Predicting Missing Observations in Linear Models

Karl-Rudolf Koch

Summary
Observations, which are missing in a set of data to be analyzed in a linear model, can be replaced by fictitious values. A least squares adjustment is then applied to predict the missing data. The adjustments are repeated until the residuals of the predictions become negligibly small. This method goes back to Healy and Westmacott (1956), and it is derived here by the EM (expectation maximization) algorithm. It is applied to the measurements of a laser scanner which gave wrong results for data with high intensities of the reflected laser beam. The faulty observations were introduced as missing observations and then predicted. If the linear model is capable of predicting the missing data well, only approximations are needed for the fictitious values.

Zusammenfassung
Beobachtungen, die in einer Menge von Daten fehlen, aber benötigt werden, um sie in einem linearen Modell zu analysieren, können durch fiktive Werte ersetzt werden. Eine Ausgleichung nach der Methode der kleinsten Quadrate wird dann angewendet, um die fehlenden Beobachtungen zu prädizieren. Die Ausgleichungen werden wiederholt, bis die Residuen der Prädiktionen vernachlässigbar klein werden. Diese Methode geht auf Healy and Westmacott (1956) zurück und wird hier durch den EM (Erwartungswert-Maximierungs)-Algorithmus abgeleitet. Er wird auf die Messungen eines Laserscanners angewendet, die für Daten mit hohen Intensitäten des reflektierten Laserstrahls falsche Ergebnisse lieferten. Die fehlerhaften Daten wurden als fehlende Beobachtungen eingeführt und dann prädiziert. Falls das lineare Modell fähig ist, die fehlenden Daten gut zu prädizieren, genügen Näherungen für die fiktiven Werte.

Keywords: Missing observations, linear model, expectation maximization algorithm, prediction, laser scanner

1 Introduction

When instruments measure automatically, observations might be missing due to a failure of the electronics. Moving objects might obstruct the lines of sight of the instruments which result in missing or wrong observations. Missing data might also be a consequence of laser scanning as laser scanners work best for Lambertian surfaces which diffusely reflect the laser beams, cf. Wagner (2010). If the laser hits a strongly reflecting surface like metal, the intensity of the reflection might get so high that the instrument does not register an observation or the observation is erroneous. If some experiments for the analysis of variance, for instance, in a two-way classification, cf. Koch (1999, p. 202), do not produce results, the problem of missing data is also encountered.

If the observations together with the missing ones are needed to be analyzed in a linear model by a computer, already Healy and Westmacott (1956) suggested to replace the missing data by fictitious values and then estimate the unknown parameters of the model by a least squares adjustment. The missing observations are replaced by the values predicted by the estimated parameters. The adjustments are then repeated until the residuals for the predictions become negligibly small. This is not only a convincing procedure, it can also be theoretically derived by the EM (expectation maximization) algorithm.

Dempster et al. (1977) developed the EM algorithm in its full generality after special cases had been derived before. In order to approximate the maximum likelihood estimation, the EM algorithm is applied when observations are missing. They may be physically not available as pointed out above, or the missing data can be imaginary. For the latter case, the missing observations are introduced to facilitate the maximum likelihood estimation.

The EM algorithm has been used to derive robust estimations for linear models. Lange et al. (1989) started from the t-distribution which concentrates for small degrees of freedom more probability mass at the tails of its density function than the normal distribution. Outliers can therefore be taken care of. Unknown weights for the observations were introduced as missing data with small weights for the outliers so that the variance-inflation model was used, cf. Beckman and Cook (1983). The degree of freedom for the t-distribution was considered as unknown parameter to obtain an adaptive robust estimation. The estimates follow by iterative reweighted least squares.

Aitkin and Wilson (1980) applied a mixture of two normally distributed components: the first one for the observations with the expected values defined by the linear model, the second one for an outlier with its own expectation. Thus, a heavy-tailed density was obtained based on the mean-shift model, cf. Beckman and Cook (1983). Missing data furnish the information as to which observation belongs to which component. The density functions for the missing data give the weights which are small for outliers. The estimates are found by iterative reweighted least squares.

The L_1-norm estimate, i.e. the least absolute deviations estimate, of the parameters of a linear model is robust, cf. Koch (1999, p. 262). It can be obtained by iterative reweighted least squares as was shown by Schlossmacher (1973). Phillips (2002) derived this estimate by the EM algorithm. He also presented the EM algorithm for the L_1-norm estimate of the parameters of a nonlinear model.

 DOI 10.12902/zfv-0010-2014

In geodesy, the EM algorithm seems first to have been applied by Luxen and Brunn (2003). They solved a problem of classification where missing data supply the information as to which observations belong to which class. Peng (2009) introduced the linear model with unknown variance components and used the EM algorithm for a robust estimation based on a combination of the mean-shift and variance-inflation model. Koch (2013a) generalized the method of Aitkin and Wilson (1980) by introducing a mixture of any number of normally distributed components, the first one for the observations, each of the following ones for a suspected outlier. This generalized method showed in a Monte Carlo study superior performance in comparison to the robust M-estimation by Huber (1964) and to outlier tests (Koch 2013b). Kargoll and Krasbutter (2013) generalized the method of Lange et al. (1989) by introducing observational errors correlated by an autoregressive process. Finally, Koch and Kargoll (2013) suggested to apply the EM algorithm based on the variance-inflation model (Lange et al. 1989) first to identify possible outliers and then the mean-shift model (Koch 2013a) to get the outliers confirmed.

As mentioned above, the method of Healy and Westmacott (1956) for predicting missing observations can be derived by the EM algorithm. This was pointed out by Little and Rubin (2002, p. 237) as well as by McLachlan and Krishnan (2008, p. 49) who explained that for applying the EM algorithm the conditional expectations of the first two moments of the missing observations are needed. This is not obvious so that the EM algorithm is derived here. The paper is therefore organized as follows: Section 2 presents the derivation. Section 3 gives an example using the observations of a laser scanner. The conclusions follow in Section 4.

2 EM algorithm for predicting missing observations

Let the $r \times 1$ vector $\boldsymbol{y}_o = |y_1, \ldots, y_r|'$ contain the observations, the $(n-r) \times 1$ vector $\boldsymbol{y}_m = |y_{r+1}, \ldots, y_n|'$ the missing data and the $n \times 1$ vector $\boldsymbol{y} = |\boldsymbol{y}_o', \boldsymbol{y}_m'|'$ the complete data. The linear model for the complete data is then given by

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} + \boldsymbol{e} \quad \text{with} \quad \boldsymbol{X} = |\boldsymbol{X}_o', \boldsymbol{X}_m'|', \quad E(\boldsymbol{e}) = \boldsymbol{0},\ D(\boldsymbol{y}) = \sigma^2 \boldsymbol{I} \tag{1}$$

where $\boldsymbol{X}$ denotes the $n \times u$ matrix of coefficients with full column rank, $\boldsymbol{\beta}$ the $u \times 1$ vector of unknown parameters, $\boldsymbol{e}$ the $n \times 1$ vector of errors, $\boldsymbol{X}_o$ the $r \times u$ coefficient matrix belonging to $\boldsymbol{y}_o$, $\boldsymbol{X}_m$ the $(n-r) \times u$ coefficient matrix belonging to $\boldsymbol{y}_m$ and σ^2 the unknown variance factor. The special covariance matrix $D(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}$ and the normal distribution

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \tag{2}$$

are assumed for the complete data, which implies independent observations, cf. Koch (1999, p. 122). The marginal distributions for $\boldsymbol{y}_o$ and $\boldsymbol{y}_m$ follow with (2) by

$$\boldsymbol{y}_o \sim N(\boldsymbol{X}_o\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_r),\ \boldsymbol{y}_m \sim N(\boldsymbol{X}_m\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{n-r}). \tag{3}$$

Let $\boldsymbol{\Theta}$ be the vector of the unknown parameters of the linear model, i.e.

$$\boldsymbol{\Theta} = |\boldsymbol{\beta}', \sigma^2|', \tag{4}$$

and $p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\Theta})$ the likelihood function which is regarded as a function of $\boldsymbol{\Theta}$ given $\boldsymbol{y}_o$ and $\boldsymbol{y}_m$. To apply the maximum likelihood estimation, it is simpler to maximize the natural logarithm of the likelihood function, thus $\log p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\Theta})$. The EM algorithm distinguishes between the E (expectation) step and the M (maximization) step. The E step determines the conditional expectation, generally called $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)})$, of the log-likelihood function with respect to the conditional distribution for $\boldsymbol{y}_m$ given $\boldsymbol{y}_o$ and the current estimate $\boldsymbol{\Theta}^{(t)}$ of the unknown parameters, cf. McLachlan and Krishnan (2008, p. 18),

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) &= E(\log p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\Theta})|\boldsymbol{y}_o, \boldsymbol{\Theta}^{(t)}) \\ &= \int_{\mathcal{Y}_m} \log p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\Theta}) p(\boldsymbol{y}_m|\boldsymbol{y}_o, \boldsymbol{\Theta}^{(t)}) d\boldsymbol{y}_m \end{aligned} \tag{5}$$

where $\mathcal{Y}_m$ denotes the domain of $\boldsymbol{y}_m$ and $p(\boldsymbol{y}_m|\boldsymbol{y}_o, \boldsymbol{\Theta}^{(t)})$ the conditional density for $\boldsymbol{y}_m$ given $\boldsymbol{y}_o$ and $\boldsymbol{\Theta}^{(t)}$. Thus, the basic idea of the EM algorithm is to integrate out the missing data $\boldsymbol{y}_m$ and to replace it by the conditional expectation of $\boldsymbol{y}_m$. This will be shown in the following. The M step of the EM algorithm determines the new estimate $\boldsymbol{\Theta}^{(t+1)}$ by maximizing (5)

$$\boldsymbol{\Theta}^{(t+1)} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}). \tag{6}$$

The E and M steps are iteratively applied until (5) converges (Wu 1983).

The likelihood function $p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\Theta})$ follows with (2) and (4) by

$$\begin{aligned} p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\Theta}) &= (2\pi\sigma^2)^{-\frac{n}{2}} \\ &\exp[-\frac{1}{2\sigma^2}\left(\begin{vmatrix} \boldsymbol{y}_o \\ \boldsymbol{y}_m \end{vmatrix} - \begin{vmatrix} \boldsymbol{X}_o \\ \boldsymbol{X}_m \end{vmatrix}\boldsymbol{\beta}\right)'\left(\begin{vmatrix} \boldsymbol{y}_o \\ \boldsymbol{y}_m \end{vmatrix} - \begin{vmatrix} \boldsymbol{X}_o \\ \boldsymbol{X}_m \end{vmatrix}\boldsymbol{\beta}\right)] \\ = (2\pi\sigma^2)^{-\frac{n}{2}} &\exp[-\frac{1}{2\sigma^2}(\boldsymbol{y}_o'\boldsymbol{y}_o + \boldsymbol{y}_m'\boldsymbol{y}_m - 2\boldsymbol{y}_o'\boldsymbol{X}_o\boldsymbol{\beta} \\ &-2\boldsymbol{y}_m'\boldsymbol{X}_m\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_o'\boldsymbol{X}_o\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta})] \end{aligned} \tag{7}$$

and $\log p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\Theta})$ by

$$\begin{aligned} \log p(\boldsymbol{y}_o, \boldsymbol{y}_m|\boldsymbol{\Theta}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 \\ &-\frac{1}{2\sigma^2}(\boldsymbol{y}_o'\boldsymbol{y}_o - 2\boldsymbol{y}_o'\boldsymbol{X}_o\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_o'\boldsymbol{X}_o\boldsymbol{\beta}) \\ &-\frac{1}{2\sigma^2}(\boldsymbol{y}_m'\boldsymbol{y}_m - 2\boldsymbol{y}_m'\boldsymbol{X}_m\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}). \end{aligned} \tag{8}$$

The conditional distribution for $\boldsymbol{y}_m$ given $\boldsymbol{y}_o$ is obtained with (2) and (3) by, cf. Koch (1999, p. 121),

$$\boldsymbol{y}_m|\boldsymbol{y}_o \sim N(\boldsymbol{X}_m\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{n-r}) \tag{9}$$

and $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)})$ in (5) with (8) by

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) &= [-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 \\ &\quad -\frac{1}{2\sigma^2}(\boldsymbol{y}_o'\boldsymbol{y}_o - 2\boldsymbol{y}_o'\boldsymbol{X}_o\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_o'\boldsymbol{X}_o\boldsymbol{\beta})] \\ &\quad \int_{\mathcal{Y}_m} p(\boldsymbol{y}_m|\boldsymbol{y}_o, \boldsymbol{\Theta}^{(t)})d\boldsymbol{y}_m \\ &\quad -\frac{1}{2\sigma^2}\int_{\mathcal{Y}_m}(\boldsymbol{y}_m'\boldsymbol{y}_m - 2\boldsymbol{y}_m'\boldsymbol{X}_m\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}) \\ &\quad p(\boldsymbol{y}_m|\boldsymbol{y}_o, \boldsymbol{\Theta}^{(t)})d\boldsymbol{y}_m \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}_o'\boldsymbol{y}_o \\ &\quad -2\boldsymbol{y}_o'\boldsymbol{X}_o\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_o'\boldsymbol{X}_o\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}) \\ &\quad +\frac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}_m'\int_{\mathcal{Y}_m}\boldsymbol{y}_m p(\boldsymbol{y}_m|\boldsymbol{y}_o, \boldsymbol{\Theta}^{(t)})d\boldsymbol{y}_m \\ &\quad -\frac{1}{2\sigma^2}\int_{\mathcal{Y}_m}\boldsymbol{y}_m'\boldsymbol{y}_m p(\boldsymbol{y}_m|\boldsymbol{y}_o, \boldsymbol{\Theta}^{(t)})d\boldsymbol{y}_m. \end{aligned} \tag{10}$$

The last two integrals in (10) represent the conditional expectations of the first and second moments of the missing observations $\boldsymbol{y}_m$, which were already mentioned in Section 1.

We obtain with (9), with $\boldsymbol{y}_m = (y_i)$, $\boldsymbol{X}_m = (\boldsymbol{x}_i')$ for $i \in \{r+1, \ldots, n\}$ and with

$$p(\boldsymbol{y}_m|\boldsymbol{y}_o) = \prod_{i=r+1}^{n} p(y_i|\boldsymbol{y}_o) \tag{11}$$

the conditional expectation of the first moment of the component y_i of $\boldsymbol{y}_m$ by

$$\begin{aligned} E(y_i|\boldsymbol{y}_o) &= \int_{\mathcal{Y}_{r+1}}\ldots\int_{\mathcal{Y}_n} y_i p(y_{r+1}|\boldsymbol{y}_o)\ldots p(y_n|\boldsymbol{y}_o) \\ &\quad dy_{r+1}\ldots dy_n = \int_{\mathcal{Y}_i} y_i p(y_i|\boldsymbol{y}_o)dy_i = \boldsymbol{x}_i'\boldsymbol{\beta} \end{aligned} \tag{12}$$

and of the second moment, cf. Koch (1999, p. 119),

$$\begin{aligned} E(\boldsymbol{y}_m'\boldsymbol{y}_m|\boldsymbol{y}_o) &= \int_{\mathcal{Y}_{r+1}}\ldots\int_{\mathcal{Y}_n}(y_{r+1}^2 + \ldots + y_n^2) \\ &\quad p(y_{r+1}|\boldsymbol{y}_o)\ldots p(y_n|\boldsymbol{y}_o)dy_{r+1}\ldots dy_n \\ &= \sum_{i=r+1}^{n}\int_{\mathcal{Y}_i} y_i^2 p(y_i|\boldsymbol{y}_o)dy_i = \sum_{i=r+1}^{n}[(\boldsymbol{x}_i'\boldsymbol{\beta})^2 + \sigma^2] \\ &= \boldsymbol{\beta}'\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta} + (n-r)\sigma^2 \end{aligned} \tag{13}$$

where $\mathcal{Y}_{r+1}$ to $\mathcal{Y}_n$ denote the domains of y_{r+1} to y_n. By substituting (12) and (13) in (10) we get

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}_o'\boldsymbol{y}_o \\ &\quad -2\boldsymbol{y}_o'\boldsymbol{X}_o\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_o'\boldsymbol{X}_o\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}) \\ &\quad +\frac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}^{(t)} - \frac{1}{2\sigma^2}(\boldsymbol{\beta}^{(t)\prime}\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}^{(t)} \\ &\quad +(n-r)\sigma^{2(t)}). \end{aligned} \tag{14}$$

The M step according to (6) leads to

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\beta}}Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) &= \frac{1}{\sigma^2}(\boldsymbol{X}_o'\boldsymbol{y}_o - \boldsymbol{X}_o'\boldsymbol{X}_o\boldsymbol{\beta} \\ &\quad -\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta} + \boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}^{(t)}) = \boldsymbol{0} \end{aligned} \tag{15}$$

and to the normal equations for the estimate $\boldsymbol{\beta}^{(t+1)}$

$$(\boldsymbol{X}_o'\boldsymbol{X}_o + \boldsymbol{X}_m'\boldsymbol{X}_m)\boldsymbol{\beta}^{(t+1)} = \boldsymbol{X}_o'\boldsymbol{y}_o + \boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}^{(t)}. \tag{16}$$

The missing observations $\boldsymbol{y}_m$ are therefore replaced by their predictions

$$\boldsymbol{y}_m^{(t)} = \boldsymbol{X}_m\boldsymbol{\beta}^{(t)}. \tag{17}$$

Furthermore, we find

$$\begin{aligned} \frac{\partial}{\partial\sigma^2}Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{y}_o'\boldsymbol{y}_o \\ &\quad -2\boldsymbol{y}_o'\boldsymbol{X}_o\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_o'\boldsymbol{X}_o\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta} \\ &\quad +\boldsymbol{\beta}^{(t)\prime}\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}^{(t)} + (n-r)\sigma^{2(t)}) \\ &\quad -\frac{1}{(\sigma^2)^2}\boldsymbol{\beta}'\boldsymbol{X}_m'\boldsymbol{X}_m\boldsymbol{\beta}^{(t)} = \boldsymbol{0} \end{aligned} \tag{18}$$

and the estimate of $\sigma^{2(t+1)}$ by

$$\begin{aligned} \sigma^{2(t+1)} &= \frac{1}{n}[(\boldsymbol{y}_o - \boldsymbol{X}_o\boldsymbol{\beta}^{(t)})'(\boldsymbol{y}_o - \boldsymbol{X}_o\boldsymbol{\beta}^{(t)}) \\ &\quad +(n-r)\sigma^{2(t)}]. \end{aligned} \tag{19}$$

At the point of convergence of the EM algorithm, it holds $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$ and $\sigma^{2(t+1)} = \sigma^{2(t)}$ so that the estimate $\sigma^{2(t+1)}$ follows with

$$\sigma^{2(t+1)} = \frac{1}{r}(\boldsymbol{y}_o - \boldsymbol{X}_o\boldsymbol{\beta}^{(t+1)})'(\boldsymbol{y}_o - \boldsymbol{X}_o\boldsymbol{\beta}^{(t+1)}). \tag{20}$$

Looking at the observation equations for the missing observations $\boldsymbol{y}_m$, we get

$$\boldsymbol{X}_m\boldsymbol{\beta}^{(t+1)} = \boldsymbol{y}_m^{(t)} + \boldsymbol{e}_m^{(t)}. \tag{21}$$

Thus, we obtain with (17) at the point of convergence

$$\boldsymbol{e}_m^{(t+1)} = \boldsymbol{0}. \tag{22}$$

At convergence, $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t+1)})$ follows from (14) after neglecting the first term, which is constant, and from (20) by

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t+1)}) &= -\frac{n}{2}\log\sigma^{2(t+1)} - \frac{1}{2\sigma^{2(t+1)}} \\ &\quad (\boldsymbol{y}_o - \boldsymbol{X}_o\boldsymbol{\beta}^{(t+1)})'(\boldsymbol{y}_o - \boldsymbol{X}_o\boldsymbol{\beta}^{(t+1)}) - \frac{(n-r)}{2} \\ &= -\frac{n}{2}(\log\sigma^{2(t+1)} + 1). \end{aligned} \tag{23}$$

The EM algorithm for determining missing observations therefore starts with choosing fictitious values. The estimates (16) of the unknown parameters $\boldsymbol{\beta}$ are then iteratively applied by replacing the missing observations by their predictions $\boldsymbol{X}_m\boldsymbol{\beta}^{(t)}$ from (17). The iterations are stopped either if the residuals in (22) are negligibly small or the conditional expectation (23) has reached a maximum. The estimate of the variance factor then follows from (20).

If a nonlinear model is given and the unknown parameters are only approximately known, one has to iterate for the linearization. With each iteration, the missing data should be predicted to avoid that the linearization is distorted if the prediction is applied after the linearization.

Dependent observations can be decorrelated by the Cholesky factorization of the covariance matrix, cf. Koch (1999, p. 154). However, this will spread the fictitious values for the missing data over the observed data due to the correlations. If nevertheless a decorrelation is needed, it has to be computed together with the prediction of the missing data for each iteration of the linearization.

3 Numerical example

The coordinates x_i, y_i, z_i with $i \in \{1, \ldots, n\}$ and the intensities of the reflected laser beam of a grid of $n = 5 \times 5$ points on a plane, vertically standing metal sheet were measured by the laser scanner HDS 3000. The coordinates refer to the local coordinate system of the instrument with the x-axis lying horizontally, the y-axis coinciding with the center of the lines of sight and the z-axis pointing to the zenith. The scans start at the lower left corner of the grid from negative to positive z-values with increasing x-values and end at the upper right corner. The distances between the points on the metal sheet are about 11 cm, the shortest y-coordinate is 539 cm, and the intensities vary between 179 and 765. Some points on the metal sheet are hit almost perpendicularly by the laser beam.

To determine the standard deviations of the observed coordinates, the grid of points was measured in addition with $n_w = 25$ repetitions. This takes only a short time so that time variable systematic effects can be excluded (Koch 2010). With the mean $\bar{x}_i$ of the repeatedly measured coordinates x_{ij} for $i \in \{1, \ldots, n\}, j \in \{1, \ldots, n_w\}$ from

$$\bar{x}_i = \frac{1}{n_w} \sum_{j=1}^{n_w} x_{ij}, \tag{24}$$

we obtain the variance $\sigma_{x_i}^2$ of x_i by

$$\sigma_{x_i}^2 = \frac{1}{n_w - 1} \sum_{j=1}^{n_w} (x_{ij} - \bar{x}_i)^2 \tag{25}$$

and accordingly $\sigma_{y_i}^2$ and $\sigma_{z_i}^2$ and the standard deviations σ_{x_i}, σ_{y_i}, σ_{z_i} by taking the square roots. The mean values $\bar{\sigma}_{x_i}$, $\bar{\sigma}_{y_i}$ and $\bar{\sigma}_{z_i}$ of the 25 standard deviations result with

$$\bar{\sigma}_{x_i} = 0.05\,\text{cm},\ \bar{\sigma}_{y_i} = 0.21\,\text{cm},\ \bar{\sigma}_{z_i} = 0.06\,\text{cm}\,. \tag{26}$$

The standard deviations σ_{x_i} and σ_{z_i} are much smaller than σ_{y_i} so that x_i and z_i can be considered fixed when fitting a plane to the measurements. The observation equations are

$$\beta_0 + x_i\beta_1 + z_i\beta_2 = y_i + e_i \quad \text{for} \quad i \in \{1, \ldots, n\} \tag{27}$$

Tab. 1: Residuals $\hat{e}_i$ ordered by decreasing absolute values, standard deviations σ_{y_i} and intensities

Point	$\hat{e}_i$ (cm)	σ_{y_i} (cm)	Intensities
13	2.37	0.22	757
18	1.56	0.21	763
3	−0.67	0.27	613
9	−0.44	0.22	408
5	−0.37	0.21	276

where $\beta_0, \beta_1, \beta_2$ are the unknown parameters of the plane and e_i the errors of y_i. As mentioned, the observations y_i are assumed as independent but have different variances. By multiplying each observation equation (27) by $1/\sqrt{\sigma_{y_i}^2}$, the linear model (1) is obtained, cf. Koch (1999, p. 155). The unknown parameters of the plane are estimated by a least squares adjustment. The results for the residuals $\hat{e}_i$ with maximum absolute errors ordered by decreasing values are shown in Tab. 1 for five points. The numbers of the points follow the sequence of the scans. The standard deviations σ_{y_i} from (25) and the intensities of the reflected laser beam are also given in Tab. 1.

Looking at the residuals $\hat{e}_i$ and the standard deviations σ_{y_i} of Tab. 1 for points 13 and 18 with high intensities, it becomes obvious that the measured coordinates y_i are wrong. The residual for point 3 which also has a high intensity is acceptable, it lies within the interval $3\sigma_{y_i}$. The measurements y_i for points 13 and 18 were therefore assumed as missing, and the EM algorithm was applied to iteratively predict the missing observations according to (16) and (17).

To check how the computations depend on the fictitious values for the observations, three results have been computed and shown in Tab. 2. The faulty measurements were used as fictitious values for result 1. The mean of the measurements y_i of points 12 and 14 for point 13 and of points 17 and 19 for point 18 were introduced as fictitious values for result 2. Finally, 100 cm were added to the mean values to obtain result 3. It was iterated, until the absolute values of the residuals in (22) were smaller than 0.004 cm. Tab. 2 shows the predicted observations $\boldsymbol{X}_m\boldsymbol{\beta}^{(t)}$ for point 13 and 18 from (17), the square root of the unbiased estimate of the variance factor from (20) and the number of iterations. The three results for the

Tab. 2: Predicted observations for points 13 and 18, square root of estimated variance factor and number of iterations

Result	point 13 predict. obs. (cm)	point 18 predict. obs. (cm)	sqrt. est. var. fact.	Iterations
1	541.16	540.95	1.1017	3
2	541.16	540.95	1.1017	3
3	541.16	540.95	1.1017	5

prediction of point 13 and point 18 agree. Approximate fictitious values can therefore be chosen for the missing data.

The transformation into the model (1) has been accomplished by the estimates (25) of the variances $\sigma_{y_i}^2$. If no model errors exist, the estimated variance factor should be close to one. The results of 1.1017 in Tab. 2 indicate that the surface of the metal sheet can be represented by a plane and that no outliers exist in the measured and predicted observations.

The sum of the adjusted distances from the instrument to the grid of points is sensitive to errors in the adjusted y-coordinates of the points. The expected value and the confidence interval for the sum were therefore determined from Monte Carlo methods, cf. Koch (2010). 100 000 normally distributed random variates for the y-coordinates with the variances $\sigma_{y_i}^2$ from (25) were generated., cf. Koch (2007, p. 197). The expectation of 13539.37 cm for the sum, the lower limit of the confidence interval of 13537.25 cm and the upper limit of 13541.49 cm were obtained for all three results of Tab. 2. This confirms that approximate values can be chosen as fictitious ones for the missing observations.

4 Conclusions

The method of Healy and Westmacott (1956) for predicting missing observations to be evaluated in a linear model is a convincing procedure, especially because it can be derived by the EM algorithm as was shown here. The measurements of a laser scanner, which were analyzed here, gave wrong results in case of high intensities of the reflected laser beam. The faulty measurements were introduced as missing observations into the EM algorithm. The linear model was defined by a plane, which was fitted to the y-coordinates measured by the laser scanner. It predicted the missing observations very well. This was found out by assuming different fictitious values for the missing data and by applying a Monte Carlo method to compute the expectation and the confidence interval of the sum of the adjusted distances from the instrument to the observed points on the plane. It is therefore sufficient to determine the fictitious values for the missing observations only approximately.

Some types of laser scanners do not register any measurements if the intensity of the reflected laser beam surpasses a certain limit. For such cases, the x- and z-coordinates have to be interpolated into the measured coordinates of the grid of points. Fictitious values are then chosen for the y-coordinates which are introduced as missing data for the EM algorithm.

Acknowledgement

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

References

Aitkin, M., Wilson, G.T.: Mixture models, outliers, and the EM algorithm. Technometrics, 22:325–331, 1980.

Beckman, R.J., Cook, R.D.: Outlier....s. Technometrics, 25:119–149, 1983.

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J Royal Statist Soc, B, 39:1–38, 1977.

Healy, M., Westmacott, M.: Missing values in experiments analysed on automatic computers. J Royal Statist Soc, C, 5:203–206, 1956.

Huber, P.J.: Robust estimation of a location parameter. Annals Mathematical Statistics, 35:73–101, 1964.

Kargoll, B., Krasbutter, I.: An iteratively reweighted least squares approach to adaptive robust adjustment of parameters in linear regression models with autoregressive, t-distributed observation errors. J Geodesy, submitted, 2013.

Koch, K.R.: Parameter Estimation and Hypothesis Testing in Linear Models, 2nd Ed. Springer, Berlin, 1999.

Koch, K.R.: Introduction to Bayesian Statistics, 2nd Ed. Springer, Berlin, 2007.

Koch, K.R.: Uncertainty of results of laser scanning data with correlated systematic effects by Monte Carlo methods. ZfV – Z Geodäsie, Geoinformation und Landmanagement, 135:376–385, 2010.

Koch, K.R.: Robust estimation by expectation maximization algorithm. J Geodesy, 87:107–116, 2013a.

Koch, K.R.: Comparison of two robust estimations by expectation maximization algorithms with Huber's method and outlier tests. J Applied Geodesy, 7:115–123, 2013b.

Koch, K.R., Kargoll, B.: Expectation maximization algorithm for the variance-inflation model by applying the t-distribution. J Applied Geodesy, 7:217–225, 2013.

Lange, K.L., Little, R.J.A., Taylor, J.M.G. : Robust statistical modeling using the t distribution. J Am Statist Ass, 84(408):881–896, 1989.

Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd Ed. Wiley, Hobokon, New Jersey, 2002.

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

McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd Ed. Wiley, Hoboken, New Jersey, 2008.

Peng, J.: Jointly robust estimation of unknown parameters and variance components based on expectation-maximization algorithm. J Surveying Engineering, 135:1–9, 2009.

Phillips, R.F.: Least absolute deviations estimation via the EM algorithm. Statistics and Computing, 12:281–285, 2002.

Schlossmacher, E.J.: An iterative technique for absolute deviations curve fitting. J Am Statist Ass, 68:857–859, 1973.

Wagner, W.: Radiometric calibration of small-footprint full-waveform airborne laser scanner measurements: Basic physical concepts. ISPRS J Photogrammetry and Remote Sensing, 65:505–513, 2010.

Wu, C.F.J.: On the convergence properties of the EM algorithm. Ann Statist, 11:95–103, 1983.

Author's adress

Prof. Dr.-Ing., Dr.-Ing. E.h. mult. Karl-Rudolf Koch (em.)
Institute for Geodesy and Geoinformation, Theoretical Geodesy
University of Bonn
Nussallee 17, 53115 Bonn, Germany
koch@geod.uni-bonn.de

This article also is digitally available under www.geodaesie.info.