Observation error model selection by information criteria vs. normality testing

Author

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

Abstract

To extract the best possible information from geodetic and geophysical observations, it is necessary to select a model of the observation errors, mostly the family of Gaussian normal distributions. However, there are alternatives, typically chosen in the framework of robust M-estimation. We give a synopsis of well-known and less well-known models for observation errors and propose to select a model based on information criteria. In this contribution we compare the Akaike information criterion (AIC) and the Anderson Darling (AD) test and apply them to the test problem of fitting a straight line. The comparison is facilitated by a Monte Carlo approach. It turns out that the model selection by AIC has some advantages over the AD test.

Keywords

maximum likelihood estimation; robust estimation; Gaussian normal distribution; Laplace distribution; generalized normal distribution; contaminated normal distribution; Akaike information criterion; Anderson Darling test; Monte Carlo method

1 Introduction

In geodesy, geophysics and many other scientific branches we are confronted with observations affected by observation errors. Since the operation of these errors is generally very complex and not well understood, their effect is mostly treated as random. Consequently, for more than 200 years geodesists and geophysicists take advantage of stochastics and partly also contribute to this field of mathematics. See (*Kutterer 2001*) for general remarks on the role of statistics in geodetic data analysis, also with a view to related concepts of uncertainty assessment.

To extract the best possible information from these observations by parameter estimation, e.g. by the concept of maximum likelihood (ML) estimation , it is necessary to make an assumption on the stochastical properties of the observation errors. These properties are completely derived from a

probability distribution of these errors. However, in practical applications such a probability distribution is never exactly known. Fortunately, there are some methods of parameter estimation, which do not need the full distribution, but only some moments like expectation and variance. But when we arrive at the basic problem of testing statistical hypotheses, we can hardly do without the assumption of a full stochastic observation error model.

The normal distribution, mostly credited to C.F. Gauss, is the best known model of geodetic and geophysical observation errors. (As usual, when we speak about 'distributions', we often mean a 'family' of distributions, which is clear from the context.) Due to its well known nice mathematical properties, first and foremost the property of being a stable distribution, it greatly simplifies the parameter estimation problem. Its choice is further motivated by both the central limit theorem as well as the maximum entropy principle. The application of the normal error distribution in practical geodesy and geophysics is also not without success. The common hypothesis tests like t-test, τ-test, χ^2-test and F-test are all based on this distribution, and critical values of these tests are found in widespread statistical lookup tables or are computed by popular scientific software (e.g. *Teunissen 2000*).

Already in the 19th century it was realized that typical error distributions of real observations are more peakshaped and thickertailed than the Gaussian bell (see *Hampel 2001* for a historical synopsis of such investigations). This gave rise to the development of robust estimation methods like L1 norm minimization or more generally in the framework of M-estimation (e.g. *Huber 2009*). However, only until recently, there was not enough computer power to actually compute robust estimates for real-life data sets. Peakshapedness of a probability distribution is measured by the standardized fourth moment of the distribution, known as kurtosis. Distributions with kurtosis >3 are called leptokurtic. Kurtosis minus 3, which is the kurtosis of the normal distribution, is also called excess kurtosis. Thus, typical error distributions of real observations seem to exhibit a positive excess kurtosis, i.e., they are leptokurtic. *Wisniewski (2014)* considers M-estimations with probabilistic models of geodetic observations including the asymmetry and the excess kurtosis, which are basic anomalies of empiric distributions of errors of geodetic, geophysical or astrometric observations.

This poses the problem of deciding, whether the normal distribution is an applicable observation error model nonetheless or if it must be replaced by something better adapted to the observations. This problem may be formalized as a stochastical hypothesis. Therefore, besides graphical methods like the famous Q-Q-plot, hypothesis testing is the most popular approach. Many hypothesis tests for normality have been proposed:

- D'Agostino's K^2 test (*D'Agostino 1970*)
- Jarque–Bera test (*Jarque and Bera 1980*)
- Anderson–Darling test (*Anderson and Darling 1952, 1954*)
- Cramér–von Mises criterion (*Cramér 1928; von Mises 1931*)
- Lilliefors test (*Lilliefors 1967*)
- Kolmogorov–Smirnov test (*Kolmogorov 1933; Smirnov 1948*)
- Shapiro–Wilk test (*Shapiro and Wilk 1965*)
- Pearson's chi-squared test (*Pearson 1900*)
- Shapiro–Francia test (*Shapiro and Francia 1972*)

However, all of them only work with samples of one random variable. Some of them require a known mean and variance. The tests differ with respect to computational simplicity and statistical power. Some of them are powerful only in case of certain types of deviation from normality (kurtosis, skewness, etc.), i.e. with respect to a certain alternative hypothesis. *Razali and Wah (2011)* found in a Monte Carlo simulation "that Shapiro-Wilk test is the most powerful normality test, followed by Anderson-Darling test, Lilliefors test and Kolmogorov-Smirnov test."

There is an ongoing interest in the adaption of distribution models to observations, e.g. in the field of GNSS observations. *Tiberius and Borre (2000)* analyzed the distribution of GPS code and phase observations evaluating sample moments and applying different statistical hypothesis tests. The authors conclude that the normal distribution assumption seems to be reasonable for the data from short baselines. However, deviations from normality arose for long baselines, and were attributed to multipath effects and unmodeled differential atmospheric delays. *Verhagen and Teunissen (2005)* present and evaluate the joint probability density function of the multivariate integer GPS carrier phase ambiguity residuals. *Cia et al. (2007)* propose the von Mises normal distribution for GNSS carrier phase observations. *Luo et al. (2011)* and *Luo (2013)* investigate the distribution of the same type of observations by sample moments, various statistical hypothesis tests, and graphical tools. The results based on a large and representative data set of GPS phase measurements showed various deviations from normality.

In the more typical situation arising in geodesy and geophysics, when the observations are part of a Gauss Markov model (GMM) or similar linear model, no rigorous test for normality is known. Practically it is often tried to apply the test for normality to the residuals of the models because they inherit their normality from the observation errors (e.g. *Luo et al. 2011)*. But this does not say much about the normality of the observation errors themselves, as will be further explained in section 3.

Deciding, which model for observation errors should be assigned to a set of observations can be viewed as a problem of model selection. From information theory we know of different approaches of model selection based on information criteria. The oldest and best known is the Akaike Information Criterion (*Akaike 1974*):

$$AIC = 2k - 2\log L(\hat{\theta};\ l) \tag{1}$$

where L denotes the likelihood function of the model, which is maximized by the maximum likelihood (ML) estimate $\hat{\theta}$ of the k-vector of parameters θ with respect to the observations l. Note that θ should comprise all parameters, i.e. also unknown variance factors or variance components. The criterion is: Among all models under consideration the one with the least AIC is to be selected. It has high likelihood and at the same time not too many parameters k, which prevents over-parametrization. If different models give AIC values very close to the minimum, it is generally recommended to avoid the selection, if possible (*Burnham and Anderson 2002*). Some geodetic applications of information criteria are presented recently for the selection of transformation models by *Lehmann (2014)* and in the framework of geodetic multiple outlier detection by *Lehmann and Lösler (2015)*. Another scope of application is the auto regressive moving-average process (e.g. *Klees et al. 2002*) especially in the framework of GNSS time series analysis (cf. *Luo et al. 2011*). In section 4 we will develop a strategy to apply information criteria for observation error model selection.

The paper is organized as follows: After introducing well and less well known models of observation errors we briefly review the Anderson-Darling (AD) test in its special form as a test for normality. Opposed to this we propose the strategy of observation error model selection by AIC. Finally, the Monte Carlo method is used to investigate and compare both strategies applied to the model of a straight line fit.

2 Models for observation errors

We start with the well-known Gaussian normal distribution $N(\mu, \sigma^2)$ with expectation μ and standard deviation σ. Its probability density function (PDF) reads

$$f_N(y|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{2}$$

A common measure of peakshapedness and tail-thickness is the excess kurtosis

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \tag{3}$$

where $\mu_4 = E\{(y - E\{y\})^4\}$ denotes the fourth central moment of the distribution. The excess kurtosis uses the normal distribution as a benchmark for peakshapedness, such that it becomes $\gamma_2 = 0$ for this distribution.

More typical error distributions of real observations seem to be leptokurtic, i.e. $\gamma_2 > 0$. The most simple leptokurtic error distribution is the Laplace distribution $L(\mu, \sigma^2)$ with expectation μ and standard deviation σ. Its PDF reads

$$f_L(y|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2}}\exp\left(-\frac{|y-\mu|}{\sigma}\sqrt{2}\right) \tag{4}$$

It has excess kurtosis $\gamma_2 = 3$, which is often overshooting the mark. It would be better to have a distribution model with a shape parameter, that can be tuned to the kurtosis of the real error distribution. Such a model is the generalized normal distribution $G(\mu, \alpha, \beta)$ with expectation μ, a scale parameter $\alpha > 0$ and a shape parameter $\beta > 0$.

$$f_G(y|\mu,\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)}\exp\left(-\frac{|y-\mu|^\beta}{\alpha^\beta}\right) \tag{5}$$

Γ denotes the Gamma function. This distribution includes normal and Laplace distribution as special cases with $\beta = 2$ and $\beta = 1$, respectively. Variance and kurtosis read

$$\sigma^2 = \alpha^2\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}, \qquad \gamma_2 = \frac{\Gamma(5/\beta)\Gamma(1/\beta)}{\Gamma(3/\beta)^2} - 3 \tag{6}$$

A different medium between normal and Laplace distribution can be derived from a common loss function in M-estimation introduced by *Huber (1964)*. It is a composite distribution $\mathrm{H}(\mu, c, d)$, consisting of a Gaussian peak and two Laplacian tails. It has three parameters: the expectation μ, a scale parameter $d > 0$ and a shape parameter $k > 0$. The PDF reads

$$f_H(y|\mu,d,k) = C(d,k)\begin{cases} \exp\left(-\frac{(y-\mu)^2}{2d^2}\right) & \text{for } |y-\mu| < k \\ \exp\left(\frac{k^2 - 2k|y-\mu|}{2d^2}\right) & \text{for } |y-\mu| \geq k \end{cases} \tag{7}$$

where $C(d,k)$ is a normalization function. The composition is such that f_H is continuous at the connection points $\mu \pm k$, where also the first derivatives are continuous. Nonetheless, this composition character makes numerical computations rather costly. Variance and excess kurtosis cannot be computed without a costly numerical quadrature.

An alternative leptokurtic error model is Student's-t distribution. Here we introduce it in its three parameter version $t(\mu,\gamma,\upsilon)$ with expectation μ, a scale parameter $\gamma > 0$ and a shape parameter $\nu > 0$. The PDF reads

$$f_t(y|\mu,\gamma,\upsilon) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\gamma\sqrt{\nu\pi}}\left(1+\frac{(y-\mu)^2}{\gamma^2\upsilon}\right)^{-\frac{\upsilon+1}{2}} \tag{8}$$

Variance and excess kurtosis may be computed by

$$\sigma^2 = \gamma^2\frac{\upsilon}{\upsilon-2} \text{ if } \upsilon > 2\,,\ \gamma_2 = \frac{6}{\upsilon-4} \text{ if } \upsilon > 4 \tag{9}$$

The Student's-t distribution can be used as a model of an extremely leptokurtic distribution. For $\upsilon \leq 4$ the excess kurtosis is even no longer finite.

The scale contaminated normal distribution $SC(\mu,\sigma_1^2,\sigma_2^2,\varepsilon)$ is a further generalization of the normal distribution, first discussed by *Tukey (1960)*. Geodetic applications for robust estimation and outlier detection are discussed by *Lehmann (2012, 2013)*. This distribution describes a normal population contaminated by a small number of members of a different normal population with much larger variance (gross errors).

It has expectation μ, the variances σ_1^2 of the original distribution and σ_2^2 of the contaminating distribution and a weight parameter $0 \leq \varepsilon \leq 1$, specifying the degree of contamination. The PDF reads

$$f_{SC}(y|\mu,\sigma_1^2,\sigma_2^2,\varepsilon) = \frac{1}{\sqrt{2\pi}}\left(\frac{1-\varepsilon}{\sigma_1}\exp\left(-\frac{(y-\mu)^2}{2\sigma_1^2}\right)+\frac{\varepsilon}{\sigma_2}\exp\left(-\frac{(y-\mu)^2}{2\sigma_2^2}\right)\right) \tag{10}$$

Table 1 gives a synopsis of the most important models for observation errors.

Table 1 Models for observation errors (*expressions for terms in brackets are intricate here)

Distribution	Generalized normal (5)	Huber (7)	Student's-t (8)	Scale contaminated normal (10)
relevant special cases	normal: $\beta = 2$ Laplace: $\beta = 1$	normal: $k \to \infty$ Laplace: $k \to 0$	normal: $\nu \to \infty$ Cauchy: $\nu = 1$	normal: $\sigma_1 = \sigma_2$ or $\varepsilon = 0$ (or $\varepsilon = 1$)
parameters for $\gamma_2 = 1$	$\beta = 1.406$	e.g. $d = 0.911$, $k = 1.511$, gives $\sigma = 1$	$\nu = 10$	e.g. $\varepsilon = 0.1, \sigma_1 = 0.899, \sigma_2 = 1.653$ or $\varepsilon = 0.03, \sigma_1 = 0.948, \sigma_2 = 2.070$ or $\varepsilon = 0.01, \sigma_1 = 0.971, \sigma_2 = 2.597$ or $\varepsilon = 0.003, \sigma_1 = 0.984, \sigma_2 = 3.395$ or $\varepsilon = 0.001, \sigma_1 = 0.991, \sigma_2 = 4.388$ yield $\sigma = 1$
parameters for $\gamma_2 = 6$	$\beta = 0.7785$	not possible	$\nu = 5$	e.g. $\varepsilon = 0.03, \sigma_1 = 0.867, \sigma_2 = 3.001$ or $\varepsilon = 0.003, \sigma_1 = 0.960, \sigma_2 = 5.175$ yield $\sigma = 1$
closed expressions*	f, F, F^{-1}, σ, γ_2, H	$f, (F, F^{-1})$	$f, F, (F^{-1}), \sigma, \gamma_2$	f, F, σ, γ_2
importance	target density in M-estimation	target density in M-estimation	generalization of statistical test distribution	instructive gross error modeling according to the variance inflation model (cf. *Lehmann 2012, 2013*)

3 Anderson-Darling normality test (AD test)

Anderson and Darling (1952,1954) developed a statistical hypothesis test for testing the distribution of a stochastical sample. The test statistic basically measures the difference between the empirical distribution of the sample and the hypothesized distribution, giving more weights to the tails of the distribution than similar tests, e.g. Cramér–von Mises criterion.

In this investigation we focus on the Anderson-Darling (AD) test because it is recommended by *Razali and Wah (2011)* as a very powerful test, but is at the same time relatively easy to implement. The test procedure is as follows:

Let $Y_1 < Y_2 < \cdots < Y_n$ be the ordered sequence of sample values, then the test statistic is defined as

$$A_2 = -n - \frac{1}{n}\sum_{i=1}^{n}\Big((2i-1)\log\big(F(Y_i)\big) + (2(n-i)+1)\log\big(1-F(Y_i)\big)\Big) \tag{11}$$

where F is the hypothesized cumulative distribution function (CDF). If the distribution of Y differs significantly from the hypothesized distribution then A_2 tends to assume large values.

The AD test is oftentimes used as a test for normality, e.g. as a pretest to check the presumption of normality before a test requiring the sample to be normally distributed, like the t-test or the F-test, is applied. In this case F is the CDF of the normal distribution. Critical values for the normality test of samples are given by *Stephens (1974)*.

When observation errors in a GMM or similar linear model should be tested for normality, the AD test cannot be applied directly. One could try to use the residuals and test them for normality, but they are often found to be normally distributed, although the observation errors themselves are not. This can be understood as follows: Assume that the observations of a linear model are not normally distributed. The residuals are linear functions of the observations, and according to the central limit theorem, they tend towards normality, as the number of observations increases. A normality test applied to the residuals may not be rejected, although the observation errors are far from being normally distributed. This results in a type II decision error.

Fortunately, a hypothesis test like χ^2 or F-test, where the test statistic is a function of the residuals, is often relatively unsusceptible to non-normal error distributions. This is why such tests "work" even though observation errors are not normal.

As a correction to this error, one must compute new critical values for each linear model. Fortunately, today enough computer power is available to accomplish this, and it will be done in section 5.

If the hypothesis of normality is rejected, it is not clear, which model of alternative distribution models should be employed. This could perhaps be done in a multiple test, where e.g. a test for generalized normality is invoked next. But as in any multiple test, there are pitfalls (e.g. *Miller 1981*). In this contribution we do neither recommend nor pursue such an approach.

4 Observation error model selection by Akaike's information criterion (AIC)

As pointed out in the introduction, the selection of an observation error model can be viewed as a general model selection problem, for which information theory provides so-called information criteria. We already introduced the Akaike Information AIC by (1). A corrected version of AIC is

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{12}$$

which is supposed to work better for small sample sizes. If n is small or k is large then AICc is strongly recommended rather than AIC (*Burnham and Anderson 2004*). It is important that parameters in the sense of (1) and (12) are also unknown variances and variance components. k also counts these quantities.

There are many alternatives to AIC, which seem to work better in special situations. We only mention the Bayesian Information Criterion (BIC), which uses a further modification of (1).

If we decide to select an observation error model by information criteria, we could proceed as follows:

1. Compute the model parameters by a ML estimation from all candidate observation error model, e.g. normal distribution, generalized and contaminated normal distributions, Laplace distribution, Huber's distribution, Student's-t distribution etc.
2. For all of the results, compute the information criterion, e.g. AIC by (1) or AICc by (12).

3. Select the model, where the information criterion assumes a minimum (possibly only if it is significantly below the second smallest value).
4. Proceed with the parameters estimated from the selected model (if any).

Step 1 is the most time consuming and difficult step. To begin with, computing ML estimates of normal, Laplace and Huber's distribution is still relatively easy and well understood. In estimation they are known L2 and L1 norm minimizations as well as M-estimation by Huber's influence function.

Computing ML estimates of generalized normal distribution is harder. In our contribution we use a kind of brute force method, which must be refined before problems of practical dimensions can be tackled:

1. Use the normal distribution as initial guess, i.e. take the solution of the L2 norm minimization problem computed before and let $\beta = 2$.
2. Perform a line search optimization for the shape parameter β in (5) using proper bounds. Here we use $0 < \beta \leq 2$, because a leptokurtic distribution is desired. The remaining parameters are held fixed.
3. Fix β now and optimize the remaining parameters, i.e. solve the Lβ norm minimization problem.
4. Return to step 2 until convergence.

Computing ML estimates of scale contaminated normal distribution is the hardest piece of work.

1. Again, use the normal distribution as initial guess, i.e. let $\varepsilon = 0$.
2. Initially guess some variance of contamination, e.g. $\sigma_2 = 10\sigma_1$ is used here.
5. Perform a line search optimization for the shape parameter ε in (10) using bounds $0 \leq \varepsilon \leq 1$. The remaining parameters are held fixed.
3. Fix ε now and optimize the remaining parameters by solving a general non-linear minimization problem.
4. Return to step 3 until convergence.

Computing ML estimates of Student's-t distribution is not discussed here.

5 Simulated observations and candidate observation error models

To compute the success rate of observation error model selection, a Monte Carlo method must be applied. For this purpose we generate $M = 10000$ observation vectors of a selected error model by a pseudo random number (PRN) generator. It has been investigated that the results presented here do not change significantly when the computations are repeated with different PRN, such that $M = 10000$ is sufficiently large to support the conclusions made below. It has been taken care that the PRN generator is reseeded each time.

Four different observation error distributions are generated here:

- standard normal distribution $N(0,1)$
- standard Laplace distribution $L(0,1)$
- weakly scale contaminated normal distribution $SC(0,1,3^2,0.1)$
- strongly scale contaminated normal distribution $SC(0,1,10^2,0.1)$

For the standard normal PRN we can use directly MATLAB 8.1's PRN generator "normrnd". In the standard Laplace case we use uniformly distributed PRN generated by MATLAB 8.1's PRN generator "unidrnd" and apply a transformation by the inverse CDF (cf. *Tanizaki 2004 p. 122 ff.*). In the scale contaminated cases we generate a normal PRN with $\sigma_1 = 1$ and contaminate it with probability $\varepsilon = 0.1$ by a second normal PRN with either $\sigma_2 = 3$ or $\sigma_2 = 10$. This simulates a normal error model with a gross error rate of 10% of an either 3 times or 10 times larger standard deviation.

As a functional model we choose the straight line fit with $n = 30$ and $n = 100$ data points at fixed equidistant abscissa:

$$y_i = x_0 + ix_1 + \varepsilon_i, \qquad i = 1, \dots, n \tag{13}$$

This model is of general relevance in various fields of geodesy, geophysics and related sciences as well as engineering disciplines. Examples are

- extracting a linear trend from a geodetic or geophysical time series
- fitting a linear calibration function for calibration of measuring devices
- surveying points on a spatial straight line, which deviate from a straight line due to observation errors

As candidate observation error models we choose

- normal distribution $N(0, \sigma^2)$ with unknown scale parameter σ
- Laplace distribution $L(0, \sigma^2)$ with unknown scale parameter σ
- generalized normal distribution $G(0, \alpha, \beta)$ with unknown scale and shape parameters α, β
- scale contaminated normal distribution $SC(0, \sigma_1^2, \sigma_2^2, \varepsilon)$ with two unknown scale parameters σ_1^2, σ_2^2 and an unknown contamination parameter ε

The first two models have in total three parameters $\theta = (x_0, x_1, \sigma^2)^T$, the third has four parameters $\theta = (x_0, x_1, \alpha, \beta)^T$ and the last has five parameters $\theta = (x_0, x_1, \sigma_1^2, \sigma_2^2, \varepsilon)^T$. The computations below employ both a Anderson-Darling normality test for the residuals of the straight line fit as well as a the model selection by AIC (1). Note that none of the presented results depend in any way on the actual true parameters x_0, x_1.

6 Results

First, we need to compute critical values for the Anderson-Darling normality test applied to the residuals of the straight line fit. For this purpose a least squares fit is computed to each normal observation error PRN vector and the corresponding value A_2 statistic in (11) is derived. From the resulting frequency distribution of A_2 displayed in Fig. 1 we extract the quantiles as a good approximation to the critical values for various type I error rates α. They are given in table 2.

The critical values are significantly smaller than those given by *Stephens (1974)* for samples. For example, the critical value for a type I error rate of α=0.01 for samples is 1.09, while we found 1.03. This confirms the assertion of section 3 that the normality test is more often positive for the residuals than for the corresponding observation errors. This effect is now taken into account by the new critical values.

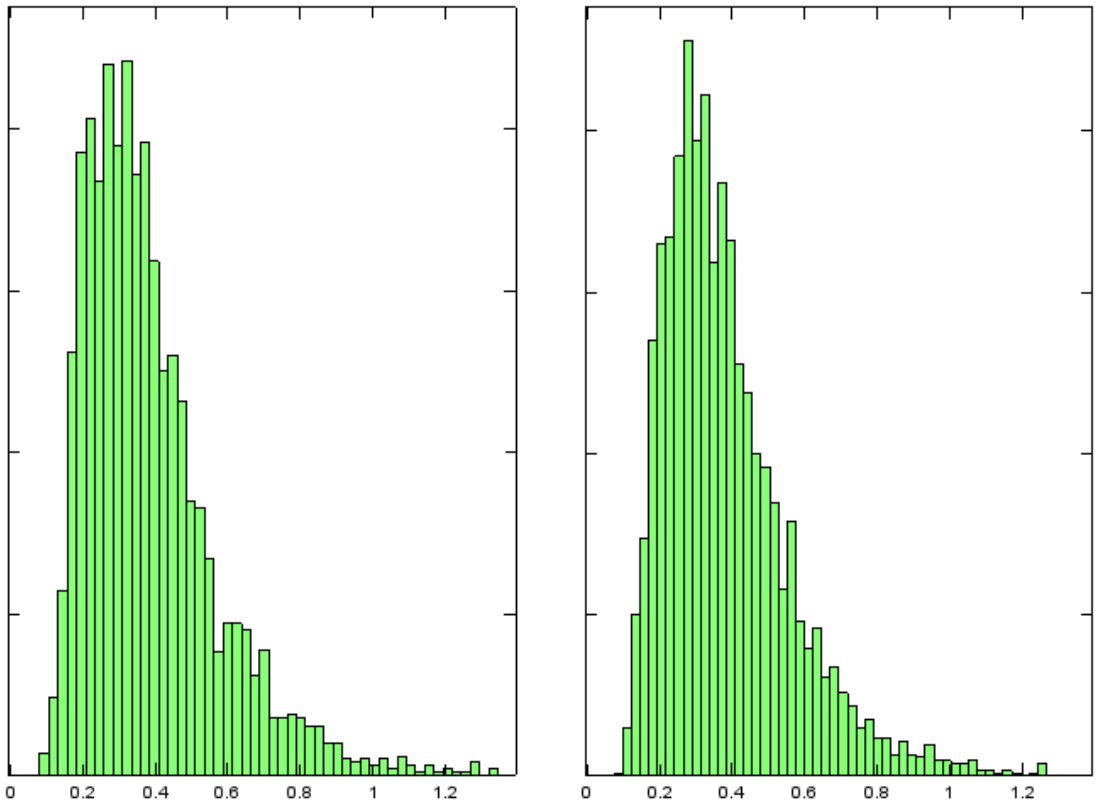

Fig. 1. Histograms of the Anderson-Darling normality test statistic A_2 in (11) applied to the residuals of the straight line fit with normal observation errors , left: $n = 30$, right: $n = 100$ observations.

Table 2. Critical values and statistical powers of the Anderson-Darling normality test for different number of observations n. L: Laplace distribution, SC: scale contaminated normal distribution.

	$n = 30$			$n = 100$		
Type I error rate α	0.10	0.05	0.01	0.10	0.05	0.01
Critical values	0.63	0.74	1.03	0.62	0.73	1.03
Statistical power for $L(0,1)$	0.40	0.31	0.14	0.87	0.81	0.63
Statistical power for $SC(0,1,3^2,0.1)$	0.36	0.29	0.16	0.74	0.67	0.51
Statistical power for $SC(0,1,10^2,0.1)$	0.86	0.85	0.81	0.99	0.99	0.99

Moreover, in contrast to the Anderson-Darling normality test, these values slightly depend on n. They are smaller for larger n because as the model size increases, the residuals tend towards normality.

These critical values are now used to test the hypothesis of normality of the non-normal observation error PRN vectors. The results are displayed in Table 2 in terms of statistical power, which is the rate of rejection of the (now known to be) false H_0. First of all, we observe that the powers are larger for $n = 100$ than for $n = 30$, which is a plausible result: Statistical inference is easier with more observations. For the AD test it is more difficult to reject the false H_0 in the case of $SC(0,1,3^2,0.1)$ than in the case of $SC(0,1,10^2,0.1)$ because in the latter case the gross errors have larger magnitude. In other words, $SC(0,1,10^2,0.1)$ is statistically more discriminable from $N(0,1)$ than $SC(0,1,3^2,0.1)$. For $L(0,1)$ the powers are mostly between $SC(0,1,3^2,0.1)$ and $SC(0,1,10^2,0.1)$.

Table 3. Rates of model selection (bold numbers are success rates). PRN - pseudo random number generator,N: normal distribution, L: Laplace distribution, G: generalized normal distribution, SC: scale contaminated normal distribution.

PRN	$N(0,\sigma^2)$	$L(0,\sigma^2)$	$G(0,\alpha,\beta)$	$SC(0,\sigma_1^2,\sigma_2^2,\varepsilon)$
	Rates of selected models for $n = 30$			
$N(0,1)$	**0.83**	0.17	0.00	0.00
$L(0,1)$	0.31	**0.64**	0.04	0.01
$SC(0,1,3,0.1)$	0.42	0.44	0.03	**0.11**
$SC(0,1,10,0.1)$	0.10	0.19	0.09	**0.62**
	Rates of selected models for $n = 100$			
$N(0,1)$	**0.96**	0.04	0.00	0.00
$L(0,1)$	0.07	**0.88**	0.03	0.02
$SC(0,1,3,0.1)$	0.20	0.53	0.01	**0.26**
$SC(0,1,10,0.1)$	0.00	0.00	0.01	**0.99**

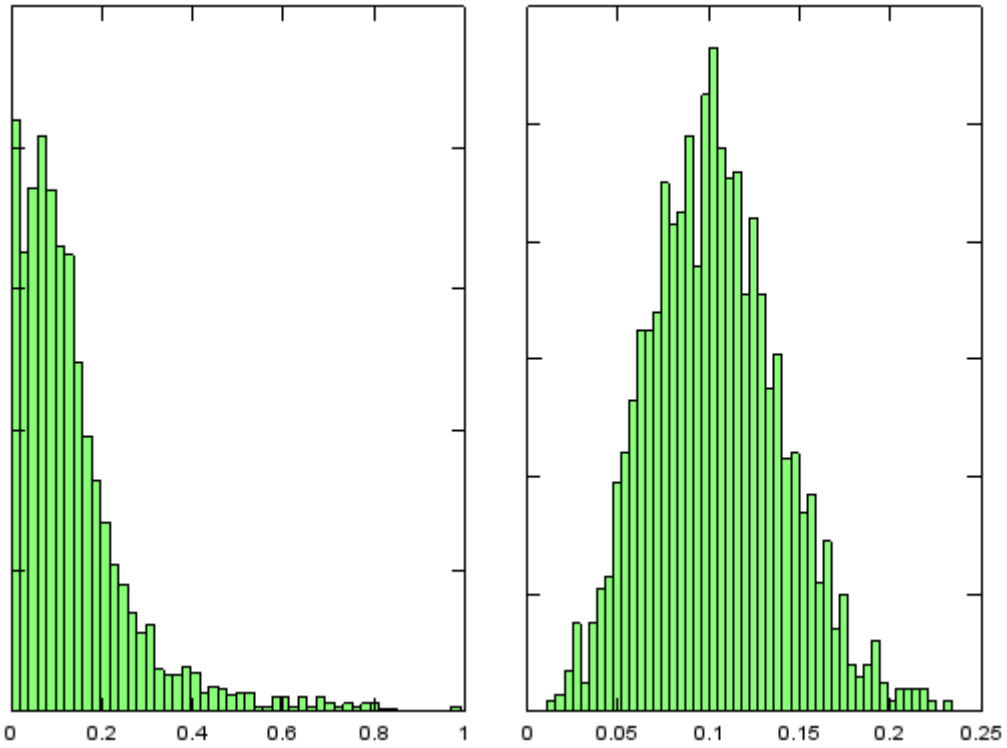

Fig. 2. Histograms of maximum likelihood estimates of the contamination parameter ε, see (10), from $n = 100$ observations. **Left**: weak contamination $SC(0,1,3^2,0.1)$ **right**: strong contamination $SC(0,1,10^2,0.1)$

Second, the observation error model selection by AIC is tried. For each observation vector we compute the ML solution $\hat{\theta}$ and therefrom the AIC by (1). The model with the minimum AIC is selected. The rates of selected models are given in Table 3. First of all, we observe that the rates of selecting the correct model are larger for $n = 100$ than for $n = 30$, which is again a plausible result.

Model selection is widely successful except for weakly scale contaminated observation errors $SC(0,1,3^2,0.1)$. Here the Laplacian observation error model is most often selected. A reason for this behavior can be concluded from figure 2. It is shown there that in the case of weak scale contamination the ML estimate of the contamination parameter ε is poor: These estimates scatter in the interval 0.0…0.5. Remember that the true value of ε is always 0.1. A similar drawing with $n = 30$ would even show a larger scattering. Summarizing, the parameters of $SC(0,1,3^2,0.1)$ are poorly recovered from the observations.

Next, it is interesting to compare the results of the AD test with the model selection by AIC. This is most easy for $n = 100$, where the selection rate of the normal model for the normal observation errors is 0.96, thus nearly matches the type I error rate of 0.05 for the AD test. The statistical power of 0.81 for $L(0,1)$ is exceeded by the corresponding success rates of model selection of 0.88. Moreover, not only is H_0 rejected, but also the correct alternative model is selected. For $SC(0,1,10^2,0.1)$ the power and success rate are both very high, such that here no advantage can be concluded for either method. For $SC(0,1,3^2,0.1)$ the AIC selects a normal distribution only with a rate of 0.20, while for the AD test it is $1 - 0.67 = 0.33$. This clearly is an advantage of AIC. However, the selection of the proper alternative model is less successful for the reasons explained above. This corresponds to what in statistics is called a type III error.

Finally, we must investigate, what the effect of model selection is on the parameter estimation of intercept x_0 and slope x_1. We expect that the estimated parameters are closer to their true values when the model is properly selected. For this investigation we restrict ourselves to $n = 100$ observations.

We compute the RMS of the estimation errors of the intercept x_0 and slope x_1 parameters

1. under the assumption that the correct model has always been chosen (which of course would practically be impossible),
2. after chosing the model by AD test with $\alpha = 0.10, 0.05, 0.01$ in such a way that the Laplace distribution is used whenever normality is rejected, and
3. after selecting the model by AIC

The RMS values are given in this order in columns 2-6 of Table 4. We see that in the case of normally distributed observation errors both AD test as well as model selection by AIC give satisfactory results. The results improve when α is chosen smaller because then the normal model is selected more often.

In the case of Laplacian observation errors the AIC gives the best results. They are even slightly better than using the L1 norm throughout, which might be surprising. The reason is that even though we generated Laplacian observation errors, in some occasions the L2 norm could produce a better fit and hence give better estimates of the parameters. The AIC would then select the better fitting normal model.

In the case of scale contaminated observation errors the AD test gives poor results because we selected an improper alternative model. This is particularly true when the contamination is strong. Here AIC performs better. For strong contamination we always select the true model such that the values in the second and sixth column coincide. It might be surprising that the estimation even gives

better results when the contamination is strong. The reason is that the estimation of the contamination parameter is easier in this case, see again Fig. 2.

It may also be surprising that fitting Laplacian observation errors is more successful by L1 norm than by L2 norm, when measured by RMS. Remember that L2 norm minimization as a "best linear unbiased estimation" (BLUE) is expected to give the least RMS values for the estimated parameters, independent of the error distribution. However, the emphasis in on "linear". A non-linear estimation like L1 norm minimization could perform better, even when measured by RMS. And here it does.

Table 4. Root mean square (RMS) values of the estimation errors of the intercept parameter x_0 and slope parameter x_1 in (13). PRN: pseudo random number generator, AD: Anderson-Darling, α: error rate, N: normal distribution, L: Laplace distribution, SC: scale contaminated normal distribution.

PRN	True model always used	AD test with α=0.10	AD test with α=0.05	AD test with α=0.01	Model selection by AIC
x_0					
$N(0,1)$	0.202	0.202	0.202	0.202	0.202
$L(0,1)$	0.194	0.194	0.194	0.195	0.190
$SC(0,1,3^2,0.1)$	0.230	0.270	0.270	0.270	0.251
$SC(0,1,10^2,0.1)$	0.224	0.529	0.529	0.529	0.224
x_1					
$N(0,1)$	0.00346	0.00353	0.00351	0.00349	0.00349
$L(0,1)$	0.00320	0.00342	0.00325	0.00331	0.00316
$SC(0,1,3^2,0.1)$	0.00399	0.00465	0.00464	0.00465	0.00435
$SC(0,1,10^2,0.1)$	0.00384	0.00821	0.00821	0.00821	0.00384

7 Conclusions

It has been shown that a proper observation error model can be selected not only by a statistical hypothesis test, but also by an information criterion like AIC.

The advantages of model selection by information criteria over hypothesis tests are:

1. It is not necessary to choose a significance level 1-α, where α is the type I decision error rate.
2. It is not necessary to compute any critical values.
3. In the case that the normal error model is not appropriate, the model selection by information criteria also yields the proper non-normal model like generalized or contaminated distributions. It is not necessary to invoke a multiple hypothesis test.

But there are also disadvantages of model selection by information criteria:

1. It does not support a statement like: "If the observation errors are truly normally distributed, this error model is chosen with probability 1-α."
2. The computational complexity is rather high. Not only the least squares (L2) fitting must be computed, but also other ML solutions like the L1 fitting (Laplace), the generalized normal and / or contaminated model fittings. The latter can be computationally demanding.

The first disadvantage should not be taken too seriously. We believe that such a statement is often dispensable because in any case the opposite statement regarding a type II error rate is not obtained. Also the second disadvantage is no longer an obstacle because today computing power is not the bottleneck. In the future, the numerical procedures for generalized normal and contaminated model fittings should be refined.

Therefore, we encourage geodesists, geophysicists and all other scientists and applied engineers to select error models by AIC. In the future we should investigate similar information criteria for observation error model selection.

References

Anderson TW, Darling DA (1952) Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Ann. Math. Stat. 23: 193–212. doi:10.1214/aoms/1177729437

Anderson TW, Darling DA (1954) A Test of Goodness-of-Fit. Journal of the American Statistical Association 49: 765–769. doi:10.2307/2281537

Akaike H (1974) A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19: 716–723

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: A Practical Information-theoretic Approach. Springer, Berlin. doi:10.1007/b97636

Cai J, Grafarend E, Hu C (2007) The statistical property of the GNSS carrier phase observations and its effects on the hypothesis testing of the related estimators. In: Proceedings of ION GNSS 2007, Fort Worth, TX, USA, Sept 25–28, 2007, pp 331–338

Cramér H (1928) On the composition of elementary errors. Skandinavisk Aktuarietidskrift, 11, 13-74, 141-180

D'Agostino RB (1970) Transformation to normality of the null distribution of g1. Biometrika, 57, 679-681, DOI: 10.1093/biomet/57.3.679

Hampel FR (2001) Robust statistics: A brief introduction and overview. In: Carosio A, Kutterer H (Eds) Proc. First Internat. Symposium on Robust Statistics and Fuzzy Techniques in Geodesy and GIS, Zurich March 2001

Huber PJ (1964) Robust Estimation of a Location Parameter, Ann. Stat. 53: 73–101

Huber PJ (2009) Robust Statistics (2nd ed.) John Wiley & Sons Inc, New York. ISBN 978-0-470-12990-6

Jarque CM, Bera AK (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ. Lett., 6, 255-259, doi:10.1016/0165-1765(80)90024-5.

Klees R, Ditmar P, Broersen P (2002)How to handle colored observation noise in large least-squares problems. J. Geodesy 76:629-640. doi:10.1007/s00190-010-0392-4

Kolmogorov A (1933) Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari, 4, 83-91 (in Italian)

Kutterer H (2001) Uncertainty assessment in geodetic data analysis. In: Carosio A, Kutterer H (Eds) Proc. First Internat. Symposium on Robust Statistics and Fuzzy Techniques in Geodesy and GIS, Zurich March 2001

Lehmann R (2012) Geodätische Fehlerrechnung mit der skalenkontaminierten Normalverteilung. Allgemeine Vermessungs-Nachrichten 5/2012. VDE-Verlag Offenbach (in German)

Lehmann R (2013) On the formulation of the alternative hypothesis for geodetic outlier detection. J. Geodesy 87(4) 373–386

Lehmann R (2014) Transformation model selection by multiple hypothesis testing. J. Geodesy 88(12)1117-1130. doi:10.1007/s00190-014-0747-3

Lehmann R, Lösler M (2015) Multiple outlier detection - hypothesis tests versus model selection by information criteria. J. Surv. Eng. (just released) doi:10.1061/(ASCE)SU.1943-5428.0000189

Lilliefors HW (1967) On the Kolmogorov-Smirnov for normality with mean and variance unknown. J. Am. Stat. Assoc., 62, 399-402

Luo X (2013) GPS Stochastic Modelling – Signal Quality Measures and ARMA Processes. Springer Berlin Heidelberg. doi:10.1007/978-3-642-34836-5

Luo X, Mayer M, Heck B (2011) On the probability distribution of GNSS carrier phase observations. GPS Solutions 15(4)369-379. doi:10.1007/s10291-010-0196-2

Miller RG (1981) Simultaneous statistical inference. Springer New York. ISBN:0-387-90548-0

Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag. Ser. 5, 50(302) 157-175, doi:10.1080/14786440009463897

Razali NM, Wah YB (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J. Stat. Model. Anal. 2(1)21-33

Shapiro SS, Francia RS (1972) An approximate analysis of variance test for normality. J. Am. Stat. Assoc., 67, 215-216, doi:10.1080/01621459.1972.10481232

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrica, 52, 591-611

Smirnov N (1948) Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat., 19, 279-281, doi:10.1214/aoms/1177730256

Stephens MA (1974) EDF Statistics for Goodness of Fit and Some Comparisons J. Amer. Stat. Assoc. 69: 730–737. doi:10.2307/2286009

Tanizaki H (2004) Computational methods in statistics and econometrics. Marcel Dekker, New York. ISBN-13: 978–0824748043

Teunissen PJG (2000) Testing theory; an introduction. Series on mathematical geodesy and positioning, 2nd Ed., Delft Univ. of Technology, Delft, Netherlands

Tiberius CCJM, Borre K (2000) Are GPS data normally distributed. In: KP Schwarz (Ed.) Geodesy Beyond 2000. International Association of Geodesy Symposia Volume 121, 2000, pp 243-248

Tukey JW (1960) A survey of sampling from contaminated distributions. In: Olkin I. (Ed.): Contributions to Probability and Statistics. University Press Stanford California

Verhagen S, Teunissen PJG (2005) On the probability density function of the GNSS ambiguity residuals. GPS Solutions 10(1)21-28. doi:10.1007/s10291-005-0148-4

von Mises R (1931) Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und Theoretischen Physik. F. Deutike, Leipzig, Germany (in German)

Wisniewski Z (2014) M-estimation with probabilistic models of geodetic observations. J. Geodesy 88:941–957. doi:10.1007/s00190-014-0735-7