J Geod (2012) 86:1137–1146
DOI 10.1007/s00190-012-0569-0

ORIGINAL ARTICLE

Improved critical values for extreme normalized and studentized residuals in Gauss–Markov models

Rüdiger Lehmann

Received: 1 December 2011 / Accepted: 18 April 2012 / Published online: 9 May 2012

Abstract We investigate extreme studentized and normalized residuals as test statistics for outlier detection in the Gauss–Markov model possibly not of full rank. We show how critical values (quantile values) of such test statistics are derived from the probability distribution of a single studentized or normalized residual by dividing the level of error probability by the number of residuals. This derivation neglects dependencies between the residuals. We suggest improving this by a procedure based on the Monte Carlo method for the numerical computation of such critical values up to arbitrary precision. Results for free leveling networks reveal significant differences to the values used so far. We also show how to compute those critical values for non-normal error distributions. The results prove that the critical values are very sensitive to the type of error distribution.

Keywords Outlier detection · Gauss–Markov model · Hypothesis testing

1 Introduction

One of today's major challenges of geodetic data analysis is gross error detection. In geodesy, a gross error is a measurement deviation that is assumed to be generated by a stochastic process of a significantly different characteristic from what is assumed in the stochastic model of the parameter estimation. Typically, the statistical dispersion of this process is much larger. Gross errors are also referred to as outliers in applied statistics. There are some more or less tentative definitions of an outlier. We quote one of the most popular definitions given be Hawkins (1980): "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". This is very much equivalent to a gross error in geodesy. We do not dwell on this terminological issue, but use the term "outlier" throughout the rest of the paper.

The quality of parameter estimation using the classical method of least squares is very much affected by outliers. In the last few decades, a toolbox of robust estimation methods has evolved. They are able to produce reasonable results even if the observations contain some outliers. We do not go into details, but refer to Andersen (2008) for details.

In geodesy, we often try to identify outliers in the framework of a statistical hypothesis testing and down-weight or remove them from the input data set. If the number of outliers is small and if our data set is sufficiently redundant, then we may not lose a significant amount of information. But we retain the advantage that we can employ simple and efficient tools of geodetic data analysis like classical least squares estimation. In the following, we exclusively pursue this approach.

The best-established method for identification of outliers in geodetic data analysis is *data snooping*. This method by Baarda (1968) was later extended by Pope (1976) to the case that the accuracy of the observations is a priori unknown. Although data snooping was first introduced for the adjustment of geodetic networks, it is a generally applicable method. First, we perform the so-called *global test* of the model. Here, we use the weighted sum of squares of the least squares residuals as a test statistic, i.e. as a value indicating outliers or other inconsistencies in the functional or stochastic model. If this test statistic is smaller than some critical value, i.e. a quantile value of its probability distribution, then we accept the hypothesis that the observations do not contain outliers.

R. Lehmann (✉)
Faculty of Spatial Information, Dresden University of Applied Sciences, Friedrich-List-Platz 1, 01069 Dresden, Germany
e-mail: r.lehmann@htw-dresden.de

Otherwise, and if other deficiencies of the functional and stochastic model can be ruled out, we have to detect and eliminate one or more outliers. In order to actually find the outlying observation, we perform the individual or *local test*, where the individual residuals are considered. The aim of this paper is to compute improved critical values of the test statistics used here.

Since data snooping is based on a statistical hypothesis testing, it may lead to a false decision as follows:

Type I error	Rejection of a true hypothesis	Probability level α
Type II error	Acceptance of a false hypothesis	Probability level β

The computation of the critical values of the test statistics for this test makes a severe neglect, which is no longer necessary to make, when fast computers are available. We will show how to improve the computation of such critical values.

Lehmann (2010) and Lehmann and Scheffler (2011) pose the problem how to determine the optimal levels of type I error probabilities for global and local tests in data snooping. If these levels are chosen too low, then we get too large critical values and many outliers remain undetected. If on the contrary these levels are chosen too high, then we get too small critical values and it is likely that good observations are eliminated. How do we strike a balance between these impairments of parameter estimation? In the papers quoted above, it is shown how to use the Monte Carlo method for this purpose.

A different approach introduces fuzzy sets for outlier detection (Aliosmanoglu and Akyilmaz 2001; Wieser 2001). The set of outliers is considered as a fuzzy set and the membership function is derived from the difference between test statistic and critical value. Also, here improved critical values of the test statistic would be needed.

The outline of the paper is as follows. First, we derive extreme normalized and studentized residuals as test statistics for outliers in the Gauss–Markov model, possibly not of full rank. Then, we show how to derive critical values by means of the Monte Carlo method. Next, we study three standard geodetic Gauss–Markov models. It will turn out that there are significant differences between the classical critical values and the related quantities computed by the Monte Carlo method.

2 Gauss–Markov model not of full rank

To relate the parameters to the observations, we begin with the familiar Gauss–Markov model in the linearized form:

$$y = A\xi + e, \quad E\{e\} = 0, \quad D\{e\} = \sigma_0^2 P^{-1} \tag{1}$$

y is the $n \times 1$-vector of known observations, ξ is the $u \times 1$-vector of unknown parameters, A is the $n \times u$-matrix relating observations and parameters. e is the $n \times 1$-vector of unknown observations errors. E and D denote the operators of expectation and dispersion. P is the known $n \times n$-matrix of weights. σ_0^2 is a variance factor, which may or may not be known. e is a vector of random variates and so is y.

For the sake of universal applicability, the system may or may not be of full rank:

$$\text{rank}(A) = q \leq u$$

The least squares solution for the estimated parameters reads

$$\hat{\xi} = (A^{\mathrm{T}} P A)^{-} A^{\mathrm{T}} P y \tag{2}$$

with superscript "−" denoting some generalized inverse matrix. If $q = u$, this solution is also a best linear unbiased estimate, otherwise $\hat{\xi}$ may not be unique. The residuals (estimated observation errors) are always unique and we obtain them by

$$\begin{aligned}\hat{e} &= y - A\hat{\xi} \\ &= y - A(A^{\mathrm{T}} P A)^{-} A^{\mathrm{T}} P y \\ &= A\xi + e - A(A^{\mathrm{T}} P A)^{-} A^{\mathrm{T}} P(A\xi + e) \\ &= e - A(A^{\mathrm{T}} P A)^{-} A^{\mathrm{T}} P e = Re \end{aligned} \tag{3}$$

This rewriting invokes a property of generalized inverses (cf. Koch 1999). R is known as the redundancy matrix. The dispersion of the residuals reads

$$\begin{aligned} D\{\hat{e}\} &= \sigma_0^2 (P^{-1} - A(A^{\mathrm{T}} P A)^{-} A^{\mathrm{T}}) \\ &= \sigma_0^2 R P^{-1} = \sigma_0^2 Q_{\hat{e}\hat{e}} \end{aligned} \tag{4}$$

where $Q_{\hat{e}\hat{e}}$ denotes the cofactor matrix of the residuals. Therefrom, we may derive an unbiased estimate of the variance factor σ_0^2 as

$$\hat{\sigma}_0^2 = \frac{\hat{e}^{\mathrm{T}} P \hat{e}}{n - q} \tag{5}$$

All estimates $\hat{\xi}, \hat{e}, \hat{\sigma}_0^2$ are functions of e and consequently random variates or random vectors.

3 Normalized and studentized residuals as test statistics for outliers

For hypothesis testing, it is important to assume a distribution of the observation errors e not being outliers. This is typically a central Gaussian distribution:

$$e \sim N(0, \sigma_0^2 P^{-1}) \tag{6}$$

Under this assumption, the least squares solution (2) is also a maximum likelihood solution (Koch 1999).

A test statistic is a quantity which assumes extreme values in the case that a certain null hypothesis H_0 is not fulfilled. In the case of outlier detection, this null hypothesis is:

H_0: No observation $y_i, i = 1, \dots, n$ is affected by outliers.

A possible alternative hypothesis is

$H_A^{(i)}$: The observation y_i for some fixed i is an outlier.

According to the pioneering work of Baarda (1968), the decision can be based on the value of the *normalized residual*

$$w_{i,\text{norm}} = \frac{\hat{e}_i}{\sigma_0 \sqrt{q_{e_i e_i}}} \tag{7}$$

as a test statistic, where $q_{e_i e_i}$ denotes the ith diagonal element of $Q_{\hat{e}\hat{e}}$ in (4). $w_{i,\text{norm}}$ is by definition (7) a random variate. If H_0 holds true, then $w_{i,\text{norm}} \sim N(0,1)$ is easily derived. For example, individual values of $|w_{i,\text{norm}}| > 1.96$ occur only with a probability of $\alpha = 0.05$. If we find $|w_{i,\text{norm}}| > 1.96$ for some y_i, we may reject H_0. This induces a probability of type I decision error (rejecting H_0 if it is true) of $\alpha = 0.05$. Other critical values are 1.64 for $\alpha = 0.10$ and 2.58 for $\alpha = 0.01$. In general, the critical value reads

$$c_{\text{norm}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{8}$$

where Φ denotes the cumulative distribution function (cdf) of $N(0,1)$. Note that we perform a two-sided test of the form $|w_{i,\text{norm}}| \le c_{\text{norm}}$. That is why we have $\alpha/2$ in (8).

If σ_0 is not known, we may use a test statistic of the form

$$w_{i,\text{stud}} = \frac{\hat{e}_i}{\hat{\sigma}_0 \sqrt{q_{e_i e_i}}} \tag{9}$$

suggested by Pope (1976). It is known as *studentized residual*. $w_{i,\text{stud}}$ is by definition (9) a random variate. If H_0 holds true, it has been shown by Pope (1976) that $w_{i,\text{stud}} \sim \tau(1, n-q-1)$. This distribution is known as τ-distribution with 1 and $n-q-1$ degrees of freedom.

Quantiles of the τ-distribution can be computed from quantiles of Fisher's F-distribution, or in our case, where the first degree of freedom is 1, also from Student's t distribution. In general, the critical value for (9) can be computed as

$$c_{\text{stud}} = \sqrt{\frac{(n-q)t^2_{n-q-1}(\alpha/2)}{n-q-1+t^2_{n-q-1}(\alpha/2)}} \tag{10}$$

where t denotes the quantile of the Student's t distribution with $n-q-1$ degrees of freedom (Pope 1976). Note that we perform a two-sided test of the form $|w_{i,\text{stud}}| \le c_{\text{stud}}$. That is why we have $\alpha/2$ in (10). From (10) we deduce

$$0 \le c_{\text{stud}} < \sqrt{n-q}. \tag{11}$$

4 Testing against *n* alternative hypotheses

Usually it is not known which observation y_i may be an outlier. Therefore, a more appropriate alternative hypothesis would be

H_A: There is at least one outlier in the vector of observations y.

Since $H_A = H_A^{(1)} \vee H_A^{(2)} \vee \dots \vee H_A^{(n)}$ where each $H_A^{(i)}$ denotes an alternative hypothesis from the preceding section, this is equivalent to testing H_0 against a sequence of alternative hypotheses

$$H_A^{(1)}, \dots, H_A^{(n)}$$

Let $w_i, i = 1, \dots, n$ denote either normalized or studentized residuals from (7) or (9). H_0 is rejected if $|w_i|$ exceeds a critical value c for any $i = 1, \dots, n$. Otherwise, it is accepted. Therefore, the test statistic coming into effect is

$$w = \max_{i=1,\dots,n} |w_i| \tag{12}$$

A true H_0 is rejected after n tests, if it is rejected in any of the n tests. The probability that a true H_0 is accepted in test i is $1-\alpha, i = 1, \dots, n$. If there is an outlier in any observation and since R in (3) is not a diagonal matrix, it is likely that more than one residual is large. Therefore, H_0 is likely to be rejected in more than one test, i.e. the test results depend on each other to some degree. For the sake of simplicity, those dependencies are usually neglected. Then the probability that a true H_0 is accepted in each test is approximately $(1-\alpha)^n$. Hence, the probability α' that a true H_0 is rejected in any test would be

$$\alpha' = 1 - (1-\alpha)^n \tag{13}$$

Since α is small, we get $\alpha' \approx n\alpha$. This is accounted for by computing the critical value c either by

$$c_{\text{norm}} = \Phi^{-1}\left(1 - \frac{\alpha'}{2n}\right) \tag{14}$$

or by

$$c_{\text{stud}} = \sqrt{\frac{(n-q)t^2_{n-q-1}(\alpha'/(2n))}{n-q-1+t^2_{n-q-1}(\alpha'/(2n))}} \tag{15}$$

if α' is the desired probability of type I decision error with respect to $H_A = H_A^{(1)} \vee H_A^{(2)} \vee \dots \vee H_A^{(n)}$. This procedure is well known and was suggested by Stefansky (1972) for the first time.

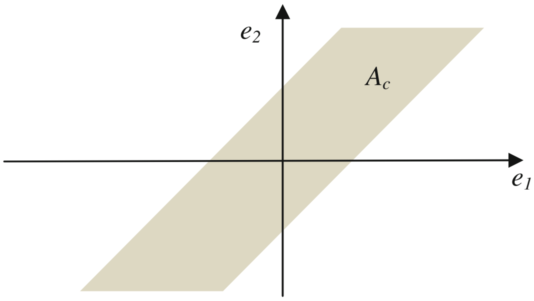

Fig. 1 Set of acceptance A_c in the simplest case of two independent repeated observations $y_1 = x + e_1$, $y_2 = x + e_2$ of identical variance, i.e. $A = (1 \quad 1)^T$ and $P = I$

5 Extreme normalized and studentized residuals as test statistics

As an alternative procedure, one could treat the extreme (i.e. maximum absolute) normalized or studentized residuals w in (12) directly as a test statistic. One advantage is that we do not need to neglect dependencies in the hypotheses as in (13). However, the distributions of the test statistic cannot be derived from well-known test distributions like normal or t-distribution anymore. Therefore, critical values cannot be taken from a statistical table, but must be computed numerically.

If F_w denotes the cdf of the test statistic w and α is a given probability of type I decision error, we get the desired critical value c by

$$c = F_w^{-1}(1 - \alpha)$$

Note that using (12) as a test statistic, we perform a one-sided test of the form $w \leq c$.

Since $w = w(e)$ is a function of a random vector e (observation errors), its distribution is propagated from (6). Let E^n denote the n-dimensional Euclidian space of observation errors e and let $A_c \subset E^n$ denote the subset of acceptance of H_0, i.e (Fig. 1).

$$A_c := \{e \in E^n : w(e) < c\} \tag{16}$$

The probability of $w(e) < c$ is the same as of $e \in A_c$. Therefore, values of the cdf F_w can be computed by

$$1 - \alpha = F_w(c) = \int\limits_{A_c} \varphi_e(e)\mathrm{d}e \tag{17}$$

where φ_e denotes the probability density function (pdf) of e. In the case of (6), this reads(19)

$$F_w(c) = \frac{\det(P)^{1/2}}{(2\pi)^{n/2}\sigma_0^n} \int\limits_{A_c} \exp\left(-\frac{e^T Pe}{2\sigma_0^2}\right)\mathrm{d}e \tag{18}$$

Let $I_c(e)$ denote the indicator function of A_c (i.e. assuming values 1 for $e \in A_c$ and 0 otherwise). We get

$$F_w(c) = \int\limits_{E^n} I_c(e) \cdot \varphi_e(e)\mathrm{d}e = E\{I_c\} \tag{19}$$

Since I_c is not a simple function, the integral in (19) cannot be evaluated analytically.

6 Monte Carlo approach

Monte Carlo methods are able to compute statistical quantities numerically. They are used whenever the functional relationships are analytically not tractable, as is the case for data snooping. The basic idea is to approximate probability distributions by frequency distributions of computer random experiments performed using pseudo random numbers. The convergence is only of the order of $m^{-1/2}$, if m is the number of experiments performed, but unlike other methods, this order does not depend on the dimension of the data space (Tanizaki 2004), i.e. the number of observations n in geodesy, which is typically very large.

Since the advent of fast computers, we use Monte Carlo methods in geodesy (e.g. Lehmann 1994; Alkhatib et al. 2009). Those methods have already been applied in outlier detection (e.g. Koch 2007; Lehmann and Scheffler 2011).

In essence, the expectation in (19) is approximated by the arithmetic mean

$$F_w(c) = E\{I_c\} \approx \mathrm{Mean}\{I_c\} = \frac{1}{m}\sum_{k=1}^{N} I_c(e_k) \tag{20}$$

computed by pseudo random vectors e_k, $k = 1, \ldots, m$ of the desired distribution.

According to (6), we chose the multivariate normal distribution for e. We apply the Box–Muller method (Box and Muller 1958), a common and efficient method for generation of normal pseudo random numbers. It needs uniform pseudo random numbers in the interval [0, 1) coming from a modern Mersenne twister generator with period $2^{19937} - 1$.

7 Computation procedure

Firstly, note that the distributions of normalized or studentized residuals are independent of σ_0. Consequently, the same holds true for w in (12). If we compute a pseudo random vector for e and we do not know σ_0, we may use any positive value instead. The value will be canceled in (7) or (9).

Secondly, normalized or studentized residuals depend on the observation errors e, but not on the true values $A\xi$ of the observed quantities. This has been shown in (3).

Thirdly, the joint distribution of the vector of normalized or studentized residuals depends on the elements of $Q_{\hat{e}\hat{e}}$. Only the effect of the main diagonal elements $q_{e_i e_i}$ is canceled by the denominator of (7) or (9). If we compute the absolute maximum w in (12), we retain the effect of the off-diagonal elements of $Q_{\hat{e}\hat{e}}$. Therefore, the distribution of w will in general depend on those elements and in this way on the matrices A and P, see (3), (4).

The following procedure yields *arbitrarily precise* approximations of the desired critical values of w in (12):

1. Compute a sequence of m pseudo random vectors e_k, $k = 1, \ldots, m$ of the desired distribution, e.g. from (6). m is known as the number of Monte Carlo experiments.
2. For each e_k, $k = 1, \ldots, m$ estimate residuals $\hat{e}_k$ by (3) and compute the test statistic w_k by (7) or (9) and (12). The frequency distribution of w_k is an approximation of the probability distribution of w.
3. For some critical value c, we would get $I_c(e_k) = 1$ for $w_k < c$ and $I_c(e_k) = 0$ otherwise. Hence, the desired mean$\{I_c\}$ is the fraction of w_k smaller than c. c must be determined such that this fraction becomes $1 - \alpha$. For this purpose, sort vector w_k numerically, getting a sorted vector w'_k such that $w'_1 \leq w'_2 \leq \cdots \leq w'_m$. Determine the critical value
$$c = \frac{1}{2}\left(w'_{[(1-\alpha)m]} + w'_{[(1-\alpha)m]+1}\right) \tag{21}$$
where $[\bullet]$ denotes rounding down to the next integer. Note that this can be done for a sequence of values α in parallel.
4. In order to ensure that the approximate values (21) are close to the true values $F_w^{-1}(1-\alpha)$, one should observe the convergence of the procedure as $m \to \infty$ and implement a suitable termination criterion. We suggest that the procedure should be terminated as soon as c in (21) fluctuates by no more than 1 %.

Such a strategy has been used already by Lemeshko and Lemeshko (2005), but only for extreme studentized residuals and only for outlier detection in samples, which is equivalent to the case of repeated observations in geodesy. In Sects. 8–10, we will display the results for a less trivial geodetic Gauss–Markov model.

8 A free leveling network

Consider a free leveling network with square loops forming a checked pattern (see Fig. 2). Observations are uncorrelated leveled height differences of identical variance σ_0^2. We apply the Gauss–Markov model (1) with $P = I$ and A-matrix of the form

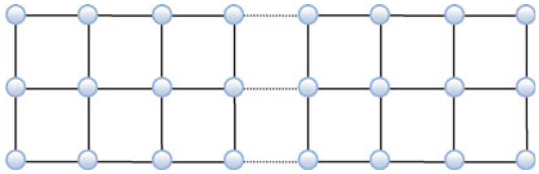

Fig. 2 Leveling network of variable size with square loops forming a checked pattern

$$A = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -1 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

For a free leveling network, we get $\operatorname{rank}(A) = q = u-1$. The right inverse in (2)–(4) is computed as usual by the constraint $\Sigma\xi = 0$ (see Koch 1999). Note that the actual observations are not required to compute improved critical values (21).

Example In the case of two loops and an "off the reel" ordering of the observations, we get

$$A = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} \tag{22}$$

The cofactor matrix of the residuals reads

$$Q_{\hat{e}\hat{e}} = \begin{pmatrix} +0.27 & +0.27 & +0.27 & -0.07 & -0.07 & -0.07 \\ +0.27 & +0.27 & +0.27 & -0.07 & -0.07 & -0.07 \\ +0.27 & +0.27 & +0.27 & -0.07 & -0.07 & -0.07 \\ +0.20 & +0.20 & +0.20 & +0.20 & +0.20 & +0.20 \\ -0.07 & -0.07 & -0.07 & +0.27 & +0.27 & +0.27 \\ -0.07 & -0.07 & -0.07 & +0.27 & +0.27 & +0.27 \\ -0.07 & -0.07 & -0.07 & +0.27 & +0.27 & +0.27 \end{pmatrix} \tag{23}$$

First of all, we intend to investigate the convergence of the Monte Carlo procedure described in Sect. 7. Figure 3 shows how the approximate critical values (21) converge to their true value as m increases. The computation is performed for 2 × 3 leveling loops, where $n = 15$ and $u = 12$. At $m =$ 10,000, we observe only minor changes of c at a relative magnitude of about 1 %.

To be on the safe side, we decided to double the computational effort to $m =$ 20,000 throughout the following computations.

Next, we investigate the critical values of extreme normalized residuals for various sizes of networks starting from

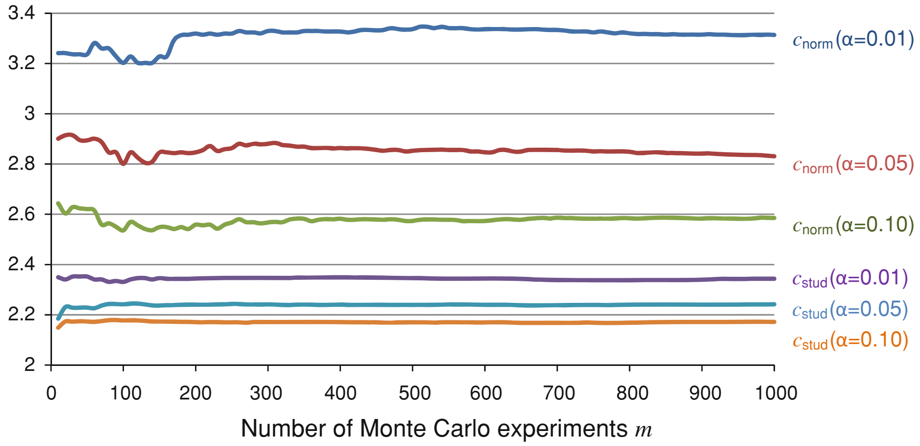

Fig. 3 Critical values (21) of extreme normalized and studentized residuals for 2 × 3 leveling loops versus number of Monte Carlo experiments m

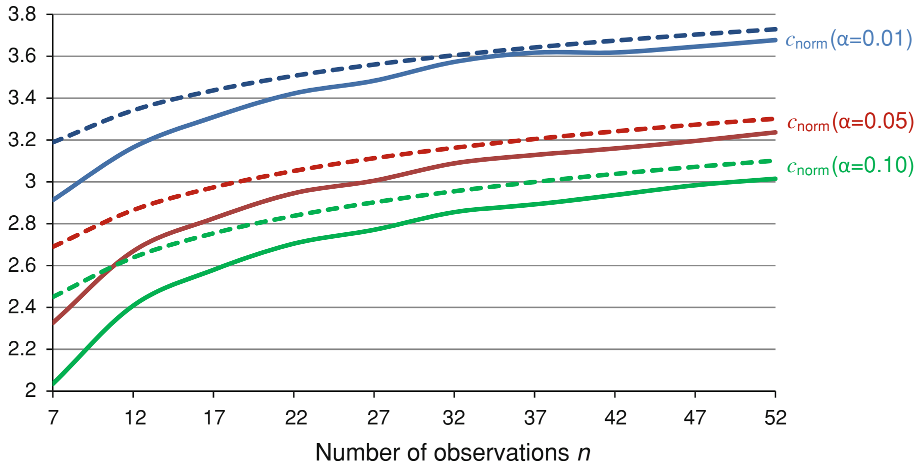

Fig. 4 Critical values of extreme normalized residuals for the free leveling network. *Solid curves* obtained from (21) by $m = 20{,}000$ Monte Carlo experiments. *Dashed curves* obtained from (14)

2 × 1 loops. Adding two new loops increases the numbers of observations $n \times 5$ and the number of parameters $u \times 3$. The result is displayed in Fig. 4. We observe that the improved critical values (21) are always smaller than those values computed by (14) neglecting dependencies between the n tests. In other words, under the condition that H_0 is true, it is less likely to get large extreme normalized residuals than predicted by (14). For a leveling network, this can be understood as follows.

If there is according to (6) a large measurement error, then we get large loop misclosures in the adjoining loops and all residuals in these loops tend to be large in magnitude. If aligned in the same sense of direction, the residuals also tend to have the same sign. If all measurement errors happen to be small, then all residuals tend to be small. Hence, the residuals are correlated. In (13), we have neglected such correlations. In this case and if one residual is small, we still suppose that the neighboring residual in this loop can be large in magnitude. In this way using (14), we overestimate the probability of large extreme normalized residuals and the dashed curve in Fig. 4 is always above the solid curve. This line of reasoning applies to positive and negative correlations. (In fact, one can change positive into negative correlations and vice versa by changing the sign of one observation.) Therefore, the dashed curve is always above the solid curve in Fig. 4. The effect is less pronounced for larger networks, because the correlations attenuate with distance and in a large network almost all pairs of observations are spatially separated.

Consequently at the same level of error probability, the improved critical values (21) detect more outliers than those computed by (7).

We repeat this computation for extreme studentized residuals. The result is displayed in Fig. 5. First of all, we note that the critical values are smaller than for the extreme normalized residuals. This is comprehensible from definition (9). If a residual e_i is large, then it is probable that also $\hat{\sigma}_0$ in (5) is large, partly mitigating the effect in (9). Consequently, studentized residuals tend to be smaller than normalized residuals. But as the number of observations n increases, this effect more and more vanishes by averaging of the residuals in (5). Indeed, we observe in Figs. 4 and 5 that the extreme studentized residuals increase faster than the extreme normalized residuals with the number of observations n.

Another difference between Figs. 4 and 5 is that in small networks, the critical values of the studentized residuals hardly depend on α. The reason is that here a large part of

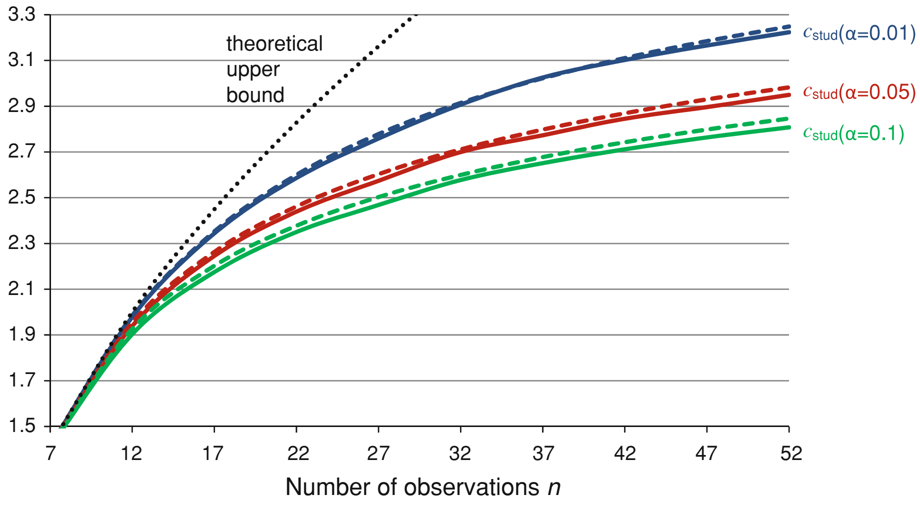

Fig. 5 Critical values of extreme studentized residuals for the free leveling network. *Solid curves* obtained from (21) by $m = 20{,}000$ Monte Carlo experiments. *Dashed curves* obtained from (15). *Dotted curve* theoretical upper bound according to (11)

studentized residuals concentrate near the theoretical upper bound of c according to (11); see dotted curve in Fig. 5. Overestimation as before is hardly possible. Differences between the improved critical values (21) and those values computed by (15) come into effect when the network gets larger. Again, the values obtained from (14) are larger for the same reason as in the case of normalized residuals. The difference is less pronounced here because of weaker correlations in larger networks, cf. discussion above.

Larger differences can be expected for models where the correlations do not diminish with size, e.g. if in a geodetic network also long traverses have been observed.

9 Correlated observations

All derivations are valid also for correlated observations where P is not a diagonal matrix. The generation of correlated pseudo random numbers e with a desired covariance matrix $D\{e\} = \sigma_0^2 P^{-1}$ in (1) from uncorrelated pseudo random numbers e' with covariance matrix I can be done by the transform

$$e = Ue'$$

with any $n \times n$-matrix U fulfilling

$$UU^{\mathrm{T}} = D\{e\} \tag{24}$$

This property is a consequence of covariance propagation.

We repeat the computations of the preceding section with a covariance matrix

$$D\{e\} = \sigma_0^2 \begin{pmatrix} 1 & \rho & \rho & \dots & \rho & \rho \\ \rho & 1 & \rho & \dots & \rho & \rho \\ \rho & \rho & 1 & \dots & \rho & \rho \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \rho & \ddots & 1 & \rho \\ \rho & \rho & \rho & \dots & \rho & 1 \end{pmatrix} \tag{25}$$

where all pairs of observations are correlated with a correlation coefficient of $0 \le \rho < 1$. This correlation can be assumed to be caused by using the same leveling equipment throughout the whole network. A simple matrix U having property (24) is a $n \times n$-matrix with elements equal to

$$u_{ij} = \frac{1}{n}\left(\sqrt{1 + (n-1)\rho} + \sqrt{1-\rho}\right) + \delta_{ij}\sqrt{1-\rho}$$

where δ_{ij} denotes Kronecker's delta.

Example For $\rho = 0.9$, the cofactor matrix of the residuals (23) is modified to

$$Q_{\hat{e}\hat{e}} = \begin{pmatrix} +0.29 & +0.29 & +0.29 & -0.04 & -0.04 & -0.04 \\ +0.29 & +0.29 & +0.29 & -0.04 & -0.04 & -0.04 \\ +0.29 & +0.29 & +0.29 & -0.04 & -0.04 & -0.04 \\ +0.25 & +0.25 & +0.25 & +0.25 & +0.25 & +0.25 \\ -0.04 & -0.04 & -0.04 & +0.29 & +0.29 & +0.29 \\ -0.04 & -0.04 & -0.04 & +0.29 & +0.29 & +0.29 \\ -0.04 & -0.04 & -0.04 & +0.29 & +0.29 & +0.29 \end{pmatrix} \tag{26}$$

Although $\rho = 0.9$ is a quite large correlation coefficient, (23) and (26) are not very much different. We conclude that correlations of the residuals are much more caused by the functional than by the stochastic relationships of the observations. So, it is no surprise that the critical values do not strongly depend on the stochastic correlations introduced by (25). In Tables 1 and 2, we only display the results for $\alpha = 0.05$. Other values of α yield a similar outcome. We can conclude from these tables that in this case, the critical values are practically the same for all correlations.

10 Non-normal error distributions

Formulas (14) and (15) are only valid if observation errors are normally distributed according to (6).

Table 1 Critical values of extreme normalized residuals for the free leveling network and correlated observations (25), $\alpha = 0.05$

n	c from (14)	c from (21) $\rho = 0$	c from (21) $\rho = 0.3$	c from (21) $\rho = 0.6$	c from (21) $\rho = 0.9$
7	2.69	2.34	2.31	2.34	2.36
12	2.87	2.68	2.68	2.67	2.68
17	2.97	2.83	2.82	2.84	2.84
22	3.05	2.94	2.91	2.94	2.93
27	3.11	3.02	3.02	3.03	3.01
32	3.16	3.07	3.06	3.08	3.08
37	3.20	3.12	3.11	3.12	3.12
42	3.24	3.17	3.16	3.15	3.17
47	3.27	3.20	3.21	3.21	3.21
52	3.30	3.22	3.23	3.23	3.24

Table 2 Same as Table 1, but studentized rather than normalized residuals

n	c from (15)	c from (21) $\rho = 0$	c from (21) $\rho = 0.3$	c from (21) $\rho = 0.6$	c from (21) $\rho = 0.9$
7	1.41	1.41	1.41	1.41	1.41
12	1.95	1.94	1.94	1.94	1.94
17	2.26	2.24	2.25	2.24	2.24
22	2.46	2.44	2.44	2.45	2.43
27	2.60	2.59	2.58	2.59	2.58
32	2.71	2.68	2.69	2.68	2.69
37	2.80	2.78	2.78	2.78	2.77
42	2.87	2.85	2.85	2.84	2.85
47	2.93	2.91	2.90	2.91	2.91
52	2.98	2.96	2.96	2.95	2.96

A further advantage of the Monte Carlo method is that we are not restricted to assumption (6). The only requirement is that we must be able to generate pseudo random numbers of the non-normal error distribution. This is ensured for the most common alternatives like symmetric triangular distribution or Laplace distribution. In both cases, we get the inverse cdf explicitly. This enables us to apply the inverse transformation method for the generation of pseudo random numbers according to the symmetric triangular and Laplace distribution (Tanizaki 2004, p. 116). Although neglected here, a further interesting alternative is the scale-contaminated normal distribution (cf. Lehmann 2012).

We repeat the computations of Sect. 8 using (see Fig. 6)

1. the central symmetric triangular distribution and
2. the central Laplace distribution.

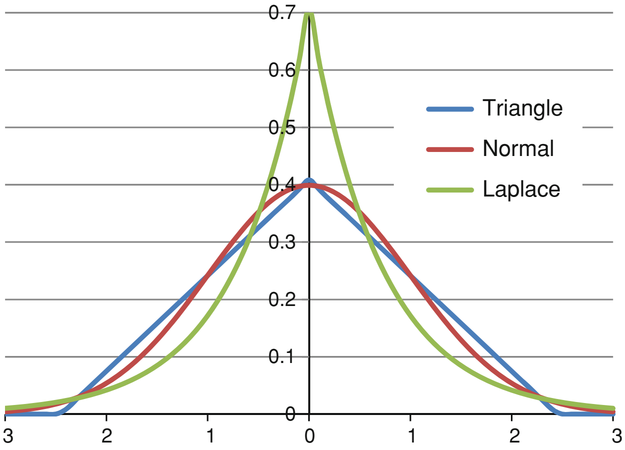

Fig. 6 Probability density functions of distributions used (zero expectation, unit variance)

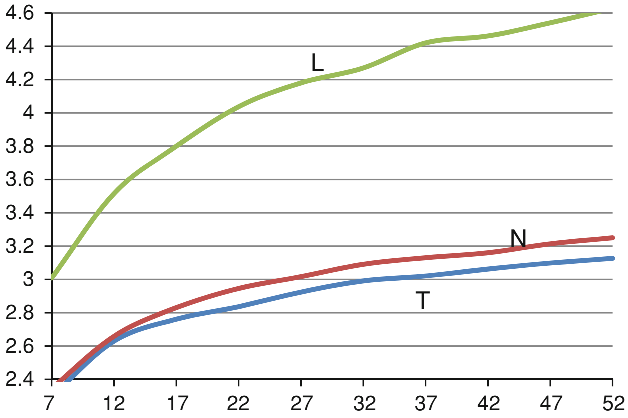

Fig. 7 Critical values of extreme normalized residuals for the free leveling network computed by (21) using the triangular (T) distribution (*blue*), normal (N) distribution (*red*) and Laplace (L) distribution (*green curve*) for $\alpha = 0.05$

Note that again the variance of the pseudo random numbers can be chosen arbitrarily because it cancels in (7) or (9). In Fig. 7 we only display the results for $\alpha = 0.05$. Other values of α yield a similar outcome.

Figures 7 and 8 show that the critical values depend very much on the type of error distribution, even more than on the functional model. This effect becomes more pronounced as the number of observations increases. The triangular distribution has bounded observation errors. Under the hypothesis H_0, this makes large extreme normalized or studentized residuals very unlikely. On the contrary, the Laplace distribution has heavy "tails". Here we can expect to get large extreme normalized or studentized residuals even if there are no outliers (H_0 holds true). As expected, we observe that $c_\mathrm{T} < c_\mathrm{N} < c_\mathrm{L}$ holds with subscripts as in Figs. 7 and 8.

11 Considerations regarding the computational workload

In each Monte Carlo experiment, we have to generate n pseudo random numbers e_k and evaluate $\hat{e} = Re$. Note that

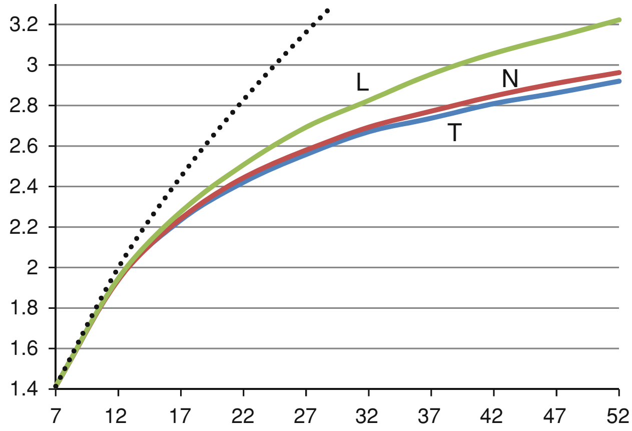

Fig. 8 Same as Fig. 7, but studentized rather than normalized residuals. *Dotted curve* theoretical upper bound according to (11)

R must be computed *only once*. In the worst case that R is a dense (not sparse) matrix, we need $2n^2$ floating point operations for $\hat{e} = Re$. The computation of the test statistic has only $O(n)$ operations. Sorting the vector of test statistics typically requires $O(n \log n)$ operations. If n is large, only the evaluation of $\hat{e} = Re$ with dense R deserves consideration.

If we assume a Gauss Markov model with $n = 1{,}000$ unknowns and $m = 20{,}000$ Monte Carlo experiments, this amounts to 4×10^{10} floating point operations for only the evaluation of $\hat{e} = Re$. For example on a common Intel Core i7 processor with up to 100 GFLOPS, this computation takes less than a second. In general, we can expect that minor changes of the functional and stochastic model may not change the improved critical values considerably.

Hence, we are not overwhelmed with computational workload. Also, remember that (21) can be evaluated efficiently for a sequence of values α in parallel.

12 Monte Carlo-based data snooping

We make a statement concerning the use of the approach proposed in Lehmann (2010) and Lehmann and Scheffler (2011) called “Monte Carlo-based data snooping” and show how it relates to the subject of this paper. It finds the optimum level of error probability α as follows: for a number of trial levels $\alpha_i, i = 1, \ldots, M$ the posterior variance of the estimated parameters is computed and the optimum α^*, i.e. the value α_i, for which the posterior variance of the estimated parameters is minimum, is selected and possibly refined by some interpolation. M is a chosen integer large enough to ensure finding the optimum α^*. The use of a geometric progression like $\alpha_i = 2^{-2-i}, i = 1, \ldots, 7$ is suggested. One can even use the trial critical values $c_i, i = 1, \ldots, M$ instead of error levels, finding the optimum c^* in the same way. In this case, c_i is not computed in any way from α_i by (14) or (15). Therefore, Monte Carlo-based data snooping does not require the improved critical values derived here. Computation time cannot be saved, however, because Monte Carlo-based data snooping is also moderately time-consuming.

13 Conclusions

The improved critical values promised by the title of this paper cannot exactly be listed in a table as in the classical case (see e.g. Pope 1976). This would only be possible in special cases (see Lemeshko and Lemeshko 2005). They depend more or less on the functional and stochastic model. All that can be given is a procedure for computing the improved values for each model.

The improved critical values (21) are always smaller than the critical values (14), (15) used so far. This means that in the case of no outliers, it is less likely that extreme normalized or studentized residuals get very much larger than those predicted by the classical approach of Sect. 4. For example, if H_0 is true and is rejected against $H_A^{(1)}$, it is more likely that it is rejected also against $H_A^{(2)}$, etc. because the test statistics are correlated. This makes the approximation (13) relatively coarse.

Let us consider the case of iterative outlier elimination, i.e. after each detected and eliminated outlier, the model is reprocessed and the outlier detection is restarted. This is a standard procedure in geodesy. It is obvious that with the improved critical values, we get *the same sequence* of outliers as in the classical approach of Sect. 4. It only truncates at a later cycle of iteration. In other words, we get the same outliers as with the classical approach and a larger level of error probability α. However, it is not easy to derive this larger α somehow from the value α used before.

Finally, we suggest applying Monte Carlo methods also in other fields of outlier detection, e.g. there are test statistics for outlier detection, which do not even approximately permit an analytical derivation of a related cdf. Therefore, such test statistics have not yet been used in geodesy. As an example, we mention the kurtosis of the normalized residuals (cf. Verma et al. 2008), which is able to detect multiple outliers without the need to a priori specify them. Using the Monte Carlo method, critical values for such test statistics can be computed in the same way as for the extreme residuals.

References

Aliosmanoglu S, Akyilmaz O (2001) A comparison between statistical and fuzzy techniques in outlier detection. International association of geodesy symposia symposia, vol 125. Springer, Berlin, pp 382–387

Alkhatib H, Neumann I, Kutterer H (2009) Uncertainty modeling of random and systematic errors by means of Monte Carlo and fuzzy techniques. J Appl Geod 3(2):67–79

Andersen R (2008) Modern methods for robust regression. Sage Publications, London

Baarda W (1968) A testing procedure for use in geodetic networks. Netherlands Geodetic Commission, Publication on Geodesy, vol 2, No. 5, Delft, The Netherlands

Box GEP, Muller ME (1958) A note on the generation of random normal deviates. Ann. Math. Stat. 29(2):610–611

Hawkins D (1980) Identification of outliers. Chapman and Hall, London

Koch KR (1999) Parameter estimation and hypothesis testing in linear models. Springer, Berlin

Koch KR (2007) Outlier detection in observations including leverage points by Monte Carlo simulations. Allgemeine Vermessungsnachrichten. VDE Verlag Berlin Offenbach

Lemeshko BY, Lemeshko SB (2005) Extending the application of Grubbs-type tests in rejecting anomalous measurements. Meas Tech 48(6):536–547

Lehmann R (1994) Adjustment in non-linear models by means of the adaptive Monte-Carlo-Integration. Allgemeine Vermessungsnachrichten, vol 7/1994. Herbert Wichmann Verlag GmbH Heidelberg (in German)

Lehmann R (2010) Normalized residuals—how large is too large? Allgemeine Vermessungsnachrichten, vol 2/2010. VDE Verlag Berlin Offenbach (in German)

Lehmann R (2012) Geodetic error calculus by the scale contaminated normal distribution. Allgemeine Vermessungsnachrichten. VDE Verlag Berlin Offenbach (in German) (in press)

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

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

Stefansky W (1972) Rejecting outliers in factorial designs. Technometrics 14:469–479

Tanizaki H (2004) Computational methods in statistics and econometrics. Marcel Dekker, New York

Verma SP, Quiroz-Ruiz A, Díaz-González L (2008) Critical values for 33 discordancy test variants for outliers in normal samples up to sizes 1000, and applications in quality control in Earth Sciences. Revista Mexicana de Ciencias Geológicas 25(1):82–96

Wieser A (2001) Robust and fuzzy techniques for parameter estimation and quality assessment in GPS. Ph.D. dissertation, Graz University of Technology, Austria