J Geod (2013) 87:107–116
DOI 10.1007/s00190-012-0582-3

ORIGINAL ARTICLE

Robust estimation by expectation maximization algorithm

Karl Rudolf Koch

Received: 13 March 2012 / Accepted: 26 June 2012 / Published online: 12 July 2012

Abstract A mixture of normal distributions is assumed for the observations of a linear model. The first component of the mixture represents the measurements without gross errors, while each of the remaining components gives the distribution for an outlier. Missing data are introduced to deliver the information as to which observation belongs to which component. The unknown location parameters and the unknown scale parameter of the linear model are estimated by the EM algorithm, which is iteratively applied. The E (expectation) step of the algorithm determines the expected value of the likelihood function given the observations and the current estimate of the unknown parameters, while the M (maximization) step computes new estimates by maximizing the expectation of the likelihood function. In comparison to Huber's M-estimation, the EM algorithm does not only identify outliers by introducing small weights for large residuals but also estimates the outliers. They can be corrected by the parameters of the linear model freed from the distortions by gross errors. Monte Carlo methods with random variates from the normal distribution then give expectations, variances, covariances and confidence regions for functions of the parameters estimated by taking care of the outliers. The method is demonstrated by the analysis of measurements with gross errors of a laser scanner.

Keywords Linear model · Mixture of normal distributions · Huber's M-estimation · Monte Carlo methods · Confidence intervals

K.R. Koch (✉)
Institute of Geodesy and Geoinformation, Theoretical Geodesy
University of Bonn, Nussallee, 17, 53115 Bonn, Germany
e-mail: koch@geod.uni-bonn.de

1 Introduction

Tukey (1960) used the contaminated normal distribution as a statistical model for measurements containing outliers. It consists of a mixture of two normally distributed components, the first one representing with a large probability the measurements, the second one with a small probability the outliers. To derive the well known robust M-estimation, Huber (1964) started from the contaminated normal distribution and obtained the least informative distribution, which consists in the middle of the standard normal distribution and at the tails of the Laplace distribution (Huber 1981, p. 71). He solved the robust estimation of the parameters of a linear model by starting from a least squares estimate and then iteratively down-weighting large residuals. Absolute values of the residuals determine the weights (Huber 1981, p. 184).

This method has been used in geodesy for many years. For instance, Chang and Guo (2005) proposed a recursive Newton method for GPS positioning instead of the re-weighting; Khodabandeh et al. (2012) used Huber's method to analyze mathematical models for time series of GPS coordinates. The Danish method of robust estimation replaces Huber's weights by exponential functions of the squares of the residuals (Krarup et al. 1980). Koch and Yang (1998) derived a robust Kalman filter by Huber's M-estimation. The contaminated normal distribution was used for data snooping by Lehmann and Scheffler (2011).

Huber's M-estimation works well in linear models with observations, which control each other. Nevertheless, its breakdown point is zero, i.e., the smallest percentage of outliers, which causes the estimator to break down by producing wrong results is 0 %. This happens if outliers appear in leverage points in addition to outliers in the remaining points, cf. Koch (1999, p. 264). Yohai (1987) therefore proposed the robust MM-estimation. It starts with an estimation with a high breakdown point; for instance the LMS-estimation of Rousseeuw (1984), followed by a robust estimate of scale and then a robust M-estimation. Applying Monte Carlo methods, subsets of data are found without outliers in leverage points. The outlier search can therefore be started with an approximate parameter estimation not distorted by gross errors in leverage points (Koch, 2007b).

Variances, covariances and confidence regions of functions of unknown parameters estimated in a linear model can be readily computed by Monte Carlo methods. Confidence regions of parameters estimated by Huber's M-estimation have been computed by this method. Unfortunately, no procedures are known to generate random variates directly from Huber's distribution. The Gibbs sampler together with the rejection method and the Cauchy distribution as envelope for the standard normal distribution was therefore applied (Koch 2007a, p. 229). To find a simpler procedure, an alternative to Huber's M-estimation is developed here by the Expectation Maximization (EM) algorithm.

Dempster et al. (1977) derived the EM algorithm in its full generality after special cases had been proposed before. It is usually applied to the problem of missing or incomplete data, cf. Little and Rubin (2002, p. 166). The first problem arises when due to limitations of the observation process, data are missing. For the second one, incomplete data are assumed in order to make certain parameter estimations tractable. This is typical for the classification in pattern recognition where the observations do not contain the information as to which pattern they belong. The same task appears when observations with gross errors belong to a mixture of densities and the information is not available as to which component of the mixture they have to be attributed. This is the problem to be dealt with here.

The EM algorithm works iteratively with the E (expectation) step and the M (maximization) step. The E step determines the expectation of the likelihood function resulting from the distribution for the observed and missing data under the condition of given observations and current estimates of the parameters. A new set of parameters is estimated in the M step by maximizing the expected value of the likelihood function. Examples of the EM algorithm in the geodetic literature are the extraction of straight lines and parabolas from digital images by Luxen and Brunn (2003) and the robust estimation of parameters and variance components by Peng (2009). Outliers are introduced in the latter example as additive parameters with different variances. The outliers, however, have to be identified for each iteration of the EM algorithm.

A simpler model for the robust estimation in linear models is introduced here. It assumes a mixture of normal densities with a component for each suspected outlier. The first component possessing a large probability takes care of the observations without gross errors and the remaining components with small probabilities represent the outliers. The information as to which observation belongs to which component is furnished by the missing data. The EM algorithm for estimating the parameters of a linear model with two components, i.e., with one outlier, has been given without derivation by Aitkin and Wilson (1980). The estimates by the EM algorithm for a linear model with any number of outliers is derived here. This method identifies outliers by introducing small weights for large residuals like Huber's method. However, the outliers are estimated in addition and can be corrected by the parameter estimation. The resulting functions of the parameters are freed from the distortions of the outliers so that their variances, covariances and confidence regions can be computed by Monte Carlo methods based on the normal distributions. This is demonstrated for the data of a laser scanner with outliers.

The paper is organized as follows: Section 2 derives the EM algorithm for observations belonging to a mixture of distributions. Section 3 applies the EM algorithm to estimate the parameters of a linear model. Section 4 presents the equations for the analysis of the data of a laser scanner. Section 5 gives the numerical results and Sect. 6 the conclusions.

2 EM algorithm

Let $\boldsymbol{y}_{\text{obs}}$ be the vector of observations, $\boldsymbol{y}_{\text{mis}}$ the vector of missing data, $\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}$ the complete data, $\boldsymbol{\Theta}$ the vector of unknown parameters and $p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\Theta})$ the likelihood function, which is regarded as a function of $\boldsymbol{\Theta}$ for given $\boldsymbol{y}_{\text{obs}}$ and $\boldsymbol{y}_{\text{mis}}$. To apply the maximum likelihood estimation, it is often analytically easier to maximize the natural logarithm of the likelihood function so that we will work with $\log p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\Theta})$.

The E (expectation) step of the EM algorithm determines the conditional expectation, usually called $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t-1)})$, of the log-likelihood function $\log p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\Theta})$ given $\boldsymbol{y}_{\text{obs}}$ and the current estimate $\boldsymbol{\Theta}^{(t-1)}$ of the unknown parameters, cf. Little and Rubin (2002, p. 168), DasGupta (2011, p. 706),

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t-1)}) = \mathrm{E}[\log p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\Theta})|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\Theta}^{(t-1)}]. \quad (1)$$

The M (maximization) step of the EM algorithm determines the new estimate $\boldsymbol{\Theta}^{(t)}$ by maximizing (1)

$$\boldsymbol{\Theta}^{(t)} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t-1)}). \quad (2)$$

These two steps are repeated until $\boldsymbol{\Theta}^{(t)}$ converges, cf. Wu (1983).

The conditional expectation $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t-1)})$ is a function of $\boldsymbol{y}_{\text{mis}}$ since $\boldsymbol{y}_{\text{obs}}$ and $\boldsymbol{\Theta}^{(t-1)}$ are given. Let $p(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\Theta}^{(t-1)})$ be the conditional density function for $\boldsymbol{y}_{\text{mis}}$, the conditional expectation is then determined by:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t-1)}) = \int_{\mathcal{Y}_{\text{mis}}} \log p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\Theta}) \, p(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\Theta}^{(t-1)}) d\boldsymbol{y}_{\text{mis}} \tag{3}$$

where $\mathcal{Y}_{\text{mis}}$ denotes the domain of $\boldsymbol{y}_{\text{mis}}$. Thus, $\boldsymbol{y}_{\text{mis}}$ is integrated out so that the conditional expectation is substituted for the missing data.

Let the observations y_i with $\boldsymbol{y}_{\text{obs}} = (y_i)$ and $i \in \{1, \ldots, n\}$ be independent. Given $\boldsymbol{y}_{\text{obs}}$, the likelihood function $p(\boldsymbol{y}_{\text{obs}} |\boldsymbol{\Theta})$ then follows with:

$$p(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\Theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\Theta}). \tag{4}$$

As mentioned in the introduction, we assume a mixture of distributions for the observations. Let $p_j(y_i|\boldsymbol{\theta}_j)$ be the density for y_i of the component j with $j \in \{1, \ldots, m\}$ of the mixture, $\boldsymbol{\theta}_j$ the vector of unknown parameters of the distribution and α_j the unknown probability with which the component j contributes to the mixture, the likelihood function given y_i then follows by:

$$p(y_i|\boldsymbol{\Theta}) = \sum_{j=1}^{m} \alpha_j p_j(y_i|\boldsymbol{\theta}_j) \quad \text{with} \quad \sum_{j=1}^{m} \alpha_j = 1\,. \tag{5}$$

The vector $\boldsymbol{\Theta}$ of unknown parameters is defined by:

$$\boldsymbol{\Theta} = |\alpha_1, \ldots, \alpha_m, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m|'. \tag{6}$$

Substituting (5) in (4) yields

$$p(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\Theta}) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_j p_j(y_i|\boldsymbol{\theta}_j). \tag{7}$$

The missing data shall indicate as to which component j the observation y_i belongs to. The $n \times 1$ vector $\boldsymbol{y}_{\text{mis}}$ of missing data is defined by:

$$\boldsymbol{y}_{\text{mis}} = |\bar{y}_1, \ldots, \bar{y}_n|' \tag{8}$$

where $\bar{y}_i$ denotes a discrete random variable having the values $\bar{y}_i \in \{1, \ldots, m\}$. For instance, $\bar{y}_i = j$ means that the observation i belongs to component j. To derive the density for the discrete random variable $\bar{y}_i$ in (8), the probability $\alpha_{\bar{y}_i}$ is assumed as prior probability for the posterior probability $p(\bar{y}_i|y_i, \boldsymbol{\Theta}^{(t-1)})$ that the observation i with $\bar{y}_i = j$ comes from component j (Lange and Sinsheimer 1993). We therefore obtain with the likelihood function $p_{\bar{y}_i}(y_i|\boldsymbol{\theta}_{\bar{y}_i})$ for $\bar{y}_i$ by Bayes' theorem, cf. Koch (2007a, p. 35),

$$p(\bar{y}_i|y_i, \boldsymbol{\Theta}^{(t-1)}) = \frac{\alpha_{\bar{y}_i} p_{\bar{y}_i}(y_i|\boldsymbol{\theta}_{\bar{y}_i})}{\sum_{k=1}^{m} \alpha_{\bar{y}_k} p_{\bar{y}_k}(y_i|\boldsymbol{\theta}_{\bar{y}_k})} \tag{9}$$

and with (8) because of independency

$$p(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\Theta}^{(t-1)}) = \prod_{i=1}^{n} p(\bar{y}_i|y_i, \boldsymbol{\Theta}^{(t-1)}). \tag{10}$$

The log-likelihood function $\log p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\Theta})$ given the complete data $\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}$ is obtained with (7) and (10) by:

$$\begin{aligned}\log p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\Theta}) &= \log p(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\Theta}) p(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\Theta}^{(t-1)}) \\ &= \log \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_j p_j(y_i|\boldsymbol{\theta}_j) = \sum_{i=1}^{n} \log(\alpha_j p_j(y_i|\boldsymbol{\theta}_j)) \end{aligned} \tag{11}$$

where the summation over j disappears because given the missing data $\boldsymbol{y}_{\text{mis}}$ the component j is determined and $p(\boldsymbol{y}_{\text{mis}} |\boldsymbol{y}_{\text{obs}}, \boldsymbol{\Theta}^{(t-1)})$ is a constant, which is ignored.

Since $\boldsymbol{y}_{\text{mis}}$ is a discrete random vector, the integration in (3) is replaced by a summation when substituting (10) and (11)

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t-1)}) = \sum_{\bar{y}_1=1}^{m} \ldots \sum_{\bar{y}_n=1}^{m} \sum_{i=1}^{n} \log(\alpha_j p_j(y_i|\boldsymbol{\theta}_j)) \prod_{i=1}^{n} p\left(\bar{y}_i|y_i, \boldsymbol{\Theta}^{(t-1)}\right). \tag{12}$$

By replacing the random variable $\bar{y}_i$ by the value j, it can take on, we obtain (Bilmes 1998)

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t-1)}) &= \sum_{j=1}^{m} \sum_{i=1}^{n} \log(\alpha_j p_j(y_i|\boldsymbol{\theta}_j)) p(j|y_i, \boldsymbol{\Theta}^{(t-1)}) \\ &= \sum_{j=1}^{m} \sum_{i=1}^{n} \log(\alpha_j) p(j|y_i, \boldsymbol{\Theta}^{(t-1)}) \\ &\quad + \sum_{j=1}^{m} \sum_{i=1}^{n} \log(p_j(y_i|\boldsymbol{\theta}_j)) p(j|y_i, \boldsymbol{\Theta}^{(t-1)}). \end{aligned} \tag{13}$$

This expression has to be maximized for the M step. The first term of (13) containing the unknown parameters α_j in (6) can be maximized independently from the second term containing the unknown parameters $\boldsymbol{\theta}_j$. Because of $\sum_{j=1}^{m} \alpha_j = 1$ in (5), we introduce the Lagrange multiplier λ and set the derivative equal to zero to find an extreme value

$$\frac{\partial}{\partial \alpha_k} \left[\sum_{j=1}^{m} \sum_{i=1}^{n} \log(\alpha_j) p(j|y_i, \boldsymbol{\Theta}^{(t-1)}) + \lambda \left(\sum_{j=1}^{m} \alpha_j - 1 \right) \right] = 0 \tag{14}$$

which gives:

$$\sum_{i=1}^{n} p(k|y_i, \boldsymbol{\Theta}^{(t-1)}) = -\alpha_k \lambda. \tag{15}$$

One gets $\lambda = -n$ because of (9) by summing both sides of (15) over k. The estimate $\hat{\alpha}_j$ of the unknown probability α_j in (6), with which the component j contributes to the mixture, therefore follows with:

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} p(j|y_i, \boldsymbol{\Theta}^{(t-1)}) \tag{16}$$

as mean of the posterior probabilities that the observations y_i belong to the components j.

To maximize the second term of (13) for estimating the unknown parameters $\boldsymbol{\theta}_j$, the relation between the observation y_i and $\boldsymbol{\theta}_j$ and the density $p_j(y_i|\boldsymbol{\theta}_j)$ have to be defined.

3 Outliers in observations of a linear model

Let the linear model be given

$$\begin{aligned} &\tilde{\boldsymbol{X}}\boldsymbol{\beta} = E(\tilde{\boldsymbol{y}}) = \tilde{\boldsymbol{y}} + \tilde{\boldsymbol{e}} \quad \text{with} \\ &\quad E(\tilde{\boldsymbol{e}}) = \boldsymbol{0} \quad \text{and} \quad D(\tilde{\boldsymbol{y}}) = \sigma^2\boldsymbol{\Sigma} \end{aligned} \tag{17}$$

where $\tilde{\boldsymbol{X}}$ denotes the $n \times u$ matrix of coefficients with full column rank, $\boldsymbol{\beta}$ the $u \times 1$ vector of unknown parameters, $\tilde{\boldsymbol{y}}$ the $n \times 1$ vector of observations, $\tilde{\boldsymbol{e}}$ the $n \times 1$ vector of errors, σ^2 the unknown variance factor and $\sigma^2\boldsymbol{\Sigma}$ the $n \times n$ positive definite covariance matrix of the observations. The Cholesky factorization of $\boldsymbol{\Sigma}^{-1}$

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{G}\boldsymbol{G}', \tag{18}$$

where $\boldsymbol{G}$ denotes a lower triangular matrix, is applied for a decorrelation of the vector $\tilde{\boldsymbol{y}}$ of observations. Without changing the vector $\boldsymbol{\beta}$, the linear model (17) is transformed by:

$$\boldsymbol{X} = \boldsymbol{G}'\tilde{\boldsymbol{X}},\ \boldsymbol{y}_{\text{obs}} = \boldsymbol{G}'\tilde{\boldsymbol{y}},\ \boldsymbol{e}_{\text{obs}} = \boldsymbol{G}'\tilde{\boldsymbol{e}} \tag{19}$$

into the model

$$\begin{aligned} &\boldsymbol{X}\boldsymbol{\beta} = E(\boldsymbol{y}_{\text{obs}}) = \boldsymbol{y}_{\text{obs}} + \boldsymbol{e}_{\text{obs}} \quad \text{with} \\ &\quad E(\boldsymbol{e}_{\text{obs}}) = \boldsymbol{0} \quad \text{and} \quad D(\boldsymbol{y}_{\text{obs}}) = \sigma^2\boldsymbol{I}. \end{aligned} \tag{20}$$

Independent observations $\tilde{y}_i$ with $\tilde{\boldsymbol{y}} = (\tilde{y}_i)$ are assumed in agreement with (4). They shall have different variances, thus, cf. Koch (1999, p. 155),

$$\boldsymbol{\Sigma} = \text{diag}(\sigma^2_{\tilde{y}_1}, \ldots, \sigma^2_{\tilde{y}_n}) \text{ and } \boldsymbol{G} = \text{diag}(1/\sigma_{\tilde{y}_1}, \ldots, 1/\sigma_{\tilde{y}_n}). \tag{21}$$

The estimate $\hat{\boldsymbol{\beta}}$ of the unknown parameters $\boldsymbol{\beta}$ in model (20) follows from:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_{\text{obs}}. \tag{22}$$

It is identical with the estimate of $\boldsymbol{\beta}$ in model (17) for substituting (19) in (22) gives

$$\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{y}}. \tag{23}$$

The vector $\hat{\boldsymbol{e}}_{\text{obs}} = (\hat{e}_i)$ of residuals is obtained with $\boldsymbol{X} = (\boldsymbol{x}_i')$ and $\boldsymbol{y}_{\text{obs}} = (y_i)$ from:

$$\hat{\boldsymbol{e}}_{\text{obs}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}_{\text{obs}} \quad \text{and} \quad \hat{e}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} - y_i. \tag{24}$$

In agreement with the EM algorithm, the variance factor σ^2 is estimated with $\hat{\sigma}^2$ by the maximum-likelihood method

$$\hat{\sigma}^2 = \hat{\boldsymbol{e}}'_{\text{obs}}\hat{\boldsymbol{e}}_{\text{obs}}/n. \tag{25}$$

It gives slightly smaller values than the unbiased estimate where n is replaced by $n - u$. The residuals $\hat{\tilde{\boldsymbol{e}}}$ in model (17) are computed with (19) by:

$$\hat{\tilde{\boldsymbol{e}}} = (\boldsymbol{G}')^{-1}\hat{\boldsymbol{e}}_{\text{obs}}. \tag{26}$$

Since independent observations $\tilde{y}_i$ with different variances are introduced, the decorrelation of the vector $\tilde{\boldsymbol{y}}$ by (19) is applied only to obtain observations y_i with equal variances of one. The decomposition of the matrix $\boldsymbol{\Sigma}^{-1}$ into its eigenvalues and eigenvectors could have been used for the decorrelation by a transformation like (19). This gives a different coefficient matrix $\boldsymbol{X}$ and different vectors $\boldsymbol{y}_{\text{obs}}$ and $\boldsymbol{e}_{\text{obs}}$ for model (20). However, the estimate $\hat{\boldsymbol{\beta}}$ of the unknown parameters from (22) is identical because substituting the transformation in (22) gives the estimate (23) for model (17). Thus, the estimate $\hat{\boldsymbol{\beta}}$ is invariant with respect to the method of decorrelation. The Cholesky factorization is to be preferred because it is simpler to compute than the decomposition into eigenvalues. To estimate the integer ambiguity of the GPS double difference phase measurements, the decorrelation is also applied. The Cholesky factorization for the decorrelation is the most efficient one among three different methods (Wang et al. 2010).

A mixture (5) of normal distributions is assumed for $\boldsymbol{y}_{\text{obs}}$. We obtain for $p_j(y_i|\boldsymbol{\theta}_j)$ after omitting $\boldsymbol{\theta}_j$ for a simplified notation

$$\begin{array}{rcl} p_1(y_i) &=& \dfrac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\dfrac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right) \\ p_2(y_i) &=& \dfrac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\dfrac{1}{2\sigma^2}(y_i - \mu_2)^2\right) \\ \multicolumn{3}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \\ p_m(y_i) &=& \dfrac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\dfrac{1}{2\sigma^2}(y_i - \mu_m)^2\right) \end{array} \tag{27}$$

where $p_1(y_i)$ denotes the distribution for an observation y_i without a gross error and $p_2(y_i)$ to $p_m(y_i)$ the distributions for observations y_i with gross errors and with expectations from μ_2 to μ_m. The different variances for the original observations $\tilde{\boldsymbol{y}}$ with and without gross errors in model (17) have been taken care of by the transformation (19). Thus, only the variance factor σ^2 needs to be considered in (27).

The density for the missing data follows from (9). By using the simplified notation of (13) and expressing the dependency on the current estimate $\boldsymbol{\Theta}^{(t-1)}$ by:

$$\hat{p}(j|y_i) = p(j|y_i, \boldsymbol{\Theta}^{(t-1)}) \tag{28}$$

we obtain with (27)

$$\hat{p}(j|y_i) = \frac{\hat{\alpha}_j p_j(y_i)}{\sum_{k=1}^{m}\hat{\alpha}_k p_k(y_i)}. \tag{29}$$

The estimate $\hat{\alpha}_j$ of α_j from (16) yields

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{p}(j|y_i) \tag{30}$$

and it holds because of (29)

$$\sum_{j=1}^{m} \hat{p}(j|y_i) = 1, \tag{31}$$

therefore,

$$\sum_{j=1}^{m} \hat{\alpha}_j = 1. \tag{32}$$

The distributions of the second term of (13) are now determined so that it can be maximized. Substituting (27) and (28) gives:

$$\begin{aligned}&\sum_{i=1}^{n}\Big(-\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\Big)\hat{p}(1|y_i)\\ &+\sum_{i=1}^{n}\Big(-\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - \mu_2)^2\Big)\hat{p}(2|y_i)\\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\ &+\sum_{i=1}^{n}\Big(-\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - \mu_m)^2\Big)\hat{p}(m|y_i)\end{aligned} \tag{33}$$

where constants are ignored. The vector $\boldsymbol{\beta}$ and the expected values μ_2 to μ_m are the unknown location parameters and σ^2 the unknown scale parameter. Differentiating (33) with respect to $\boldsymbol{\beta}$ and setting the result equal to zero gives the estimate $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ of model (20), cf. Koch (1999, p. 161),

$$\begin{aligned}&\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}_{\text{obs}} \quad \text{with}\\ &\boldsymbol{W} = \text{diag}(\hat{p}(1|y_1), \dots, \hat{p}(1|y_n)).\end{aligned} \tag{34}$$

The estimate $\hat{\mu}_j$ of μ_j follows with:

$$\hat{\mu}_j = \sum_{i=1}^{n} y_i\hat{p}(j|y_i) \Big/ \sum_{i=1}^{n}\hat{p}(j|y_i) \quad \text{for} \quad j \in \{2, 3, \dots, m\}. \tag{35}$$

The residuals $\hat{\boldsymbol{e}}_{\text{obs}}$ are obtained with $\hat{\boldsymbol{\beta}}$ from (34) according to (24). The maximum-likelihood estimate $\hat{\sigma}^2$ of σ^2 results with (31) by:

$$\begin{aligned}\hat{\sigma}^2 = \frac{1}{n}\Big[&\sum_{i=1}^{n}\Big((y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2\hat{p}(1|y_i)\\ &+\sum_{j=2}^{m}(y_i - \hat{\mu}_j)^2\hat{p}(j|y_i)\Big)\Big].\end{aligned} \tag{36}$$

The estimates (30) and (34) to (36) are iteratively applied together with (29), which result from (27) after substituting (34) to (36).

To start the iterations, one or more observations have to be selected which might contain gross errors. Observations with largest absolute values of the residuals might be chosen as done for Huber's M-estimation. This method works well if the observations control each other like the observations of the example of Sect. 4. Otherwise, largest studentized residuals are candidates. They are computed by dividing the residuals by their standard deviations times the square root of the unbiased estimate of the variance factor. These residuals are the test statistics for the τ-test by Pope (1976), cf. Koch (1999, p. 305). If gross errors are suspected, for instance in the two observations y_k and y_l, we obtain $m = 3$ components in the mixture (5) and set as first estimate in (29)

$$\begin{aligned}&\hat{p}(1|y_i) = 0 \quad \text{for } i = k, l\\ &\hat{p}(1|y_i) = 1 \quad \text{for } i \neq k, l\\ &\hat{p}(2|y_i) = 0 \quad \text{for } \ i \neq k\\ &\hat{p}(2|y_i) = 1 \quad \text{for } \ i = k\\ &\hat{p}(3|y_i) = 0 \quad \text{for } \ i \neq l\\ &\hat{p}(3|y_i) = 1 \quad \text{for } \ i = l.\end{aligned} \tag{37}$$

Thus, the observations y_k and y_l do not contribute in the first iteration to the estimate $\hat{\boldsymbol{\beta}}$ by (34) and $\hat{\mu}_2 = y_k$ and $\hat{\mu}_3 = y_l$ are obtained by: (35).

Setting the partial derivatives of (33) with respect to the unknown parameters equal to zero leads to local extrema only. To check whether the log-likelihood function (13) is maximized and whether the iterations converge, (13) is computed for each iteration by substituting (28), (30) and (33) to (36)

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) = n\sum_{j=1}^{m}(\log(\hat{\alpha}_j)\hat{\alpha}_j) - \frac{n}{2}(\log\hat{\sigma}^2 + 1). \tag{38}$$

Wu (1983) shows that for exponential families of distributions, which are used here, $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)})$ converges to local maxima. However convergence cannot always be expected. The distributions introduced by (27) have to represent the data. If observations containing gross errors are not introduced as outliers and vice versa, divergence of the iterations has to be expected.

The number $m-1$ of outliers is unknown, but it is restricted by:

$$m - 1 < n/2 \tag{39}$$

since for a larger number of outliers one cannot distinguish between observations and outliers anymore. To detect the outliers, the EM algorithm should therefore first be applied by choosing the observation with the maximum absolute value for the residual or for the studentized residual as outlier. Then, one by one additional observations with decreasing absolute values for their residuals can be introduced as outliers. If the EM algorithm converges, it is checked whether the first or the additional observation, say y_i, is confirmed by the algorithm as an outlier. This happens, if

$$\hat{p}(1|y_i) < 0.005 \tag{40}$$

since the observation y_i is then sufficiently down-weighted in (34) so it does not contribute to the estimate $\hat{\boldsymbol{\beta}}$. The addition of outliers is stopped, if (40) is not fulfilled, if the EM algorithm diverges or if the upper limit (39) is reached.

4 Analysis of the measurements of a laser scanner

The rectangular coordinates of points on the surface of a planar metal sheet attached to a vertical board were measured by a laser scanner in the coordinate system of the instrument. The $\tilde{x}$-axis lies horizontally, the $\tilde{y}$-axis coincides with the center of lines of sight of the instrument and the $\tilde{z}$-axis points to the zenith. A grid of $n_g \times n_g = n$ points was determined so that the coordinates $\tilde{x}_i, \tilde{y}_i, \tilde{z}_i$ with $i \in \{1, \ldots, n\}$ were measured, where $\tilde{y}_i$ approximates the distance of the instrument to the metal sheet. In addition to these observations, the grid of points was measured with n_w repetitions to determine the variances of the measured coordinates. With the mean value $\tilde{y}_{mi}$ of the repeatedly measured coordinates $\tilde{y}_{ij}$ for $j \in \{1, \ldots, n_w\}$ from:

$$\tilde{y}_{mi} = \frac{1}{n_w} \sum_{j=1}^{n_w} \tilde{y}_{ij}, \tag{41}$$

we get the variance $\sigma^2_{\tilde{y}_i}$ of $\tilde{y}_i$ by:

$$\sigma^2_{\tilde{y}_i} = \frac{1}{n_w - 1} \sum_{j=1}^{n_w} (\tilde{y}_{ij} - \tilde{y}_{mi})^2 \tag{42}$$

and accordingly $\sigma^2_{\tilde{x}_i}$ and $\sigma^2_{\tilde{z}_i}$. As will be shown in Sect. 5, $\sigma^2_{\tilde{x}_i}$ and $\sigma^2_{\tilde{z}_i}$ are of an order of magnitude smaller than $\sigma^2_{\tilde{y}_i}$. Only $\tilde{y}_i$ is therefore considered as measurement and $\tilde{x}_i$ as well as $\tilde{z}_i$ are introduced as fixed quantities.

A plane metal sheet has been observed, and a plane with parameters $\beta_0, \beta_1, \beta_2$ is fitted to the measurements $\tilde{y}_i$ by the observation equations:

$$\beta_0 + \tilde{x}_i\beta_1 + \tilde{z}_i\beta_2 = \tilde{y}_i + \tilde{e}_i \quad \text{for} \quad i \in \{1, \ldots, n\}. \tag{43}$$

We set

$$\boldsymbol{\beta} = |\beta_0, \beta_1, \beta_2|', \ \tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}'_i), \ \tilde{\boldsymbol{x}}'_i = |1, \tilde{x}_i, \tilde{z}_i| \tag{44}$$

and assume the observations $\tilde{\boldsymbol{y}} = (\tilde{y}_i)$ as independent so that the linear model (17) is obtained with $\sigma^2_{\tilde{y}_i}$ in (21) from (42). It is transformed by (19) into the model (20) where the EM algorithm is applied.

By selecting the first observation, which is suspected to contain a gross error, the probabilities $\hat{p}(j|y_i)$ are determined by (37). The first iteration is started with these values by computing (30) and (34) to (36). This leads to (29) for the next iteration. This process is repeated until the iterations converge or diverge which is controlled by computing (38). In case of convergence, it is checked by (40) if the first outlier is determined. If not, the observations do not contain gross errors. If there is an outlier, a second observation possibly containing a gross error is added. The EM algorithm is applied again and in case of convergence, it is checked by (40) whether an outlier has been found.

To compare the results of the parameter estimation with and without taking care of gross errors, the expected value, the standard deviation and the confidence interval of the sum of adjusted distances is computed by Monte Carlo methods. As mentioned for (20), the observations $\tilde{y}_i$ are independent and normally distributed. Random variates $\tilde{y}_{ik}$ for $\tilde{y}_i$ with $k \in \{1, \ldots, o\}$ are therefore generated with (42) by, cf. Koch (2007a, p. 197),

$$\tilde{y}_{ik} = \hat{\sigma}\sigma_{\tilde{y}_i} z + \tilde{y}_i \tag{45}$$

where $\hat{\sigma}$ results from (25) or (36) and z denotes a random variate for the random variable $Z \sim N(0, 1)$.

If we do not take care of the gross errors, random variates $\boldsymbol{\beta}_k$ for the unknown parameters $\boldsymbol{\beta}$ are computed by substituting (45) in (22). Random variates s_k for the sum s of the adjusted distances from the origin of the local $\tilde{x}$-, $\tilde{y}$-, $\tilde{z}$-coordinate system of the laser scanner to the grid of n points on the adjusted plane follows with the random variates $\tilde{\boldsymbol{x}}'_i\boldsymbol{\beta}_k$ for the adjusted observations from (43) and (44) by

$$s_k = \sum_{i=1}^{n} (\tilde{x}_i^2 + (\tilde{\boldsymbol{x}}'_i\boldsymbol{\beta}_k)^2 + \tilde{z}_i^2)^{1/2}. \tag{46}$$

If outliers are determined, the random variates $\tilde{y}_{ik}$ with $i \in \{1, \ldots, n\}$ and k fixed are used in (34), (35), (36) for each iteration of the EM algorithm. This is repeated for $k \in \{1, \ldots, o\}$. Random variates s_k for the sum s then follow by (46). It means that the outliers are eliminated because they are down-weighted in (34).

The Monte Carlo estimate $\hat{E}(s)$ of the expectation $E(s)$ of the sum s follows from:

$$\hat{E}(s) = \frac{1}{o} \sum_{k=1}^{o} s_k \tag{47}$$

and the estimate $[\hat{V}(s)]^{1/2}$ of the standard deviation $[V(s)]^{1/2}$ of s from:

$$[\hat{V}(s)]^{1/2} = \left[\frac{1}{o} \sum_{k=1}^{o} (s_k - \hat{E}(s))^2\right]^{1/2}. \tag{48}$$

If there is more than one function of the parameters, covariances are estimated accordingly. A 0.95 Bayesian confidence interval is determined for the sum s, since it has minimum length, cf. Koch (2007a, p. 71), and it is readily determined

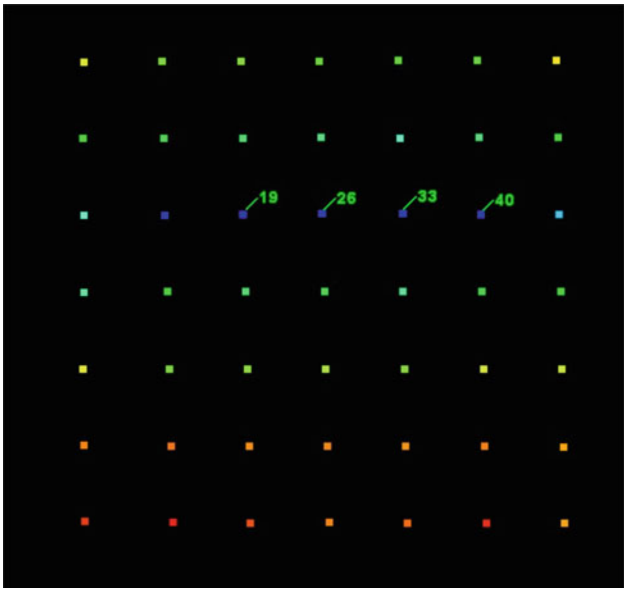

Fig. 1 Grid of 7 × 7 measured points on the metal sheet seen from the instrument. The points with largest residuals are numbered, see Fig. 2

by Monte Carlo methods. The probability of s lying within a cell of suitable width is computed by the relative frequency from the random variates s_k. The probabilities at both ends of the histogram are added such that they remain equal until the probability α of the $1 - \alpha = 0.95$ confidence interval is reached. For o random variates, $o - 1$ cells are chosen. To get the confidence limits with at least two significant digits, $o = 100{,}000$ random variates are generated (Koch et al. 2010, 2012).

5 Numerical results

A surface best qualified to be observed by a laser scanner is a Lambertian surface, i.e., an ideal diffusely reflecting surface, cf. Wagner (2010). To study the influence of a strongly reflecting surface on the measurements, the coordinates $\tilde{x}_i, \tilde{y}_i, \tilde{z}_i$ of a grid of 7×7 points on a planar, vertically standing metal sheet were measured by the laser scanner Leica HDS 3000. The points as seen from the instrument are shown in Fig. 1. The measured intensities of the reflected laser beam are depicted by colors. The highest intensity is 764 shown in dark blue in Fig. 1; the smallest one is 66 in dark red.

The distances between the points on the metal sheet amount to about 8 cm. The smallest $\tilde{y}_i$-coordinate is 5.38 m. The standard deviations for the measured coordinates are determined for $n_w = 25$ repetitions by (42) with 0.01 mm $\leq \sigma_{\tilde{x}_i} \leq$ 0.13 mm, 1.2 mm $\leq \sigma_{\tilde{y}_i} \leq$ 2.2 mm, 0.01 mm $\leq \sigma_{\tilde{z}_i} \leq$ 0.29 mm. The coordinates $\tilde{x}_i$ and $\tilde{z}_i$ are therefore considered fixed contrary to the coordinate $\tilde{y}_i$.

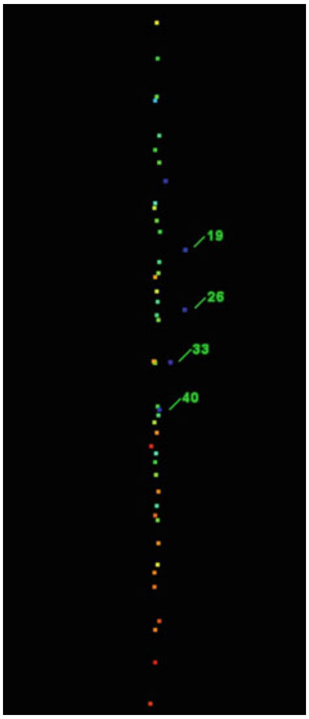

Fig. 2 Residuals of fitting a plane to the 7 × 7 points seen along the plane from the lower left corner of Fig. 1. The points with largest residuals are numbered

Table 1 Residuals $\hat{\tilde{e}}_i$ ordered by decreasing absolute values, standard deviations $\sigma_{\tilde{y}_i}$ and intensities

Point	$\hat{\tilde{e}}_i\,(cm)$	$\sigma_{\tilde{y}_i}$ (cm)	Intensities
26	2.76	0.21	733
33	2.64	0.19	724
19	1.41	0.20	763
40	0.62	0.22	764
13	−0.49	0.15	452
1	−0.48	0.15	103
46	−0.25	0.16	413

The plane fitted to the 49 $\tilde{y}_i$-coordinates by (22) is well determined even if some outliers exist in the data. The residuals $\hat{\tilde{\boldsymbol{e}}} = (\hat{\tilde{e}}_i)$ from (26) are therefore used to identify possible outliers. They are ordered by decreasing absolute values and given for 7 points in Table 1 together with the standard deviations $\sigma_{\tilde{y}_i}$ and the intensities. The residuals for the first 4 points are also shown in Fig. 2 by a view from the lower left corner of Fig. 1 along the fitted plane. The numbering of the points starts from the lower left corner of Fig. 1 and continues along the columns from left to right.

The three largest residuals in Table 1 indicate outliers. They are obviously caused by the high intensities of the reflected laser beam and lead to measured coordinates $\tilde{y}_i$, which are too short. The high intensities cause only a small

Table 2 Analysis of measurements

$\hat{\sigma}$	Sum s [m] of adjusted distances		
	$\hat{E}(s)$	$[\hat{V}(s)]^{1/2}$	0.95 Confidence interval
(a) Fitting a plane, gross errors included			
3.1806	265.425	0.037	$265.352 \le s \le 265.497$

Gross error in observ. #		Iter.	$\hat{\sigma}$	Sum s [m] of adjusted distance		
Introduced	Detected			$\hat{E}(s)$	$[\hat{V}(s)]^{1/2}$	0.95 Confidence interval
(b) EM algorithm						
26	26,33,19	18	1.4333	265.471	0.022	$265.429 \le s \le 265.513$
26,33	26,33,19	24	1.0472	265.474	0.022	$265.434 \le s \le 265.513$
26,33,19	26,33,19	12	1.0472	265.481	0.013	$265.454 \le s \le 265.506$
26,33,19,40		Div.				

increase of the standard deviations $\sigma_{\tilde{y}_i}$. Outliers cannot be detected by $\sigma_{\tilde{y}_i}$ or by the intensities. Fitting a plane by (22) gives the square root $\hat{\sigma}$ of the estimate $\hat{\sigma}^2$ of the variance factor from (25) which is considerably larger than 1.0, see Table 2a. It is caused by the large residuals. Table 2a also shows the expectation (47), the standard deviation (48) and the 0.95 confidence interval from Monte Carlo methods for the sum s from (46) of adjusted distances.

To detect outliers, the EM algorithm was applied. First, only the point 26 with the largest residual was introduced as outlier. The probabilities $\hat{p}(j|y_i)$ were therefore chosen according to (37) as first estimates. The EM algorithm converged after 18 iterations as stated in the first line of Table 2b. The probabilities $\hat{p}(1|y_i)$ then gave values smaller than 0.004 for points 26, 33 and 19 and values equal to 1.000 for the rest of the points. Thus, points 26, 33 and 19 were identified as outliers by (40). The probabilities $\hat{p}(2|y_i)$ were equal to 1.00 for points 26, 33 and 19 and smaller than 0.0002 for the rest of the points. The square root $\hat{\sigma}$ of the estimate $\hat{\sigma}^2$ of the variance factor from (36) in the first line of Table 2b is considerably smaller than that of Table 2a since the observations of points 26, 33 and 19 are down-weighted by the estimate (34).

Furthermore, the expectation (47), the standard deviation (48) and the confidence interval for the sum s of adjusted distances from (46) are given in Table 2b. The expectation is 4.6 cm larger than that of Table 2a because the measured coordinates $\tilde{y}_i$ of points 26, 33 and 19, which are too short, are corrected with (46) by the EM algorithm. The standard deviation is smaller than that of Table 2a and the confidence interval is shorter. This documents the gain in accuracy by correcting the measurements containing gross errors.

The points 26 and 33 were then introduced as outliers after choosing the first estimates of the probabilities $\hat{p}(j|y_i)$ according to (37). Again, the points 26, 33 and 19 were identified as outliers, which is indicated in the second line of Table 2b. The value for $\hat{\sigma}$ now approximates 1.0, and the expectation of s only changes by 0.3 cm against the previous value. The third line of Table 2b shows the result when the points 26, 33 and 19 were introduced as outliers. All three points were confirmed as having gross errors by the EM algorithm. Finally, the 4th point 40 of Table 1 was assumed as an additional outlier. The EM algorithm diverged as stated in the 4th line of Table 2b. One therefore has to conclude that only the points 26, 33 and 19 contain gross errors.

In addition to the three outliers in the measurements, six outliers were introduced by adding random values of equal probabilities between 2.8 and 1.4 cm, which are the residuals of points 26 and 19. The sign of these values was also randomly chosen. Table 3a shows the results for fitting the plane by (22) with the outliers included. The value for $\hat{\sigma}$ increases in comparison to Table 2a, the expectation of s changes by 3.4 cm, and the standard deviation and the confidence interval increase. The residuals of the fit of the plane were ordered by decreasing absolute values and the first 6 points with largest residuals and then all 9 points were introduced as outliers into the EM algorithm. Table 3b gives the results. In the case of 6 points, only 8 points were identified as outliers by (40), while introducing all 9 points confirmed these points as having gross errors. The second line of Table 3b shows that $\hat{\sigma}$ drops below 1.0 because only 40 observations $\tilde{y}_i$ now enter the estimation by (34). The expectation of s changes by only 1.1 cm in comparison to the third line of Table 2b.

Finally, Table 4 gives the results for introducing 21 outliers in addition to the three existing ones. This is the maximum number of outliers which can be chosen according to (39). The expected value of s in Table 4a differs from the one of Table 2a by 7.2 cm. If 12 outliers were introduced into the EM algorithm, all 24 outliers were detected as shown in the first line of Table 4b. The expectation of s only differs from the one in the third line of Table 2b by 0.8 cm. By comparing this value with 7.2 cm, which results from Table 4a, one rec-

Table 3 Measurements with 6 outliers in addition to the 3 existing ones

$\hat{\sigma}$	Sum s [m] of adjusted distances		
	$\hat{E}(s)$	$[\hat{V}(s)]^{1/2}$	0.95 Confidence interval
(a) Fitting a plane, gross errors included			
4.8164	265.391	0.056	$265.281 \leq s \leq 265.502$

9 Gross error in observ.		Iter.	$\hat{\sigma}$	Sum s [m] of adjusted distance		
Introduced	Detected			$\hat{E}(s)$	$[\hat{V}(s)]^{1/2}$	0.95 Confidence interval
(b) EM algorithm						
6	8	30	1.1848	265.482	0.016	$265.451 \leq \mathrm{s} \leq 265.513$
9	9	40	0.8471	265.470	0.014	$265.444 \leq \mathrm{s} \leq 265.498$

Table 4 Measurements with 21 outliers in addition to the 3 existing ones

$\hat{\sigma}$	Sum s [m] of adjusted distances		
	$\hat{E}(s)$	$[\hat{V}(s)]^{1/2}$	0.95 Confidence interval
(a) Fitting a plane, gross errors included			
8.2099	265.353	0.096	$265.165 \leq s \leq 265.541$

24 Gross error in observ.		Iter.	$\hat{\sigma}$	Sum s [m] of adjusted distance		
Introduced	Detected			$\hat{E}(s)$	$[\hat{V}(s)]^{1/2}$	0.95 Confidence interval
(b) EM algorithm						
12	24	48	0.7760	265.489	0.034	$265.446 \leq s \leq 265.562$
24		Div.				

ognizes how the outliers distort the expectation of s, while the EM algorithm corrects the outliers and keeps the expectation of s approximately the same. If all 24 outliers were introduced into the EM algorithm, it diverged as expressed in the second line of Table 4b. These many outliers were not accepted by the EM algorithm.

6 Conclusions

It is demonstrated by an example that the EM algorithm detects a large number of gross errors in the observations of a linear model. Of course, the perfect estimate of the plane by the data of the laser scanner is helpful. Outliers are not only detected but are also estimated and corrected by the estimates not distorted by gross errors. Due to the normal distributions, expected values, variances, covariances and confidence regions are readily computed by Monte Carlo methods. It is shown that the expectation of the sum of adjusted distances is distorted by outliers while it stays almost fixed if the gross errors are corrected. The standard deviation of the sum and its confidence interval increase considerably with the number of outliers, the increase is moderate if the outliers are corrected. Huber's M-estimation also identifies the outliers based on the least informative distribution. However, random variates for this distribution are difficult to generate so that subsequent Monte Carlo methods are not as simple as for the mixture of normal distributions used for the EM algorithm.

Acknowledgments The author is indebted to Boris Kargoll for valuable comments and to Ernst-Martin Blome for assistance with the measurements and the graphical depiction of the results.

References

Aitkin M, Wilson GT (1980) Mixture models, outliers, and the EM algorithm. Technometrics 22:325–331

Bilmes JA (1998) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. http://lasa.epfl.ch/teaching/lectures/ML_Phd/Notes/GP-GMM.pdf

Chang X-W, Guo Y (2005) Huber's M-estimation in relative GPS positioning: computational aspects. J Geodyn 79:351–362

DasGupta A (2011) Probability for statistics and machine learning. Springer, New York

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc B 39:1–38

Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101

Huber PJ (1981) Robust statistics. Wiley, New York

Khodabandeh A, Amiri-Simkooei AR, Sharifi MA (2012) GPS position time-series analysis based on asymptotic normality of M-estimation. J Geodyn 86:15–33

Koch KR (1999) Parameter estimationand hypothesis testing in linear models, 2nd edn. Springer, Berlin

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

Koch KR (2007b) Outlier detection in observations including leverage points by Monte Carlo simulations. Allgemeine Vermessungs-Nachrichten 114:330–336

Koch KR, Yang Y (1998) Robust Kalman filter for rank deficient observation models. J Geodyn 72:436–441

Koch KR, Kuhlmann H, Schuh W-D (2010) Approximating covariance matrices estimated in multivariate models by estimated auto- and cross-covariances. J Geodyn 84:383–397

Koch KR, Brockmann JM, Schuh W-D (2012) Optimal regularization for geopotential model GOCO02S by Monte Carlo methods and multi-scale representation of density anomalies. J Geodyn. doi:10.1007/s00190-012-0546-7

Krarup T, Juhl J, Kubik K (2012) Götterdämmerung over least squares adjustment. 14th Congress ISP Hamburg, International Archives of Photogrammetry, XXIII, B3, Commission III, pp 369–378

Lange K, Sinsheimer JS (1993) Normal/independent distributions and their applications in robust regression. J Comput Graphical Stat 2:175–198

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

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, Hobokon

Luxen M, Brunn A (2003) Parameterschätzung aus unvollständigen Beobachtungsdaten mittels des EM-Algorithmus. ZfV–Z Geodäsie, Geoinformation und Landmanagement 128:71–79

Peng J (2009) Jointly robust estimation of unknown parameters and variance components based on Expectation-Maximization algorithm. J Surv Eng 135:1–9

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 (1984) median of squares regression. J Am Stat Ass 79:871–880

Tukey JW (1960) A survey of sampling from contaminated distributions. In: Olkin I, Ghurye S, Hoeffding W, Madow W, Mann H (eds) Contributions to probability and statistics. Stanford University Press, Stanford pp 448–485

Wagner W (2010) Radiometric calibration of small-footprint full-waveform airborne laser scanner measurements: Basic physical concepts. ISPRS J Photogramm Remote Sens 65:505–513

Wang J, Feng Y, Wang C (2010) A modified inverse integer Cholesky decorrelation method and performance on ambiguity resolution. J Global Position Syst 9:156–165

Wu CFJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11:95–103

Yohai VJ (1987) High breakdown-point and high efficiency robust estimates for regression. Ann Stat 15:642–656