. If the true probability distribution of the vector of observations $\underline{\boldsymbol{y}}$ belongs to this family W and if its parameter vector $\boldsymbol{\theta}$ is an element of the set Θ then the hypothesis is true, otherwise it is false. If the set Θ comprises only one element then H is said to be a **simple hypothesis**. Otherwise H is said to be a **composite hypothesis**. Parameters from the vector $\boldsymbol{\theta}$ having a range of admissible values rather than a fixed value are called **nuisance parameters**.

The aim is to decide if H is true or false on the basis of a realization $\boldsymbol{y}$ of $\underline{\boldsymbol{y}}$. A statistical hypothesis can never be absolutely verified. (In exceptional cases it can be falsified if we observe a vector $\boldsymbol{y}$ which under H has zero probability: $P(\boldsymbol{y}|H) = 0$). If it would have been very unlikely to have observed $\boldsymbol{y}$ if H is true then it will be rejected, otherwise it will be accepted.

A standard hypothesis when trying to detect outliers in a n-vector of geodetic observations $\underline{\boldsymbol{y}}$ is

H_0: There are no outliers in $\underline{\boldsymbol{y}}$.

This is called a **null hypothesis** H_0, in contrast to the alternative hypotheses to be introduced in section 5. H_0 proposes that inliers are exclusively affected by random errors and by non-random biases such that they deviate from the true value by normal distributed errors. If we try to detect outliers in the standard Gauss-Markoff model (see Koch 1999) then a possible formulation of H_0 in compliance with (3) could read

$$H_0: \underline{\boldsymbol{y}} \sim N(\boldsymbol{A}\boldsymbol{x}, \sigma_0^2 \boldsymbol{P}^{-1}), (\boldsymbol{x}, \sigma_0^2) \in \mathbb{R}^u \times \mathbb{R}^+ \tag{4}$$

where $\boldsymbol{x}$ denotes the unknown u-vector of model parameters and $\boldsymbol{A}$ is the $n \times u$ design matrix relating observations and parameters and having $\operatorname{rank}(\boldsymbol{A}) = q$. $\boldsymbol{P}$ denotes the known $n \times n$-matrix of weights and σ_0^2 is the unknown a priori variance factor. (4) is clearly a composite hypothesis with $u + 1$ nuisance parameters $\boldsymbol{\theta} = (\boldsymbol{x}, \sigma_0^2)$. Other possible hypotheses may involve variance components as additional nuisance parameters or work with fixed σ_0^2 instead. In the latter case the formulation (4) is modified to

$$H_0: \underline{\boldsymbol{y}} \sim N(\boldsymbol{A}\boldsymbol{x}, \sigma_0^2 \boldsymbol{P}^{-1}), \boldsymbol{x} \in \mathbb{R}^u \tag{5}$$

which is clearly a composite hypothesis with u nuisance parameters $\boldsymbol{\theta} = \boldsymbol{x}$.
In practical cases there will remain small probabilities of a decision error. We distinguish between two types of decision errors, see Table 1. Our natural goal is to minimize probabilities of decision error. But the smaller the significance level α is chosen, the more frequently we are inclined to accept H_0 and the more frequently we will accept it if it is actually false. This increases the false negative rate β. Thus we need a tradeoff between both types of decision error.

Table 1 Decision errors in hypothesis testing

decision error	**type I: H_0 is true, but rejected**	**type II: H_0 is false, but accepted**
probability	false positive rate α = size of the test = significance level	false negative rate β =1 – power of the test
in case of outlier detection	false alarm: outlier(s) detected, which are good observations	failing to raise an alarm: outlier(s) remain undetected

4. Critical regions and test statistics

In the space of observations $\mathbb{R}^n$ we select a critical region $C \subset \mathbb{R}^n$ with the property that under the hypothesis H_0 the probability of $\underline{\boldsymbol{y}}$ falling in this region is independent of possible nuisance parameters θ of H_0 and is very small. This probability equals the significance level α:

$$P\left(\underline{\boldsymbol{y}} \in C \middle| H_0\right) = \alpha \tag{6}$$

Since α determines the size of C, it is also referred to as the **size** of the test. If H_0 is a simple hypothesis then $W(\boldsymbol{\theta})$ is a fully specified probability distribution and C could naturally be the complement of a $(1-\alpha)$-confidence region of $\underline{\boldsymbol{y}}$.

A hypothesis test is accomplished in five steps.

(1) Propose a null hypothesis H_0.
(2) Chose a standard value for α, say 0.1 or 0.05 or 0.01.
(3) Chose a critical region C of probability (6).
(4) Observe $\boldsymbol{y}$.
(5) If $\boldsymbol{y} \in C$ then reject H_0, otherwise accept H_0.

(In geodesy we often exchange (3) and (4), but this is dangerous. C is not allowed to be choosen depending on y. Otherwise, we get a post hoc hypothesis, i.e. a hypothesis suggested by the observations, a widespread misuse of statistics also in geodesy.)
However, if H_0 is a composite hypothesis then we often get a different confidence region for every $\boldsymbol{\theta} \in \Theta$. Unfortunately, this is the standard situation in geodetic outlier detection, cf. (4) and (5). This introduces an undesirably great degree of freedom when choosing C.
In geodesy and in many other disciplines a different approach is more common: Instead of choosing C we may chose a scalar random function $\underline{T}(\underline{\boldsymbol{y}})$ called a test statistic such that its distribution under H_0 does not depend on the nuisance parameters $\boldsymbol{\theta}$ and can be easily computed. If for an observed vector $\boldsymbol{y}$ we find that $T(\boldsymbol{y})$ is outside some confidence interval $[c_{min}, c_{max}]$ of $\underline{T}|H_0$ then it would be very unlikely to have observed $\boldsymbol{y}$ if H_0 is true. Consequently, H_0 will be rejected, otherwise it will be accepted. In some cases it is not wise to reject H_0 if, although unlikely on various occasions, $T(\boldsymbol{y})$ assumes a very small value because we believe that H_0 holds true nonetheless. Then the confidence interval is chosen as $[-\infty, c_{max}]$. A test of this kind is called a one-sided test, in contrast to the general two-sided test. c_{min}, c_{max} define the critical region C and are called critical values.
A hypothesis test is then accomplished as follows:

(1) Propose a null hypothesis H_0.

(2) Chose a standard value for α, say 0.1 or 0.05 or 0.01.
(3) Chose a test statistic $\underline{T}(\underline{\boldsymbol{y}})$ with known distribution under H_0.
(4) Find two critical values c_{min}, c_{max} such that $P\left(\underline{T}(\underline{\boldsymbol{y}}) < c_{min}\middle|H_0\right) + P\left(\underline{T}(\underline{\boldsymbol{y}}) > c_{max}\middle|H_0\right) = \alpha$. In particular, chose $c_{min} = -\infty$ for a one-sided test.
(5) Observe y.
(6) If $T(\boldsymbol{y}) < c_{min}$ or if $T(\boldsymbol{y}) > c_{max}$ then reject H_0, otherwise accept H_0.

For outlier detection in geodesy we often use the following test statistics (Baarda 1968, Pope 1976, Teunissen 2000, Lehmann 2012b):

- posterior/prior variance ratio: $\underline{T}_g = \underline{\hat{\boldsymbol{e}}}^T P \underline{\hat{\boldsymbol{e}}}/((n-q)\sigma_0^2)$
- individual normalized residuals: $\underline{T}_{n,i} = \underline{\hat{e}}_i/\left(\sigma_0\sqrt{q_{\hat{e}_i\hat{e}_i}}\right), i = 1, \dots, n$
- individual studentized residuals: $\underline{T}_{s,i} = \underline{\hat{e}}_i/\sqrt{q_{\hat{e}_i\hat{e}_i}\,\underline{\hat{\boldsymbol{e}}}^T P \underline{\hat{\boldsymbol{e}}}/(n-q)}\;\; i = 1, \dots, n$
- extreme normalized residuals: $\underline{T}_n = \max\left|\underline{T}_{n,i}\left(\underline{\boldsymbol{y}}\right)\right|$
- extreme studentized residuals: $\underline{T}_s = \max\left|\underline{T}_{s,i}\left(\underline{\boldsymbol{y}}\right)\right|$

$\underline{\hat{e}}$ is the n-vector of residuals (estimated observation errors) with elements $\underline{\hat{e}}_i$ and $q_{\hat{e}_i\hat{e}_i}$ denote the diagonal elements of the cofactor matrix

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

The superscript "-" denotes some generalized inverse matrix. In section 7 we will comment on the derivation of these test statistics.

Under the hypothesis H_0 in (5) we find that

$$(n-q)\cdot\underline{T}_g \sim \chi^2(n-q)$$

has a central χ^2-distribution with $n-q$ degrees of freedom and

$$\underline{T}_{n,i} \sim N(0,1)$$

has a standard normal distribution.

Example 4: Choosing size $\alpha = 0.05$ and since

$$P\left(|\underline{T}_{n,i}| > 1.96\middle|H_0\right) = 0.05$$

we are inclined to reject H_0 in (5) if $|\underline{T}_{n,i}|$ exceeds 1.96.

Under the hypothesis H_0 in (4) or (5) we find that

$$\underline{T}_{s,i} \sim \tau(n-q-1)$$

has a τ-distribution with $n-q-1$ degrees of freedom. This distribution is derived by Thompson (1935). It is introduced to geodesy by Pope (1976) and is later adopted by Koch (1999) and others.

The distributions of $\underline{T}_n$ and $\underline{T}_s$ are more difficult to derive. The common approximation is the following: $\underline{T}_x < c$ is equivalent to $\left|\underline{T}_{x,i}\right| < c$ for all $i = 1, \dots, n$ where $x \in \{n, s\}$. If these n random events were nearly independent then we could write

$$1-\alpha = P\left(\underline{T}_x < c\right) \approx \prod_{i=1}^{n} P\left(-c < \underline{T}_{x,i} < c\right) = (1-\alpha')^n \tag{7}$$

This is to say: The test with test statistic $\underline{T}_x$ and significance level α is replaced by a family of n tests with test statistics $\underline{T}_{x,i}, i = 1, \dots, n$ and significance level α'. Since $\alpha \ll 1$, we find with good accuracy the relationship

$$\alpha \approx n\alpha' \tag{8}$$

It is called **Bonferroni equation** (Abdi 2007).

However, $\left|\underline{T}_{x,i}\right|, i = 1, \dots, n$ are not always sufficiently independent. Lehmann (2012b) suggests using a Monte Carlo method (see section 10) for the numerical evaluation of the relevant integrals. It is shown that the true distribution may differ substantially from the approximation above.

In the geodetic method of **data snooping** according to Baarda (1968) we use a stepwise procedure:

(1) **Global test**: We invoke $\underline{T}_g$ as a test statistic for general model misspecifications.

(2) **Local test**: If the global test rejects H_0 then we localize the outlier by means of $\underline{T}_n$.
(3) **Rejection rule**: If an outlier is found then it is discarded or down-weighted or re-measured and the procedure is restarted.

If σ_0^2 is assumed to be unknown then $\underline{T}_g$ and $\underline{T}_n$ cannot be used and we are left with only the local test based on $\underline{T}_s$ and the rejection rule. (σ_0^2 being unknown is the standard assumption in most other disciplines performing outlier tests.)

5. Optimal design of tests and the alternative hypothesis

Hypothesis tests for outlier detection should be subject to re-design and optimization because today we have sufficient computing power to apply optimization techniques which were ineligible in previous decades. There are three starting points:

(1) α determines the inhibition threshold for an alarm and should be chosen with care. Otherwise we will run the risk of either losing too many good observations or of leaving outliers undetected. For data snooping the problem is first addressed in (Lehmann 2010, Lehmann and Scheffler 2011). Any standard value of α (say 0.1 or 0.05 or 0.01) is doubtful.
(2) Remember that α directly specifies the probability of a type I decision error, i.e. of discarding good observations. But in geodetic outlier detection a type II decision error, i.e. an undetected outlier, is often considered to be more harmful. It is not possible to compute the probability β when only H_0 is specified.
(3) The choice of a test statistic $\underline{T}$ in outlier tests is by no means unique. In fact, it may be intuitively more appealing than the choice of C, but it is eventually no less arbitrary. See (Barnett and Lewis 1994) or (Hawkins 1980) for very long lists of rival test statistics. One could try to minimize β or equivalently maximize the **power of the test** $1-\beta$, see section 7. Also other measures of optimality can be conceived, see section 8.

None of these three goals can be achieved without specification of an **alternative hypothesis** H_A to be adopted if H_0 is rejected.

Example 5: Compare the two alternatives

$H_{A,1}$: many rather small outliers vs. $H_{A,2}$: few very large outliers

In the first case we must not be afraid of frequent false alarms and a small α would certainly let most outliers pass. Thus, α should be chosen large in order to really detect any outlier. In the second case we will hardly fail to raise an alarm. Here α can safely be chosen small in order to prevent frequent false alarms.

A proper formulation of the two rival hypotheses as an extension of (3) is

$$H_0: \underline{\boldsymbol{y}} \sim W(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_0 \quad \text{vs.} \quad H_A: \underline{\boldsymbol{y}} \sim W(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_A \tag{9}$$

This is applicable if both $\underline{\boldsymbol{y}}|H_0$ and $\underline{\boldsymbol{y}}|H_A$ belong to the same parametric family of distributions W, which can be assumed here. Θ_0 and Θ_A are two disjoint subsets of the parameter space Θ of W.

6. Types of alternative hypotheses for outlier detection

While H_0 in (4) or (5) is generally beyond dispute in geodesy, the correct formulation of the alternative hypothesis deserves an in-depth discussion.

Hawkins' mechanisms, see section 1, give rise to different formulations of alternative hypotheses. We refer to (Barnett and Lewis 1994) for the following synopsis:

1. Inherent alternatives: According to mechanism (A), all observation errors come from the same non-normal distribution:

$$H_A: \underline{y}_i \sim W(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_A, i = 1, \dots, n \tag{10}$$

Candidates for W are leptokurtic distributions like generalized normal distribution (Nadarajah 2005) or Student's-t distribution. These distributions comprise the normal distribution as a special or limiting case such that the formalism (9) is applicable.

No gross error is directly involved here. But we can consider the leptokurtic distribution W in (10) to be the

result of a variance inflation according to (2). In fact, the outlier detection and rejection in the presence of the inherent alternative may be regarded as "thinning the tails" of distribution W in (10).

2. Deterministic alternatives: A fixed and known subset of observation errors is central normally distributed according to (4) or (5). The remaining observations are affected by gross errors and thus come from a **different** family of distributions. In other words, if we have reasons to reject H_0 then we believe to know which observations are outlying. The order of observations may be chosen "inliers first" such that we can formulate

$$H_A: \underline{y}_i \sim W(\boldsymbol{\theta}), \begin{cases} \boldsymbol{\theta} \in \Theta_0, & i = 1, \dots, n_1 \\ \boldsymbol{\theta} \in \Theta_A, & i = n_1 + 1, \dots, n \end{cases} \tag{11}$$

Here we employ Hawkins'mechanism (B). $W(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_A$ may the result of either a mean shift model (1) or a variance inflation model (2). In view of example 3, $W(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_A$ may be another normal distribution with parameters $\boldsymbol{\theta} \notin \Theta_0$.

In order to avoid formulating a post hoc hypothesis, it is important that the set of suspected outliers is not determined by inspection of the observations. E.g. it is not allowed here to simply use the observations with the extreme residuals as suspected outliers.

The deterministic alternative is the best established type of H_A in geodesy, mostly in conjunction with the mean shift model (Baarda 1968, Pope 1976, Koch 1999, Teunissen 2000, Kargoll 2012).

3. Slippage alternatives: A fixed and known number n_1 of observation errors is central normally distributed according to (4) or (5), but the remaining $n_2 = n - n_1$ observation errors come from a distribution with different parameters. Let I_1 and I_2 denote the **unknown** disjoint subsets of $\{1, \dots, n\}$ with n_1 and n_2 elements, respectively. Then we can formulate

$$H_A: \underline{y}_i \sim W(\boldsymbol{\theta}), \begin{cases} \boldsymbol{\theta} \in \Theta_0, & i \in I_1 \\ \boldsymbol{\theta} \in \Theta_A, & i \in I_2 \end{cases} \tag{12}$$

Here we again employ Hawkins' mechanism (B). $W(\boldsymbol{\theta})$ may as well be the result of either a mean shift model (1) or a variance inflation model (2). The slippage alternative is identical to the deterministic alternative except that here we do not know which observations are outlying. In order to avoid formulating a post hoc hypothesis (see section 4), the number of outliers must not be determined by inspection of the observations.

For example, if outliers are known to be rare and we are inclined to reject H_0 then we may alternatively propose that there is exactly one outlier in $\underline{y}$, i.e. $n_2 = 1$, but we do not know which one.

It is common in geodesy to replace this slippage hypothesis by a family of deterministic hypotheses with $n_2 = 1$ for each single observation. If the slippage hypothesis is true than exactly one hypothesis of the deterministic family is true and vice versa. The extreme residual as a test statistic for (12) can be replaced by the individual residuals as test statistics for (11) with the significance level divided by n. The approximate equivalence is shown by (7), (8).

4. Mixture alternatives: Any observation error comes with **fixed and known probability** $1 - \varepsilon$ from the normal distribution $N(0, \sigma_r^2)$ with PDF f_r and with small probability ε from a different distribution with PDF f_{r+g}, possibly another normal distribution with parameters different from those in f_r. Again, only the latter observations are affected by gross errors. Their number is not fixed, but random. If both populations of observation errors are unified then the total PDF can then be written as the PDF of a **contaminated distribution** (Goldstein 1982)

$$f(e) = (1 - \varepsilon) f_r(e) + \varepsilon f_{r+g}(e) \tag{13}$$

In the geodetic literature this type of distribution is used for outlier detection or robust estimation in (Yang1991, Hekimoglu and Koch 2000, Gui et al. 2011, Lehmann and Scheffler 2011). We can formulate

$$H_A: \underline{y}_i \sim W(\boldsymbol{\theta}): \boldsymbol{\theta} \in \Theta_A, i = 1, \dots, n \tag{14}$$

where W is a family of contaminated distributions with PDF of type (13). The normal distribution is included in this family through $\varepsilon = 0$ such that the formalism (9) is applicable.

Note that (14) and (10) are fully identical. In fact, the mixture alternative can be regarded as a special inherent alternative with a contaminated distribution.

Here we again employ Hawkins' mechanism (B). $W(\boldsymbol{\theta})$ may as well be the result of either a mean shift model (1) or a variance inflation model (2). In the former case $W(\boldsymbol{\theta})$ is called a location-contaminated distribution and in the latter case a scale-contaminated distribution.

Example 3 (cont'd): Using a mixture alternative, the resulting contaminated PDFs (13) are the PDF of the location-contaminated normal distribution

$$f_{LC}(y|\mu,\mu+e_g,\sigma_r^2,\varepsilon)=\frac{1}{\sigma_r\sqrt{2\pi}}\left((1-\varepsilon)\exp\left(-\frac{(y-\mu)^2}{2\sigma_r^2}\right)+\varepsilon\exp\left(-\frac{((y-\mu-e_g)^2}{2\sigma_r^2}\right)\right) \quad (15)$$

and the PDF of the scale-contaminated normal distribution (Lehmann 2012a)

$$f_{SC}(y|\mu,\sigma_r^2,\sigma_r^2+\sigma_g^2,\varepsilon)=\frac{1}{\sqrt{2\pi}}\left(\frac{1-\varepsilon}{\sigma_r}\exp\left(-\frac{(y-\mu)^2}{2\sigma_r^2}\right)+\frac{\varepsilon}{\sqrt{\sigma_r^2+\sigma_g^2}}\exp\left(-\frac{(y-\mu)^2}{2(\sigma_r^2+\sigma_g^2)}\right)\right) \quad (16)$$

7. Most powerful outlier tests

Given some value α, a nearby solution to the optimal choice of a test statistic is to minimize β or equivalently to maximize the power $1-\beta$. A test maximizing the power is called a **most powerful** (MP) test (Teunissen 2000, Kargoll 2012).

Unfortunately, for all practically relevant outlier tests the power depends on the nuisance parameters in Θ_A. A common solution is to introduce an invariance principle, which reduces the set of possible test statistics such that the power has a unique maximum. In this way we derive **uniformly most powerful invariant** (UMPI) tests. The test statistics $\underline{T}_g, \underline{T}_{n,i}, \underline{T}_{s,i}$ are UMPI test statistics with respect to the mean shift model. $\underline{T}_g$ can be derived from a slippage or deterministic alternative with $n_2=n-u$ (Teunissen 2000, Kargoll 2012).

$\underline{T}_{n,i}, \underline{T}_{s,i}$ can be derived from a deterministic alternative and a single outlier in the i-th observation, i.e. $n_2=1$. One could expect the test statistics $\underline{T}_n, \underline{T}_s$ to be related somehow to the slippage alternative with $n_2=1$. But no UMPI property has yet been rigorously derived. This would be difficult to accomplish because the extreme residuals $\underline{T}_n, \underline{T}_s$ are **non**linear functionals of $\underline{y}$. But at least approximately we can transform $\underline{T}_n, \underline{T}_s$to a family of UMPI test statistics $\underline{T}_{n,i}, \underline{T}_{s,i}, i=1,\dots,n$ by (7).

But even if for a practically useful H_A a UMPI test can be constructed, the power of a test $1-\beta$ as an optimization criterion still has the following disadvantages:

- It does not indicate how to choose α. E.g. if α is chosen too large then we lose a lot of good observations. Even if all outliers are discarded by the test, it does not yield a satisfactory result.
- It disregards the rejection rule (down-weighting, discarding, re-measuring etc.). In other words, a most powerful outlier test does not guarantee best (in whatever sense) estimated parameters.

An alternative optimization criterion proposed by Anscombe (1960) will be discussed in the next section.

8. Anscombe's premium and protection

Consider a scalar parameter x to be estimated from the observations $\underline{\boldsymbol{y}}$. If $\underline{\boldsymbol{y}}$ would be free of outliers (i.e. H_0 holds true) then an optimal estimator of x in some sense is denoted by $\underline{\hat{x}}$. But since we cannot be sure that H_0 holds, we try to detect outliers in $\underline{\boldsymbol{y}}$ by hypothesis testing and in case of rejection of H_0 we discard or down-weight outliers and reprocess the remaining observations. That is to say, we come up with a rival estimator $\underline{\hat{x}}'$ which is intended to protect against outliers in $\underline{\boldsymbol{y}}$ as specified by H_A.

However, if H_0 holds true then we expect $\underline{\hat{x}}'$ to perform rather poorly in comparison to $\underline{\hat{x}}$. This is the price we are prepared to pay for this protection. Anscombe (1960) introduces the notion of the **premium** expressing the relative **loss** of the outlier detection provided that H_0 holds true:

$$Prem=\frac{\mathrm{MSE}(\underline{\hat{x}}'|H_0)-\mathrm{MSE}(\underline{\hat{x}}|H_0)}{\mathrm{MSE}(\underline{\hat{x}}|H_0)} \quad (17)$$

MSE is the mean squared error as a measure of deviation of true and estimated parameter. If $\underline{\hat{x}}$ is optimal under H_0 in a way that $\mathrm{MSE}(\underline{\hat{x}}|H_0)$ is minimum then the premium is obviously positive. Moreover, if $\underline{\hat{x}}|H_0$ and $\underline{\hat{x}}'|H_0$ are best linear unbiased estimates, the latter estimate perhaps with some down-weighted or

discarded observations, then $\mathrm{MSE}(\underline{\hat{x}}|H_0)$ and $\mathrm{MSE}(\underline{\hat{x}}'|H_0)$ equal the variances $\mathrm{var}(\underline{\hat{x}}|H_0)$ and $\mathrm{var}(\underline{\hat{x}}'|H_0)$. And these values are known to be independent of the values of the parameters x, but are proportional to σ_0^2. Regardless of σ_0^2 being a nuisance parameter as in (4) or not as in (5), the premium can be computed without σ_0^2 because it cancels in (17). Additionally, $\mathrm{MSE}(\underline{\hat{x}}'|H_0)$ depends on α and the rejection rule. The relationship between α and the premium is simple: For $\alpha = 0$ we have $\underline{\hat{x}}' = \underline{\hat{x}}$ and consequently the premium is zero. Then it is monotonically increasing with α.

As the opposite side of the coin Anscombe (1960) introduces the notion of **protection** expressing the relative **gain** of the outlier detection estimator $\underline{\hat{x}}'$ with respect to $\underline{\hat{x}}$, provided that H_A holds true:

$$Prot = \frac{\mathrm{MSE}(\underline{\hat{x}}|H_A) - \mathrm{MSE}(\underline{\hat{x}}'|H_A)}{\mathrm{MSE}(\underline{\hat{x}}|H_A)} \tag{18}$$

Any reasonable protection would be positive, but this is not at all guaranteed (see below). If H_A is such that $\underline{\hat{x}}|H_A$ is a meaningless result then a good protection would be close to 1. For the mean shift model $\underline{\hat{x}}|H_A$ and $\underline{\hat{x}}'|H_A$ are biased estimates. $\mathrm{MSE}(\underline{\hat{x}}|H_A)$ and $\mathrm{MSE}(\underline{\hat{x}}'|H_A)$ are still independent of the values of the parameters x, but usually depend on the other nuisance parameters in H_A. Additionally, $\mathrm{MSE}(\underline{\hat{x}}'|H_A)$ depends on α and the rejection rule. And so does the protection. This situation resembles the dependence of the power of the test on the nuisance parameters in H_A, which makes it necessary to introduce UMPI tests, see previous section.

If $\boldsymbol{x} = (x_1, \dots, x_u)$ is a vector then it is suggested by Anscombe (1960) to extend (17),(18) to:

$$Prem = \frac{\Sigma_i \mathrm{MSE}(\underline{\hat{x}}_i'|H_0) - \Sigma_i \mathrm{MSE}(\underline{\hat{x}}_i|H_0)}{\Sigma_i \mathrm{MSE}(\hat{x}_i|H_A)} \tag{19}$$

$$Prot = \frac{\Sigma_i \mathrm{MSE}(\underline{\hat{x}}_i|H_A) - \Sigma_i \mathrm{MSE}(\underline{\hat{x}}_i'|H_A)}{\Sigma_i \mathrm{MSE}(\underline{\hat{x}}_i|H_A)} \tag{20}$$

However, this might not always be reasonable because parameters x_i may have different units like coordinates, parameters of orientation and scale in a horizontal geodetic network. But in any case, the sums in (19),(20) may not necessarily extend over the complete set of parameters x_i of a geodetic model, but only over those parameters of "primary interest". The latter notion is adopted from Lehmann and Scheffler (2011), where it is pointed out that we often need auxiliary parameters for establishing a model. And the estimated values of those parameters may not be required anymore after the processing of the observations. Thus, if we evaluate the sums in (19),(20) then it is proposed to skip those parameters: E.g. in a geodetic control network the sums in (19),(20) may extend only over the coordinates of control points, see (Lehmann and Scheffler 2011). In the extreme case that only one parameter x_i is of primary interest, we return to (17),(18) with $\underline{\hat{x}}$ replaced by $\underline{\hat{x}}_i$.

Clerici and Harris (1980) introduce the notions of premium and protection to geodetic outlier detection. Later they apply the concept to displacement analysis, which is in principle equivalent to the detection of outliers by the mean shift model (Clerici and Harris 1983). The protection is used in Lehmann and Scheffler (2011) as an optimization measure for data snooping, but it is not called "protection" because at that time Anscombe's term was unknown to the authors.

A good hypothesis test for outlier detection would be such that the protection is high and at the same time the premium is low. In this way one can opt for one of various rival applicants for $\underline{\hat{\boldsymbol{x}}}'$ or optimize its perfomance by tuning parameters. However, maximizing the protection while simultaneously minimizing the premium is impossible. In fact, the least premium equals zero by $\underline{\hat{\boldsymbol{x}}}' = \underline{\hat{\boldsymbol{x}}}$. But then we would not get any protection. In turn, a higher protection is likely to reject more good observations, which increases the premium. Here we propose a practical solution to this dilemma: Chose a certain premium you are willing to pay, say 10%, and try to get the best possible protection under this restriction.

The situation with $Prem$ and $Prot$ is similar to the optimization of decision error levels α, β, where we chose a certain α and try to find the test with the smallest β, see section 7. We believe that chosing $Prem$ is practically more self-evident than choosing α.

Unfortunately, the terms of the form $\mathrm{MSE}(\underline{\hat{x}}_i'|H)$ cannot be evaluated analytically because the functional relationship between $\underline{\boldsymbol{y}}$ and $\underline{\hat{\boldsymbol{x}}}'$ is too complicated. A well-established procedure for their numerical calculation is the Monte Carlo (MC) method, already applied by Lehmann and Scheffler (2011) in a similar computation. In essence the MC method replaces random variates by computer generated pseudo random numbers, probabilities by relative frequencies and expectations by arithmetic means over large sets of such numbers. A computation with one set of pseudo random numbers is a MC experiment.
An advantage of the MC method is that the MSEs can be computed for all relevant critical values c in parallel as follows: In each MC experiment we compute both $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{x}}'$, regardless of the value of test statistic T. The realization $\hat{\boldsymbol{x}}$ contributes to all MSEs for $c > T$, the realization $\hat{\boldsymbol{x}}'$ to all MSEs for $c \leq T$. This yields arbitrarily dense values of the functions Prem(c) and Prot(c) with nearly no extra computational costs. This procedure is applied for the generation of Fig. 4-6, see below.

9. Practical example: MSEs for fitting a straight line

As a practical example we chose the straight line fit with n equidistant data points, a common model not only in geodesy.

9.1 Setup and MSEs for the null hypothesis

As the null hypothesis H_0 we use in the following (5) with $\boldsymbol{P} = \boldsymbol{I}$ (unit matrix) such that $\sigma_0^2 = \sigma_r^2$.
Out of the great variety of possible alternative hypotheses we chose for illustration

(S) slippage alternatives (12) with $n_2 = 1$ and
(M) mixture alternatives (14) with $\varepsilon = 1/n$.

By (S) we assume that there is at most one outlier in $\underline{\boldsymbol{y}}$, but which one is unknown. In the following, the unknown index of the outlying observation is denoted by k. (M) means that there is in average one outlier in $\underline{\boldsymbol{y}}$, but it can happen that there is none or there are multiple outliers. Both (S) and (M) will be combined with either

(MS) a mean shift model (1) or
(VI) a variance inflation model (2) with normally distributed gross errors, see example 3.

Thus, we arrive at four different combinations, denoted as $H_A^{SMS}, H_A^{SVI}, H_A^{MMS}, H_A^{MVI}$. They can be formulated as follows:

$$H_A^{SMS}: \underline{y}_i \sim N(\boldsymbol{a}_i\boldsymbol{x} + \delta_{ik}e_g, \sigma_r^2), \boldsymbol{x} \in \mathbb{R}^u, k \in \{1, \dots, n\}, e_g \in \mathbb{R}\backslash 0\,, i = 1, \dots, n$$
$$H_A^{SVI}: \underline{y}_i \sim N(\boldsymbol{a}_i\boldsymbol{x}, \sigma_r^2 + \delta_{ik}\sigma_g^2), \boldsymbol{x} \in \mathbb{R}^u, k \in \{1, \dots, n\}, \sigma_g^2 \in \mathbb{R}^+, i = 1, \dots, n$$
$$H_A^{MMS}: \underline{y}_i \text{ follows PDF } f_{LC}(y_i|\boldsymbol{a}_i\boldsymbol{x}, \boldsymbol{a}_i\boldsymbol{x} + e_g, \sigma_r^2, \varepsilon), \boldsymbol{x} \in \mathbb{R}^u, e_g \in \mathbb{R}\backslash 0, i = 1, \dots, n$$
$$H_A^{MVI}: \underline{y}_i \text{ follows PDF } f_{SC}(y_i|\boldsymbol{a}_i\boldsymbol{x}, \sigma_r^2, \sigma_r^2 + \sigma_g^2, \varepsilon), \boldsymbol{x} \in \mathbb{R}^u, \sigma_g^2 \in \mathbb{R}^+, i = 1, \dots, n$$

where δ_{ik} is the Kronecker symbol and $\boldsymbol{a_i}$ denotes the i-th row of $\boldsymbol{A}$. The PDFs f_{LC}, f_{SC} are from (15), (16). The slippage alternatives have $u + 2$ and the mixture alternatives have $u + 1$ nuisance parameters.
The observation equations read for the slippage alternatives

$$y_i = x_1 + x_2 i + e_{r,i} + \delta_{ik}e_g, i = 1, \dots, n$$

and for the mixture alternatives

$$y_i = x_1 + x_2 i + e_{r,i} + be_{g,i}, i = 1, \dots, n$$

where b is a Bernoulli random variate with probability $\varepsilon = 1/n$. Under H_0 the least squares estimate of the parameter vector (intercept, slope) can be obtained by simple least squares calculus as

$$\underline{\hat{\boldsymbol{x}}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\underline{\boldsymbol{y}} = \frac{2}{n^2 - n}\begin{pmatrix} 2n+1 & -3 \\ -3 & \frac{6}{n+1} \end{pmatrix}\begin{pmatrix} 1 & \cdots & 1 \\ 1 & \cdots & n \end{pmatrix}\underline{\boldsymbol{y}} \tag{21}$$

It is well-known that $\underline{\hat{\boldsymbol{x}}}|H_0$ is an unbiased estimate, therefore the MSEs equal the variances to be obtained by covariance propagation applied to (21):

$$\mathrm{MSE}(\underline{\hat{x}}_1|H_0) = \mathrm{var}(\underline{\hat{x}}_1|H_0) = \frac{4n+2}{n^2 - n}\sigma_r^2 \tag{22}$$

(23)

$$\mathrm{MSE}\left(\hat{\underline{x}}_2|H_0\right) = \mathrm{var}\left(\hat{\underline{x}}_2|H_0\right) = \frac{12}{n^3 - n}\sigma_r^2$$

9.2 MSEs of the least squares estimates in the mean shift model

Due to the mean shift, $\hat{\underline{\boldsymbol{x}}}|H_A^{SMS}$ and $\hat{\underline{\boldsymbol{x}}}|H_A^{MMS}$ are biased estimates:

$$E\left\{\hat{\underline{\boldsymbol{x}}}|H_A^{SMS}\right\} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T E\left\{\underline{\boldsymbol{y}}|H_A^{SMS}\right\} = \boldsymbol{x} + (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{a}_k^T e_g, \boldsymbol{a}_k = (1\ k)$$

$$E\left\{\hat{\underline{\boldsymbol{x}}}|H_A^{MMS}\right\} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T E\left\{\underline{\boldsymbol{y}}|H_A^{MMS}\right\} = \boldsymbol{x} + (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(1 \quad \cdots \quad 1)^T \varepsilon e_g$$

where $E\left\{\underline{\boldsymbol{y}}|H_A^{MMS}\right\}$ is derived from the expectation of (15), which is by simple probability calculus obtained as $\mu + \varepsilon e_g$. With (21) this yields the following biases:

$$\mathrm{bias}\left(\hat{\underline{x}}_1|H_A^{SMS}\right) = \frac{2}{n^2 - n}(2n + 1 - 3k)$$

$$\mathrm{bias}\left(\hat{\underline{x}}_2|H_A^{SMS}\right) = \frac{2}{n^3 - n}(-3(n+1) + 6k)$$

$$\mathrm{bias}\left(\hat{\underline{x}}_1|H_A^{SMS}\right) = \varepsilon e_g$$

$$\mathrm{bias}\left(\hat{\underline{x}}_2|H_A^{MMS}\right) = 0$$

(The latter two terms can also be found directly, if one considers the straight line solution for coincident observations $\boldsymbol{y} = (1 \quad \cdots \quad 1)^T$ to be $\hat{\underline{\boldsymbol{x}}} = (1{,}0)^T$).

$\underline{\boldsymbol{y}}|H_A^{SMS}$ has the same covariance matrix as $\underline{\boldsymbol{y}}|H_0$ and consequently $\hat{\underline{x}}_i|H_A^{SMS}$ have the same variances as $\hat{\underline{x}}_i|H_0$:

$$\mathrm{var}\left(\hat{\underline{x}}_1|H_A^{SMS}\right) = \frac{4n + 2}{n^2 - n}\sigma_r^2$$

$$\mathrm{var}\left(\hat{\underline{x}}_2|H_A^{SMS}\right) = \frac{12}{n^3 - n}\sigma_r^2$$

With H_A^{MMS} the situation is different: The variance of (15) can be derived by simple calculus obtained as $\sigma_r^2 + (\varepsilon - \varepsilon^2)e_g^2$. This yields the covariance matrix of $\underline{\boldsymbol{y}}|H_A^{MMS}$ as

$$\boldsymbol{\Sigma}_{yy}|H_A^{MMS} = \left(\sigma_r^2 + (\varepsilon - \varepsilon^2)e_g^2\right)\boldsymbol{I}$$

and by covariance propagation follows

$$\mathrm{var}\left(\hat{\underline{x}}_1|H_A^{MMS}\right) = \frac{4n + 2}{n^2 - n}\left(\sigma_r^2 + (\varepsilon - \varepsilon^2)e_g^2\right)$$

$$\mathrm{var}\left(\hat{\underline{x}}_2|H_A^{MMS}\right) = \frac{12}{n^3 - n}\left(\sigma_r^2 + (\varepsilon - \varepsilon^2)e_g^2\right)$$

With these expressions we find

$$\mathrm{MSE}\left(\hat{\underline{x}}_1|H_A^{SMS}\right) = \mathrm{var}\left(\hat{\underline{x}}_1|H_A^{SMS}\right) + \mathrm{bias}\left(\hat{\underline{x}}_1|H_A^{SMS}\right)^2 = \frac{4n + 2}{n^2 - n}\sigma_r^2 + 4\left(\frac{2n + 1 - 3k}{n^2 - n}\right)^2 e_g^2 \tag{24}$$

$$\mathrm{MSE}\left(\hat{\underline{x}}_2|H_A^{SMS}\right) = \mathrm{var}\left(\hat{\underline{x}}_2|H_A^{SMS}\right) + \mathrm{bias}\left(\hat{\underline{x}}_2|H_A^{SMS}\right)^2 = \frac{12}{n^3 - n}\sigma_r^2 + 36\left(\frac{2k - n - 1}{n^3 - n}\right)^2 e_g^2 \tag{25}$$

$$\mathrm{MSE}\left(\hat{\underline{x}}_1|H_A^{MMS}\right) = \mathrm{var}\left(\hat{\underline{x}}_1|H_A^{MMS}\right) + \mathrm{bias}\left(\hat{\underline{x}}_1|H_A^{MMS}\right)^2 = \frac{4n + 2}{n^2 - n}\left(\sigma_r^2 + (\varepsilon - \varepsilon^2)e_g^2\right) + \varepsilon^2 e_g^2 \tag{26}$$

$$\mathrm{MSE}\left(\hat{\underline{x}}_2|H_A^{MMS}\right) = \mathrm{var}\left(\hat{\underline{x}}_2|H_A^{MMS}\right) + \mathrm{bias}\left(\hat{\underline{x}}_2|H_A^{MMS}\right)^2 = \frac{12}{n^3 - n}\left(\sigma_r^2 + (\varepsilon - \varepsilon^2)e_g^2\right) \tag{27}$$

9.3 MSEs of the least squares estimates in the variance inflation model

Since (16) has expectation μ (Lehmann 2012a), we clearly see that

$$E\left\{\underline{\boldsymbol{y}}|H_A^{SVI}\right\} = E\left\{\underline{\boldsymbol{y}}|H_A^{MVI}\right\} = E\left\{\underline{\boldsymbol{y}}|H_0\right\} = \boldsymbol{A}\boldsymbol{x}$$

Consequently, $\hat{\underline{\boldsymbol{x}}}|H_A^{SVI}$ and $\hat{\underline{\boldsymbol{x}}}|H_A^{MVI}$ are unbiased estimates, but have different variances. For H_A^{SVI} the covariance matrix of $\underline{\boldsymbol{y}}$ is

$$\boldsymbol{\Sigma}_{yy}|H_A^{SVI} = \mathrm{diag}\left(\sigma_r^2, \ldots, \sigma_r^2, \sigma_r^2 + \sigma_g^2, \sigma_r^2, \ldots, \sigma_r^2\right)$$

where σ_g^2 is added to the k-th diagonal element. By covariance propagation applied to (21) we obtain the covariance matrix of $\underline{\hat{x}}$ as

$$\boldsymbol{\Sigma}_{\hat{x}\hat{x}}|H_A^{SVI} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\sigma_r^2 + (\boldsymbol{A}^T\boldsymbol{A})^{-1}a_k^T a_k(\boldsymbol{A}^T\boldsymbol{A})^{-1}\sigma_g^2$$

This yields

$$\mathrm{MSE}\left(\underline{\hat{x}}_1\middle|H_A^{SVI}\right) = \mathrm{var}\left(\underline{\hat{x}}_1\middle|H_A^{SVI}\right) = \frac{4n+2}{n^2-n}\sigma_r^2 + 4\left(\frac{2n+1-3k}{n^2-n}\right)^2\sigma_g^2 \tag{28}$$

$$\mathrm{MSE}\left(\underline{\hat{x}}_2\middle|H_A^{SVI}\right) = \mathrm{var}\left(\underline{\hat{x}}_2\middle|H_A^{SVI}\right) = \frac{12}{n^3-n}\sigma_r^2 + 36\left(\frac{2k-n-1}{n^3-n}\right)^2\sigma_g^2 \tag{29}$$

By simple probability calculus we find that the variance of (16) equals $\sigma_r^2 + \varepsilon\sigma_g^2$ (Lehmann 2012a), which yields

$$\boldsymbol{\Sigma}_{yy}|H_A^{MVI} = \left(\sigma_r^2 + \varepsilon\sigma_g^2\right)\boldsymbol{I}$$

$$\boldsymbol{\Sigma}_{\hat{x}\hat{x}}|H_A^{MVI} = \left(\sigma_r^2 + \varepsilon\sigma_g^2\right)(\boldsymbol{A}^T\boldsymbol{A})^{-1}$$

$$\mathrm{MSE}\left(\underline{\hat{x}}_1\middle|H_A^{MVI}\right) = \mathrm{var}\left(\underline{\hat{x}}_1\middle|H_A^{MVI}\right) = \frac{4n+2}{n^2-n}\left(\sigma_r^2 + \varepsilon\sigma_g^2\right) \tag{30}$$

$$\mathrm{MSE}\left(\underline{\hat{x}}_2\middle|H_A^{MVI}\right) = \mathrm{var}\left(\underline{\hat{x}}_2\middle|H_A^{MVI}\right) = \frac{12}{n^3-n}\left(\sigma_r^2 + \varepsilon\sigma_g^2\right) \tag{31}$$

9.4 Supplement

It is illustrative to observe when the least squares estimate (21) is least distorted by the outliers, i.e. when

$$\mathrm{MSE}\left(\underline{\hat{x}}_i\middle|H_A\right) \approx \mathrm{MSE}\left(\underline{\hat{x}}_i\middle|H_0\right)$$

apart from the trivial cases $e_g \approx 0$ or $\sigma_g \approx 0$ or $\varepsilon \approx 0$.

For slippage alternatives this would happen for the intercept parameter x_1 if the outlier occurs at $k \approx (2n+1)/3$, cf. (24),(28), while for the slope parameter x_2 this is obtained in the center of the observations, where $k \approx (n+1)/2$, cf. (25),(29). E.g. for $n = 7$ and $k = 5$ the intercept parameter $x_1$would be completely unaffected by an outlier while for the slope parameter x_2 this would happen at $k = 4$.

Mixture alternatives do not show such a behavior, except for (27) at the theoretical value $\varepsilon = 1$, i.e. all observations are affected by the same mean shift e_g, which understandably leaves the slope invariant. But how can an estimate be best if **all** observations are outliers? This surprising behavior reveals a weakness of the mean shift model in describing the geodetic reality.

9.5 Mean MSEs for the slippage alternative

Obviously, the MSEs for H_A depend on the nuisance parameter(s) other than $\boldsymbol{x}$, i.e. either e_g or σ_g and in the case of the slippage alternatives also k. This is undesired because $Prot$ in (18),(20) depends on those unknown parameters. In order to get rid of the nuisance parameter k in the slippage alternatives we assume that every observation has the same probability to be affected by gross error. In this way we introduce the mean MSE with respect to k, symbolically

$$\overline{MSE} = \frac{1}{n}\sum_{k=1}^{n} MSE(k)$$

After some simple calculus we arrive at quite compact formulae:

$$\overline{MSE}\left(\underline{\hat{x}}_1\middle|H_A^{SMS}\right) = \frac{4n+2}{n^2-n}\left(\sigma_r^2 + e_g^2/n\right) \tag{32}$$

$$\overline{MSE}\left(\underline{\hat{x}}_2\middle|H_A^{SMS}\right) = \frac{12}{n^3-n}\left(\sigma_r^2 + e_g^2/n\right) \tag{33}$$

$$\overline{MSE}\left(\underline{\hat{x}}_1\middle|H_A^{SVI}\right) = \frac{4n+2}{n^2-n}\left(\sigma_r^2 + \sigma_g^2/n\right) \tag{34}$$

$$\overline{MSE}\left(\underline{\hat{x}}_2\middle|H_A^{SVI}\right) = \frac{12}{n^3-n}\left(\sigma_r^2 + \sigma_g^2/n\right) \tag{35}$$

Still these terms are not fully computable, but depend on the unknown nuisance parameter e_g or σ_g. Getting rid also of this dependence would mean to introduce a probability distribution also for those parameters as done e.g. by Möller (1972). But such a choice would always be disputable.

9.6 Estimates after application of the rejection rule

The rival estimator $\underline{\hat{x}}'$ to be applied if H_0 is rejected works here as follows: After computing (21) we derive the residuals $\underline{\hat{e}}$ and carry out a local test only. (The global test would not be good to perform here because it is not optimal for the chosen H_A, see section 7.) If $T_n > c$ for some critical value c then H_0 is rejected. We discard the observation with the extreme normalized residual T_n and compute (21) with the remaining $n-1$ observations. Thus, estimator $\underline{\hat{x}}'$ either equals $\underline{\hat{x}}$ or the corresponding least squares estimate with $n-1$ observations.

For slippage alternatives the terms of the form $\mathrm{MSE}(\hat{x}_i'|H_A)$ will in addition depend on the unknown nuisance parameters k and on either e_g or σ_g. The dependence on k can be removed by averaging as before, getting terms of the form $\overline{MSE}(\hat{x}_i'|H_A)$.

As pointed out in section 8, those terms must be evaluated numerically by the MC method. To be on the safe side, the number of MC experiments is here chosen to be 10^6. This is much more than needed, as can be demonstrated by reproducing exactly the same results with different pseudo random numbers. In this small scale model we can afford the computational costs of such a brute force approach, but in general the number of MC experiments should be chosen with care.

10. Results

10.1 Settings

Here we intend to demonstrate how the alternative hypothesis influences the performance of the outlier detection in terms of premium and protection. We compute premium and protection by (17),(18) for the intercept parameter x_1 and for the slope parameter x_2 separately. A joint computation by (19),(20) would not make sense here because slope and intercept have different units, see discussion in section 8.

In Fig. 4, 5 and 6 we display the results of premium and protection for $n = 10$ observations as a function of the critical value c. As stated before, the strict relationship between c and α is nontrivial (see Lehmann 2012b), but can be approximated by (7),(8). In Fig. 4 slippage alternatives is used while Fig. 5 and 6 employ mixture alternatives. For the slippage alternatives we only display the premium and protection for the slope parameter x_2 (Fig. 4). It turns out that the corresponding values of the intercept parameter x_1 are so much the same such that related curves would largely overlap, if plotted together in Fig. 4. Even Fig. 5 and Fig. 6 show highly visible similarities. This indicates that an outlier detection performing well for one parameter also performs well for another.

10.2 Premium

The premium is independent of H_A and is therefore identical in Fig. 4 and 5 for slope parameter x_2. In Fig. 6 it is given for intercept parameter x_1, but the difference of all premiums is negligible. The premium is increasing with α and consequently decreasing with c. If one wants to pay a premium of at most 10% then one has to obey $c > 2.4$ in this example. For $c > 3$ there is practically no premium anymore because then a true H_0 is very rarely rejected.

10.3 Protection

The protection depends on H_A and also on the nuisance parameters e_g or σ_g therein. Since the sign of e_g is unimportant for the protection, we restrict ourselves to positive e_g. It is clearly seen in Fig. 4-6 that the larger the gross errors in terms of either e_g or σ_g the better the protection. This is a trivial result: Protection against large gross errors is more effective than against small ones.

For outliers caused by small gross errors the local test using the extreme normalized residual T_n as a test statistic yields no protection at all. $Prot$ in (18) even becomes negative. This means that the rejection of

the outlier makes the estimation worse. In Fig. 4 we see that for H_A^{SMS} the magnitude of the gross error must be $|e_g| \geq 3\sigma_r$ while for H_A^{SVI} it must be $\sigma_g \geq 2\sigma_r$ to reach an operable protection provided by $\hat{x}_i'$, i.e. Prot>0 in (18). From Fig. 5 we conclude that for H_A^{MMS} and H_A^{SVI} the corresponding limits are $|e_g| \geq 5\sigma_r$ and $\sigma_g \geq 2\sigma_r$. Fortunately, it is rather unimportant to get protection against outliers caused by small gross errors, but most of all for H_A^{MMS} the protection is not satisfactory. The reason is that here with a probability of about 0.26 there are multiple gross errors of equal size e_g in $\underline{\boldsymbol{y}}$. But at most one outlier is discarded. This is less dramatic for H_A^{MVI}: Although multiple gross errors occur with the same probability, they are of random size, which makes it likely that at least the outlier caused by the extreme gross error is discarded. It may be surprising that the protection against larger gross errors can be worse than for smaller ones, see H_A^{MMS} in Fig. 5. This behavior can be explained by multiple gross errors masking each other. At this point there is a notable difference between the slope parameter x_2 in Fig. 5 and the intercept parameter x_1 in Fig. 6: Masking is worse for the slope parameter, it cannot occur in the slippage alternative used here because of $n_2 = 1$. This behavior is a pre-stage of the peculiarity explained at the end of subsection 9.4. If a good protection can be obtained in the mean shift model then it is good also for small c. In other words: Even smaller T_n still indicate the correct outlier to be rejected. In contrast to this, in the variance inflation model there is an optimal protection here at $c \approx 3$. This is explained as follows: Even for large gross error variance it happens that some realizations of gross errors are small. If c is small then they are wrongly detected.

10.4 Optimization

Fig. 4-6 can be used as a starting point for the optimization of outlier detection. Most of all if H_A^{MMS} applies, it is necessary to employ a different test statistic in order get an effective protection already at $|e_g| < 5\sigma_r$. This is beyond the scope of this paper. If $\underline{T_n}$ yields a good protection then we can find the optimal critical value (here $c \approx 3$, where at the same time the premium is low). Moreover, it becomes evident that here the choice of c is not too decisive.

Note that the values derived from Fig. 4-6 cannot be assumed to hold in general. For any other observation model the computations must be repeated. For example, the optimal critical value would tend to increase with n as suggested by (8).

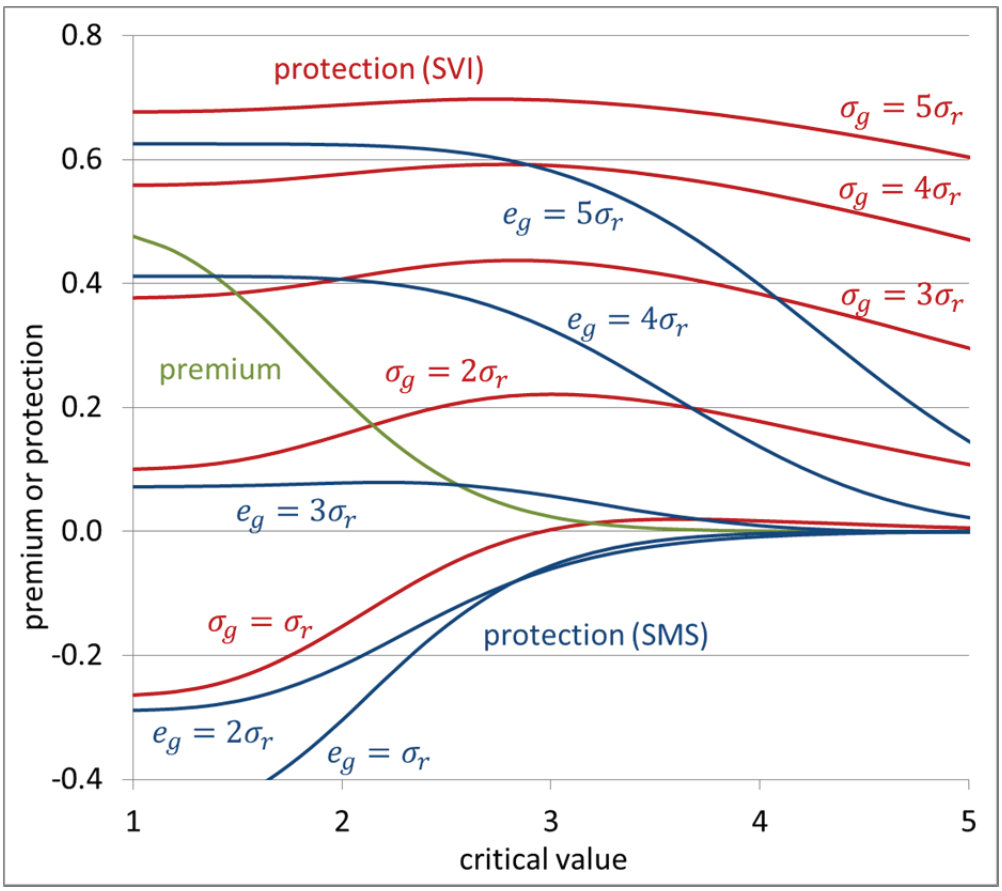

Fig. 4 Premium (green) and protection of slippage alternatives (S) in the mean shift model (MS, blue) and in the variance inflation model (VI, red) for the slope parameter of a straight line fit through 10 equidistant data points versus critical value c of the extreme normalized residual T_n

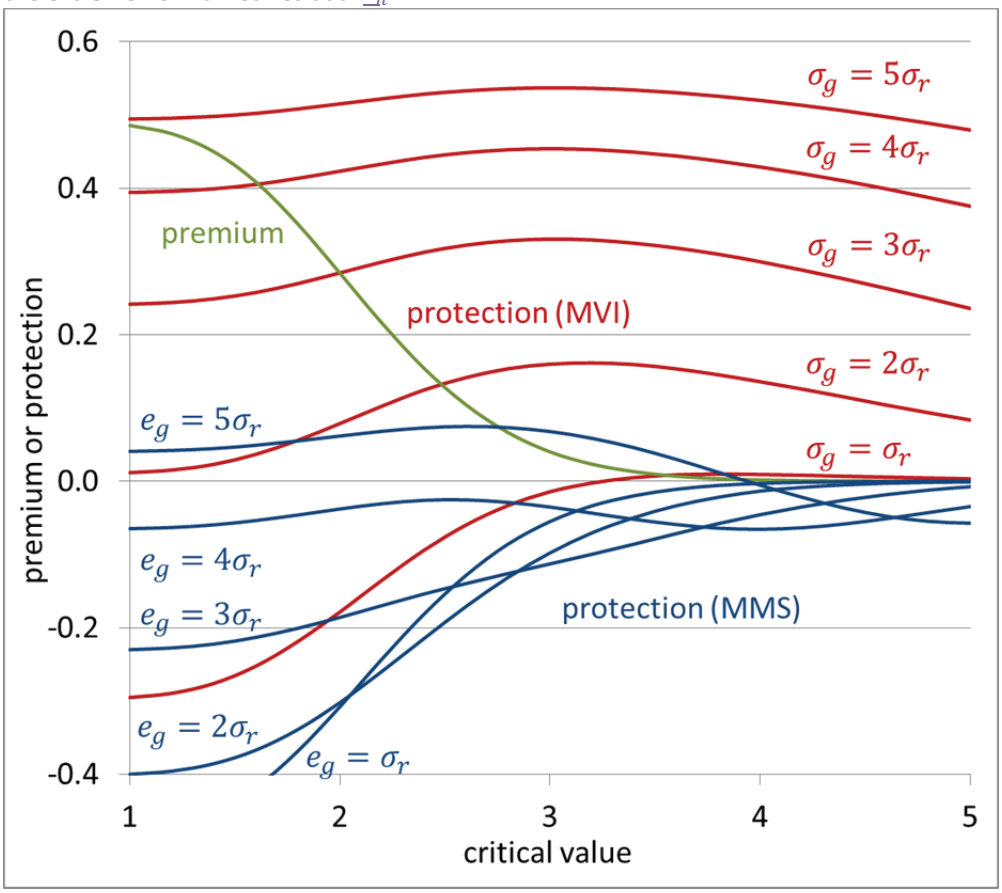

Fig. 5 Same as Fig. 4, but mixture alternatives instead of slippage alternatives

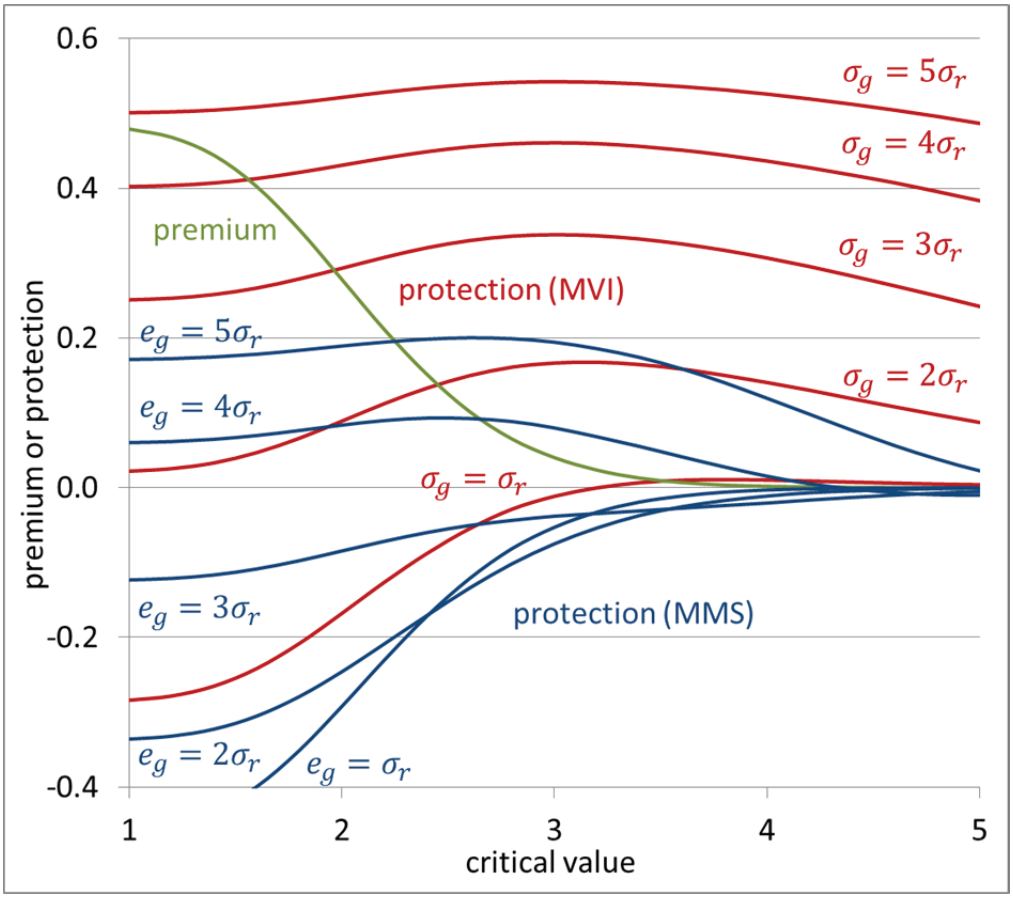

Fig. 6 Same as Fig. 5, but intercept parameter instead of slope parameter

11. Conclusions

If one is satisfied with a plausible test statistic for outlier detection such as extreme normalized or studentized residuals and with a intuitive guess or experience-based choice of the significance level α or equivalently of the critical value c then there is no need to invoke any alternative hypothesis H_A. This is why its importance is often ignored. But if one desires to measure the performance of outlier detection or even to optimize it in any way then it is necessary to decide on an appropriate H_A. Besides the power, premium and protection are very well-suited, even better-suited measures of performance for outlier tests.

There is a substantial wealth of forms of possible H_A. This gives the user the great flexibility to formulate his H_A on the basis of his experiences. Whatever is known about location, number and stochastic behavior of the gross errors causing the outliers to be detected can be incorporated and, no less important, whatever is not known can be omitted. However, only a fraction of possible H_A has even been considered, namely those, for which UMPI test statistics can be theoretically derived. It is not guaranteed that these test statistics work also for other H_A.

In former times, when critical values had to be looked up in statistical tables, it was only possible to formulate H_0 and H_A and to chose $\underline{T}$ and α in such a way that an appropriate lookup-table for the corresponding critical value is available. This substantially restricted the freedom of choice and the possibility of optimization. But since powerful computers are available everywhere, it is no longer forbidden to use the Monte Carlo method for computing measures of performance and for optimizing outlier tests. One can even dispense with the derivation of the analytical formulas given in section 9, because $\mathrm{MSE}(\underline{\hat{x}}_i|H)$ and $\overline{\mathrm{MSE}}(\underline{\hat{x}}_i|H)$ can as well be computed by the Monte Carlo method in the same way as $\mathrm{MSE}(\underline{\hat{x}}_i'|H)$ and $\overline{\mathrm{MSE}}(\underline{\hat{x}}_i'|H)$.

In the considered practical example it turned out that actually very similar alternative hypotheses show different performances. Hence, an outlier test optimized for one H_A may not be suited for another. The user interested in a good performance of the outlier detection for his problem has to repeat the

computations performed in section 9 on its own problem. He can try other test statistics and rejection rules.

Acknowledgement: This work has been completed while the author spends his sabbatical at Technische Universität Berlin with Prof. Dr.-Ing. Frank Neitzel as his host. The support is gratefully acknowledged.

References

Abdi H (2007) The Bonferonni and Šidák Corrections for Multiple Comparisons. In: Neil Salkind (Ed.) Encyclopedia of Measurement and Statistics. Sage Thousand Oaks (CA)

Anscombe FJ (1960) Rejection of outliers. Technometrics 2(2):123-147

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

Barnett V, Lewis T (1994) Outliers in statistical data, John Wiley, ISBN 0-471-93094-6,Chichester

Clerici E, Harris MW (1980) A Premium-Protection Method Applied to Detection and Rejection of Erroneous Observations. Manuscripta Geodaetica, Vol. 5, pp. 282-298.

Clerici E, Harris MW (1983) A review of the premium-protection method and its possible application in detection of displacements J Geod 57 (1-4): 1-9, DOI: 10.1007/BF02520908

Fan H (2010) Theory of Errors and Least Squares Adjustment. Royal Institute of Technology (KTH), Division of Geodesy and Geoinformatics Stockholm (Sweden), Geodesy Report No. 2015, ISBN 91-7170-200-8

Gui Q, Li X, Gong Y, Li B, Li G A (2011) Bayesian unmasking method for locating multiple gross errors based on posterior probabilities of classification variables J Geod 85:191–203

Hawkins D (1980) Identification of Outliers. Chapman and Hall London New York

Hekimoglu S, Koch KR (2000) How can reliability of the test for outliers be measured? Allgemeine Vermessungsnachrichten. VDE Verlag Berlin Offenbach, S. 247-253

Kargoll B (2012) On the Theory and Application of Model Misspecification Tests in Geodesy. Deutsche Geodätsche Kommission Reihe C, Nr. 674, München

Koch KR (1999) Parameter Estimation and Hypothesis Testing in Linear Models. Springer Verlag Berlin Heidelberg New York

Koch KR (2007) Introduction to Bayesian statistics. 2nd edn. Springer, Berlin

Lehmann R (2010) Normalized residuals – how large is too large? (in German). Allgemeine Vermessungsnachrichten Vol. 2/2010 53-61, VDE Verlag Berlin Offenbach

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

Lehmann R (2012a) Geodetic error calculus by the scale contaminated normal distribution (in German). Allgemeine Vermessungsnachrichten Vol. 5/2012, 143-149, VDE Verlag Berlin Offenbach

Lehmann R (2012b) Improved critical values for extreme normalized and studentized residuals in Gauss-Markov models. J Geod 86:1137–1146. DOI: 10.1007/s00190-012-0569-0

Möller HP (1972) Ausreißer in Stichproben aus normalverteilten Grundgesamtheiten. PhD Thesis, Cologne University

Mood AM, Graybill FA, Boes DC (1974) Introduction to the Theory of Statistics, McGraw-Hill Kogakusha, Tokyo

Nadarajah S (2005) A generalized normal distribution, Journal of Applied Statistics 32 (7) 685–694. DOI: 10.1080/02664760500079464.

Neumann I, Kutterer H, Schön St (2006) Outlier Detection in Geodetic Applications with respect to Observation Imprecision. Proceedings of the NSF Workshop on Reliable Engineering Computing - Modeling Errors and Uncertainty in Engineering Computations-. Savannah (Georgia), USA, pp. 75-90.

Pope AJ (1976) The statistics of residuals and the detection of outliers. NOAA Technical Report NOS65 NGS1, US Department of Commerce, National Geodetic Survey Rockville, Maryland

Rousseeuw PJ, Leroy AM (2003) Robust Regression and Outlier Detection. John Wiley & Sons, New Jersey, ISBN: 978-0471488552
Teunissen PJG (2000) Testing theory; an introduction. 2nd edition. Series on Mathematical Geodesy and Positioning, Delft University of Technology, The Netherlands. ISBN-13 978-90-407-1975-2
Thompson W (1935) On the Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation. Annals of Mathematical Statistics, Vol. 6, pp. 214—219
Yang Y (1991) Robust Bayesian estimation. Bull Geod 65(3):145–150
Yang Y (1999) Robust estimation of geodetic datum transformation. J Geod 73:8-274

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