Bachelor Thesis
Estimating Standard Errors of Parameters Obtained by the EM-Algorithm

Marijtje van Leeuwen

Erasmus University Rotterdam

Abstract. In this thesis, two different methods of standard error estimation when using the EM-algorithm will be compared. One of these methods is a long existing and commonly used method called Louis' method. The main idea of this method is to use the first and second derivatives of the complete-data log likelihood to obtain an estimate for the observed information matrix. The second method estimates the observed information matrix by taking the negative of the second derivative matrix of the incomplete-data log likelihood and has to my knowledge not been proposed in earlier research. A simulation experiment shows that the standard error estimates by the second method are generally closer to the simulated standard error estimates when the EM-algorithm is used to estimate parameters of a Gaussian mixture model with two components.

Table of Contents

1 Introduction

The Expectation-Maximization (EM) algorithm for finding Maximum Likelihood Estimates (MLE's) is a commonly used approach in a variety of incomplete-data problems. In this context, incomplete-data problems can be interpreted in many ways, such as data with missing values, grouped, censored or truncated data, finite mixture models, et cetera. Incomplete-data problems arise in many contexts, like the medical context (when a true disease status is not observed, see Kang et al. (2013)) or in a marketing context (Market segmentation, see Kim and Lee (2011)). Although the general formulation of the EM-algorithm was presented in Dempster et al. (1977), there are still a number of papers published on the algorithm every year, inter alia Kang et al. (2013), Kim and Lee (2011), Melnykov and Melnykov (2012), Cappé (2011) and Xu et al. (2014), suggesting that the topic is still very relevant in the recent literature.

The main idea of the EM-algorithm is to associate with the given incomplete-data problem, a complete-data problem, for which Maximum Likelihood estimation is computationally more tractable, see McLachlan and Krishnan (2008). The algorithm performs two steps on each iteration. The first step, which is known as the E-step, is to compute a particular conditional expectation of the complete-data log likelihood. The second step, known as the M-step, is to maximize this expectation over the relevant parameters. The EM-algorithm has some appealing properties relative to other methods for finding MLE's, such as monotonicity and the fact that it is easy to implement the algorithm, see McLachlan and Krishnan (2008). However, there is also some criticism associated with the algorithm. One of the commonly mentioned disadvantages of the algorithm is that the algorithm does not provide estimates of the covariance matrix of the parameter estimates. Since in most statistical problems, it is convenient to have standard errors of parameter estimates, a solution is needed.

There are several methods designed to compute standard errors of parameter estimates obtained by the EM-algorithm. Some of these methods focus on a specific incomplete data problem (Baker's method for categorical data, see Baker (1992)), and others are more general (Louis' method, see Louis (1982), Oakes' method, see Oakes (1999)). Most of the methods focus on computing the inverse of the observed information matrix, which is used as an estimate of the covariance matrix. In some papers, the standard errors are estimated using the expected information matrix, which is generally easier to compute, see Baker (1992). However, in Baker (1992) it is also stated that using the empirical information matrix for standard error estimation is inefficient and ignores the likelihood principle. An alternative method to approximate standard errors is the bootstrap approach, which is used in Efron (1994).

Among all existing methods for standard error estimation when using the EM-algorithm, Louis' method, as presented in Louis (1982), seems to be the most popular and can be seen as the Golden Standard in the field of standard error estimation when using the EM-algorithm. In Louis' method, the observed information matrix is computed on the last iteration of the EM procedure using only the complete-data gradient and second derivative matrix. The observed information matrix needs to be inverted in order to obtain estimates of the covariance matrix. One of the main advantages of Louis' method is that it is generally applicable, see McLachlan and Krishnan (2008). In McLachlan and Krishnan (2008) it is even stated that Louis' method is the one best suited for standard error estimation of parameters obtained by using the Monte Carlo EM (MCEM) algorithm, which is described in Wei and Tanner (1990), although in Kang et al. (2013) it is stated that the performance of the method for standard error estimation in MCEM is not yet clear.

Since Louis' method for estimating standard errors is quite technical (in Kang et al. (2013) it is said that obtaining the gradient and second derivative matrix is not always easy), it could be useful to apply an approach that is less technical and more easy to compute. Such an alternative approach is only useful if it works fast, but still provides good estimates in the end. In the current research, an alternative approach for standard error estimation will be discussed. The main idea of the alternative approach is to use the negative of the second derivative matrix of the incomplete-data log likelihood as an estimate of the information matrix, which can be inverted in order to obtain an estimate of the covariance matrix. In this research, we will compare the performance of standard error estimates obtained by Louis' method with the standard error estimates obtained by the alternative method in the context of Gaussian mixture models.

The methods will be exemplified using a data set that contains the real Gross Domestic Products (GDP's) per capita of 143 countries over the period 1970-2011. The distribution of the GDP's across countries seems to contain heterogeneity, see Paap and van Dijk (2009). Therefore, one can a use a finite mixture model to divide the countries as if in different groups. In Paap and van Dijk (2009), it is argued that the GDP across countries can be best modelled by using a mixture of two distributions. That is, as if there are two groups, one group of countries with a lower GDP and one group of countries with a higher GDP. In this research, we will use a Gaussian mixture model (see Hasselblad (1966)) to describe the data.

In the remaining of the paper, the characteristics of the data set will first be discussed and it will be explained why we use a Gaussian mixture model to describe the data in Section 2. After that, we will shortly explain the EM-algorithm and Louis' algorithm in Section 3. In this section, it will also be explained how the alternative method for standard error estimation works. For all three algorithms, we will explain how it could be used for the data set as well. In Section 4, the results of the different algorithms on the used data set will be discussed. In Section 5, the results of a simulation experiment with the parameters set at the same values as the estimated parameters by the EM-algorithm on the real data set will be shown. In Section 6, the results of another simulation experiment will be discussed to obtain even more insights about the differences between Louis' method and the alternative method. At last, in Section 7, we will discuss the main findings of this research and give suggestions for further research.

2 Application: Modelling the distribution of wealth

In this section, the data set that is used to test the different standard error estimation methods is discussed. The idea is to model the distribution and mobility of wealth of different countries. First, it will be explained what the numbers in the data set represent. Then, the main characteristics and earlier research on the topic is shortly discussed.

2.1 The data set

As we want to model the distribution of wealth, an appropriate measure for wealth is needed. The measure should be chosen in such a way that it is possible to compare wealth of different countries. The Penn World Table version 8.0, see Feenstra et al. (2013), provides inter alia, the real Gross Domestic Product based on national accounts for 167 countries over the time period 1950-2011. The real GDP is constructed by dividing the GDP of the countries in national currencies by estimations of Purchasing Power Parities (PPP's). This is done to correct for differences in prices across countries. The real GDP that is constructed this way only needs to be divided by the population size of the country in order to obtain an estimate of real GDP per capita. When constructing real GDP per capita like this, it becomes a measure of relative living standards across countries at different points in time. It can thus be interpreted as a measure of wealth that allows to compare between countries. For more information about how PPP's are estimated and about the data set in general, it is advised to read Feenstra et al. (2013).

In this research, the data set provided in Feenstra et al. (2013) will be used to calculate the real GDP per capita for countries for a sequence of years. However, as not all observations for all 167 countries are known over the entire period, we do not use the data for the entire time period, but only use the data for the period 1970-2011. There are 143 countries left with information about real GDP and population size for this period of time. For all years, the EM-algorithm is used to find the distribution that best describes the data.

2.2 Theory

In Paap and van Dijk (2009), it is shown that the real GDP per capita, which can be seen as an approximation of wealth, can be described by a bimodal distribution. This means that one can use a finite mixture density in order to obtain an estimate of the distribution. By using finite mixture models, one treats the sample as if it is a composition of more data sets, each with their own distribution and density function. In our case, as we assume the data can be described by a bimodal distribution, we let there be two different groups. One of the groups will represent the poor countries group and in the other group, there will be the rich countries. Although in Paap and van Dijk (2009), a mixture of a Weibull and a truncated normal distribution is used to describe the data, it is also stated that the data can be fairly well described by a mixture of two truncated normal distributions. Because our main focus is on estimating standard errors, we will use a mixture of two normal density functions, as this is more easy to implement and can still provide useful insights in standard error estimation.

3 Methodology

In this section, the basics of the EM-algorithm will shortly be discussed in subsection 3.1. An example of the algorithm applied to the data set will be given as well. In subsection 3.2, Louis' method is explained in more detail. At last, the alternative method for standard error estimation will be provided in subsection 3.3. For both standard error estimation methods, the derivations for the algorithm applied to the used data set will be given as well. That is, it will be explained how the methods should be implemented for a Gaussian mixture model with two components.

3.1 The EM-algorithm

The EM-algorithm is an approach to the iterative computation of Maximum Likelihood estimates. It can be used in a variety of incomplete-data problems, such as missing data problems or finite mixture models. The EM-algorithm uses the complete-data log likelihood to estimate the parameters of the incomplete-data problem. The algorithm consists of two steps. In the first step, the conditional expectation of the log likelihood of the complete-data problem is calculated, as the actual log likelihood function is not observable. Since the first step thus requires the calculation of an expectation, it is called the expectation step of the algorithm, which is abbreviated to 'E-step'. The second step of the algorithm is to maximize the expectation with respect to the parameters that are to be estimated. This step is also called the maximization step of the algorithm, and is abbreviated to 'M-step'. The steps are repeated until convergence.

We let $\log L(\boldsymbol{\psi}) = \log g(\boldsymbol{y}, \boldsymbol{\psi})$ be the log likelihood function of the incomplete-data problem and $\log L_c(\boldsymbol{\psi}) = \log g_c(\boldsymbol{x}, \boldsymbol{\psi})$ the log likelihood function of the complete-data problem. In the first function, there is the variable $\boldsymbol{y}$, which represents the observed data vector. The variable $\boldsymbol{x}$ represents the complete-data vector and includes the variable $\boldsymbol{y}$ as well as some unknown or missing data vector $\boldsymbol{z}$, so that in general it holds that $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$. The idea of the EM-algorithm is to estimate $\boldsymbol{\psi}$ by using $\log L_c(\boldsymbol{\psi})$. Since $\boldsymbol{x}$ is not fully observed, the first step of the algorithm requires to compute the expectation of this log likelihood function conditioned on what data is observed, namely $\boldsymbol{y}$. In the first iteration of the algorithm, the E-step thus requires calculation of $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(0)}) = E_{\boldsymbol{\psi}^{(0)}}\{\log L_c(\boldsymbol{\psi})|\boldsymbol{y}\}$. In the equation, $\boldsymbol{\psi}^{(0)}$ represents the initial value for $\boldsymbol{\psi}$. The M-step requires the maximization of $\mathrm{Q}(\boldsymbol{\psi};\boldsymbol{\psi}^{(0)})$ with respect to $\boldsymbol{\psi}$. That is, $\boldsymbol{\psi}^{(1)}$ is chosen in such a way that $\mathrm{Q}(\boldsymbol{\psi}^{(1)};\boldsymbol{\psi}^{(0)}) \geqslant \mathrm{Q}(\boldsymbol{\psi};\boldsymbol{\psi}^{(0)})$. In the second iteration, the values of the current fit $\boldsymbol{\psi}^{(1)}$ are used to calculate $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(1)})$ instead of $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(0)})$ and in the M-step the values for $\boldsymbol{\psi}^{(2)}$ are chosen such that $\mathrm{Q}(\boldsymbol{\psi};\boldsymbol{\psi}^{(1)})$ is maximized with respect to $\boldsymbol{\psi}$. The E-step and M-step are repeated until the difference between subsequent (incomplete-data) likelihood values is sufficiently small (L($\boldsymbol{\psi}^{(k+1)}$)-L($\boldsymbol{\psi}^{(k)}$)$\leqslant \epsilon$).

The theory applied to the used data set Earlier research has shown that the data set we use may be described by a bimodal distribution. A finite mixture model can thus be used to model the data set, where the number of components is set at two. One of the groups may be interpreted as a group of countries with low GDP per capita (poor countries) and the other group may be interpreted as a group with high GDP per capita (rich countries). In a finite mixture model, the distribution of each group is known, but it is unknown to which group an observation belongs. In our case, we assume that the countries within the groups can be described by normal distributions with parameters $\boldsymbol{\theta}_1$=(μ_1,σ_1) and $\boldsymbol{\theta}_2$=(μ_2,σ_2). We let π_1 and π_2 be the ex-ante probabilities that a country belongs to the first or the second group. For identifiability reasons however, one should impose a constraint stating that the one is less or equal than the other, thus that $\pi_1 \leq \pi_2$. Furthermore, as we assume that there are only two groups, π_1 and π_2 should sum up to one. This means that we only need to estimate one of them, say π_1 which will be simply referred to as π. When only using this variable π, the constraint that needs to be imposed becomes $\pi \leq \frac{1}{2}$.

The incomplete-data log likelihood of a finite mixture with two components is:

$$\log L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \pi) = \sum_{i=1}^{n} \log\{\pi\phi(y_i; \boldsymbol{\theta}_1) + (1-\pi)\phi(y_i; \boldsymbol{\theta}_2)\} \tag{1}$$

where $\phi(y_i;\theta_j) = (2\pi\sigma_j^2)^{-1/2} exp\{-\frac{1}{2}(y_i - \mu_j)^2/\sigma_j^2\}$, the density function of the normal distribution with parameters μ_j and σ_j. The corresponding complete log likelihood is

$$\log L_c(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \pi) = \sum_{i=1}^{n}\{z_i \log(\pi) + (1-z_i)\log(1-\pi) + z_i \log(\phi(y_i; \boldsymbol{\theta}_1)) + (1-z_i)\log(\phi(y_i; \boldsymbol{\theta}_2))\} \tag{2}$$

In the equation, z_i equals 0 if observation i belongs to the first group and 1 if it belongs to the second group. Whether an observation belongs to the one or the other group is decided by the algorithm as well. The E-step of the algorithm is now to determine $Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(k)}) = E_{\boldsymbol{\psi}^{(k)}}\{logL_c(\boldsymbol{\psi})|\boldsymbol{y}\}$, with $\boldsymbol{\psi} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \pi)$. In the M-step, $Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(k)})$ is maximized with respect to $\boldsymbol{\psi}$ such that $Q(\boldsymbol{\psi}^{(k+1)}, \boldsymbol{\psi}^{(k)}) \geq Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(k)})$. The E-step and M-step are repeated until convergence.

Once the estimates of parameters are obtained, it is easy to determine which parameters belong to the poor group and which belong to the rich group. One can simply check whether μ_1 is larger or smaller than μ_2. When it is larger, the parameters of the distribution of the rich group are $\boldsymbol{\theta}_1$ and the probability that a country belongs to the rich group is estimated by π. For the poor group, the distribution has parameters $\boldsymbol{\theta}_2$ and the probability that a country belongs to the group is $(1-\pi)$. When μ_1 is smaller than μ_2, it is the other way around.

3.2 Louis' method for standard error estimation

In Louis (1982), Louis' method to extract the observed information matrix of the incomplete-data problem is described. For his method, one needs to compute only the gradient vector and second derivative matrix of the complete-data problem. Furthermore, the method only needs to be executed once, after the last iteration of the EM-algorithm. When $S_c(\boldsymbol{x};\boldsymbol{\psi})$ and $S(\boldsymbol{y};\boldsymbol{\psi})$ represent the gradient vector of the complete- and, respectively incomplete-data log likelihood and $B_c(\boldsymbol{x};\boldsymbol{\psi})$ and $B(\boldsymbol{y};\boldsymbol{\psi})$ are the second derivatives matrices of the complete- and incomplete-data log likelihood, the observed information matrix can be computed as:

$$I(\boldsymbol{\psi};\boldsymbol{y}) = E_{\boldsymbol{\psi}}\{-B_c(X;\boldsymbol{\psi})|\boldsymbol{y}\} - E_{\boldsymbol{\psi}}\{S_c(X;\boldsymbol{\psi})S_c^T(X;\boldsymbol{\psi})|\boldsymbol{y}\} - S(\boldsymbol{y};\boldsymbol{\psi})S^T(\boldsymbol{y};\boldsymbol{\psi}) \tag{3}$$

In (3), it can be noticed that the conditional expectations of $S_c(X;\boldsymbol{\psi})$ and $B_c(X;\boldsymbol{\psi})$ are used. This is necessary, since the variable $\boldsymbol{x}$ is not fully observed. The last term in (3), $S(\boldsymbol{y},\boldsymbol{\psi})$, represents the gradient vector of the incomplete-data log likelihood. However, the incomplete-data log likelihood is what is maximized with the EM-algorithm, so the gradient vector, $S(\boldsymbol{y},\boldsymbol{\psi})$, will be equal to the zero vector. For that reason, the term $S(\boldsymbol{y};\boldsymbol{\psi})S^T(\boldsymbol{y};\boldsymbol{\psi})$ can be discarded and the observed information matrix can be computed as:

$$I(\boldsymbol{\psi};\boldsymbol{y}) = E_{\boldsymbol{\psi}}\{-B_c(X;\boldsymbol{\psi})|\boldsymbol{y}\} - E_{\boldsymbol{\psi}}\{S_c(X;\boldsymbol{\psi})S_c^T(X;\boldsymbol{\psi})|\boldsymbol{y}\} \tag{4}$$

$I(\boldsymbol{\psi};\boldsymbol{y})$ can be inverted to obtain an estimate of the covariance matrix of the incomplete-data problem. The square roots of the diagonal elements represent the estimates of the standard errors of the parameters.

The theory applied to the used data set The complete-data log likelihood for the application is given in (2). In order to obtain standard error estimates using Louis' method, one needs to compute the gradient vector $S_c(\boldsymbol{x};\boldsymbol{\psi})$ and negatives of the second derivative matrix $-B_c(\boldsymbol{x};\boldsymbol{\psi})$ of this complete-data log likelihood first. The gradient vector of (2) is:

$$S_c(\boldsymbol{x};\boldsymbol{\psi}) = \begin{bmatrix} \frac{\partial \log L_c}{\partial \pi} \\ \frac{\partial \log L_c}{\partial \mu_1} \\ \frac{\partial \log L_c}{\partial \mu_2} \\ \frac{\partial \log L_c}{\partial \sigma_1} \\ \frac{\partial \log L_c}{\partial \sigma_2} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \{\frac{z_i}{\pi} - \frac{1-z_i}{1-\pi}\} \\ \sum_{i=1}^n \{(1-z_i)(\frac{x_i-\mu_1}{\sigma_1^2})\} \\ \sum_{i=1}^n \{z_i(\frac{x_i-\mu_2}{\sigma_2^2})\} \\ \sum_{i=1}^n \{(1-z_i)(-\frac{1}{\sigma_1} + \frac{(x_i-\mu_1)^2}{\sigma_1^3})\} \\ \sum_{i=1}^n \{z_i(-\frac{1}{\sigma_2} + \frac{(x_i-\mu_2)^2}{\sigma_2^3})\} \end{bmatrix} \tag{5}$$

The negative of the second derivative matrix of (2) is:

$$-B_c(\boldsymbol{x};\boldsymbol{\psi}) = \sum_{i=1}^{n}$$
$$\begin{bmatrix} \frac{z_i}{\pi^2} + \frac{1-z_i}{(1-\pi)^2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1-z_i}{\sigma_1^2} & 0 & \frac{2(1-z_i)(x_i-\mu_1)}{\sigma_1^3} & 0 \\ 0 & 0 & \frac{z_i}{\sigma_2^2} & 0 & \frac{2z_i(x_i-\mu_2)}{\sigma_2^3} \\ 0 & \frac{2(1-z_i)(x_i-\mu_1)}{\sigma_1^3} & 0 & \frac{(1-z_i)(-\sigma_1^2+3(x_i-\mu_1)^2)}{\sigma_1^4} & 0 \\ 0 & 0 & \frac{2z_i(x_i-\mu_2)}{\sigma_2^3} & 0 & \frac{z_i(-\sigma_2^2+3(x_i-\mu_2)^2)}{\sigma_2^4} \end{bmatrix} \tag{6}$$

The covariance matrix can be estimated by taking the inverse of $I(\boldsymbol{\psi};\boldsymbol{y}) = -B_c(\boldsymbol{x};\boldsymbol{\psi}) + S_c(\boldsymbol{x};\boldsymbol{\psi})S_c^T(\boldsymbol{x};\boldsymbol{\psi})$ evaluated at $\widehat{\boldsymbol{\psi}}$ from the last iteration of the EM-algorithm. The diagonal elements square roots of the covariance matrix can be used as an estimation or the standard errors of $\pi,\ \mu_1,\ \mu_2,\ \sigma_1$ and σ_2 respectively. Whether the standard error belongs to the parameter of the poor or rich group still depends on the relative size of the parameter estimates μ_1 and μ_2.

3.3 An alternative method for standard error estimation

As an alternative method for standard error estimation of the parameters of an incomplete-data problem, we will use the second derivative matrix of the incomplete-data log likelihood. Since we want to compute standard error estimates of the incomplete-data problem, it seems to make sense to focus on this particular log likelihood function. For normal Maximum Likelihood estimation, it is usual to use the inverse of the negative of the second derivative matrix of the log likelihood function evaluated at the parameter estimates as an estimation for the covariance matrix. We will investigate whether using this inverse of the negative of the second derivative of the log likelihood of the incomplete-data problem is a good approximation for the covariance matrix, or more specifically, for the standard errors of parameter estimates. That is, when $\log L(\boldsymbol{\psi}) = \log g(\boldsymbol{y};\boldsymbol{\psi})$ represents the log likelihood function of the incomplete-data problem, we take the second derivative (also called Hessian) $\frac{\partial^2 \log L(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}$ and the inverse of the negative expectation of the Hessian evaluated at the parameter estimates is used as an estimate of the covariance matrix. Thus, $Cov_{\boldsymbol{\psi}} = (-E_{\boldsymbol{\psi}}[\frac{\partial^2 \log L(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}]_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}})^{-1}$ is the covariance matrix estimate obtained when using this method.

The theory applied to the used data set For the alternative method for standard error estimation, one needs to compute the second derivative matrix of the incomplete-data log likelihood. For a Gaussian mixture model with two components, which is assumed to be the distribution for the used data set, the log likelihood function is given in (1). The second derivative matrix with respect to π, μ_1, μ_R, σ_1 and σ_2 is constructed as:

$$B(\boldsymbol{y};\boldsymbol{\psi}) = \begin{bmatrix} \frac{\partial^2 \log L}{\partial \pi^2} & \frac{\partial^2 \log L}{\partial \pi \partial \mu_1} & \frac{\partial^2 \log L}{\partial \pi \partial \mu_2} & \frac{\partial^2 \log L}{\partial \pi \partial \sigma_1} & \frac{\partial^2 \log L}{\partial \pi \partial \sigma_2} \\ \frac{\partial^2 \log L}{\partial \mu_1 \partial \pi} & \frac{\partial^2 \log L}{\partial \mu_1^2} & \frac{\partial^2 \log L}{\partial \mu_1 \partial \mu_2} & \frac{\partial^2 \log L}{\partial \mu_1 \partial \sigma_1} & \frac{\partial^2 \log L}{\partial \mu_1 \partial \sigma_2} \\ \frac{\partial^2 \log L}{\partial \mu_2 \partial \pi} & \frac{\partial^2 \log L}{\partial \mu_2 \partial \mu_1} & \frac{\partial^2 \log L}{\partial \mu_2^2} & \frac{\partial^2 \log L}{\partial \mu_2 \partial \sigma_1} & \frac{\partial^2 \log L}{\partial \mu_2 \partial \sigma_2} \\ \frac{\partial^2 \log L}{\partial \sigma_1 \partial \pi} & \frac{\partial^2 \log L}{\partial \sigma_1 \partial \mu_1} & \frac{\partial^2 \log L}{\partial \sigma_1 \partial \mu_2} & \frac{\partial^2 \log L}{\partial \sigma_1^2} & \frac{\partial^2 \log L}{\partial \sigma_1 \partial \sigma_2} \\ \frac{\partial^2 \log L}{\partial \sigma_2 \partial \pi} & \frac{\partial^2 \log L}{\partial \sigma_2 \partial \mu_1} & \frac{\partial^2 \log L}{\partial \sigma_2 \partial \mu_2} & \frac{\partial^2 \log L}{\partial \sigma_2 \partial \sigma_1} & \frac{\partial^2 \log L}{\partial \sigma_2^2} \end{bmatrix} \tag{7}$$

The actual derivatives for the Gaussian mixture model with two components can be found in Appendix A. In order to obtain an estimate for the covariance matrix of the parameter estimates, one first needs to estimate $B(\boldsymbol{y};\boldsymbol{\psi})$ by using the parameter estimates of the final iteration of the EM-algorithm. The inversion of the negative of $B(\boldsymbol{y};\boldsymbol{\psi})_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}$ gives an estimate of the covariance matrix when using this alternative method. The square roots of the diagonal elements can be used as standard error estimates of the parameters.

4 Results on the used data set

In this section, the results of the EM-algorithm and the different standard error estimation methods on the data set that contains the real GDP per capita of 143 countries over the time period 1970-2011 will be discussed. The main focus will be on the differences between Louis' method and the alternative method that are both designed to obtain standard error estimates of parameters estimated while using the EM-algorithm.

As mentioned before, once the parameters are estimated, it is easy to determine which parameters belong to the distribution of the poor group and which belong to the rich group. When $\mu_1 \leq \mu_2$, the parameters μ_1 and σ_1 belong to the poor group and the ex-ante probability that a country belongs to the poor group is π. In this case, the parameters μ_2 and σ_2 belong to the distribution of the rich group and the ex-ante probability that a country belongs to the rich group is $(1-\pi)$. When $\mu_2 \leq \mu_1$, it is exactly the other way around. In this section, the parameters that belong to the distribution of the poor countries will be referred to as μ_P and σ_P and parameters of the rich group will be μ_R and σ_R. The variable π represents the ex-ante probability that a country belongs to the poor group, independent of whether $\mu_P \leq \mu_R$ or $\mu_R \leq \mu_P$.

In Fig. 1, the results of the parameter and standard error estimates are plotted against time. The EM-algorithm is applied to all years that are in the data set and it can be seen that the parameter estimates (represented by the blue lines) are clearly changing over time. There is no clear pattern in the distribution of countries to either the poor or the rich group. It can be seen that the parameter π varies between 0.54 and 0.66. In the figures for μ_P and μ_R one can see that the real GDP per capita is growing over time for both groups, although a decline can be seen for the years 1981-1986, which is probably caused by the economic

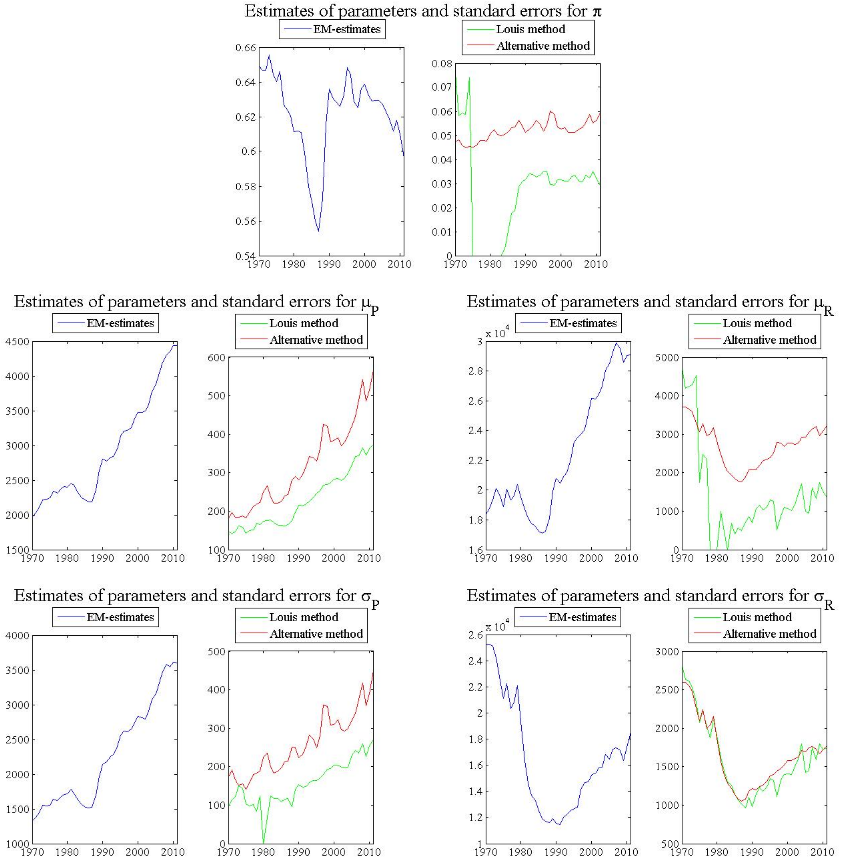


Fig. 1. Plots of the parameter and standard error estimates when using the EM-algorithm and the two methods for standard error estimation on the data set containing real Gross Domestic Product per capita of several countries for the sequential years from 1970 until 2011.

crisis of the eighties. The variance estimates, σ_P and σ_R, of the groups are quite different from each other. Where the variance of the poor group seems to be growing over time, the variance of the rich groups falls in the period 1970-1990, but has started growing after 1990 until 2011.

When looking at the standard error estimates of the parameters, it can be seen that the estimates obtained by the alternative method (represented by the red lines) are generally higher than the estimates obtained by Louis' method (represented by the green lines). The differences are most clear for the parameters μ_P and σ_P. The standard error estimates by the different methods for the parameter σ_R seem to be most similar to each other. When comparing the results of the different standard error estimation methods on the parameters π and

μ_R, it can be noticed that in the earlier years, the estimates obtained by Louis' method are higher, but after 1975, it becomes the other way around. What is perhaps even more interesting is that the standard error estimates obtained with Louis' method are zero for a couple of years for the parameters π, μ_R and σ_P. In fact, the obtained variances were negative, but that shouldn't be possible. When using the alternative method for standard error estimation, the standard error estimates are always positive.

What is clear from the results on the real GDP per capita data set, is that the standard error estimates by both methods are very different from each other. Since the actual parameter values and the actual standard errors are not known, it is hard to tell which standard error estimation method performs better. One way to be able to obtain more information about this, is to make bootstrap samples (see Tan et al. (2005)) and estimate the means and variances of the parameter estimates. The square roots of the variances can be used as another estimate of the standard errors. Furthermore, by applying the standard error estimation methods on the bootstrap samples, one can obtain variances on the standard errors. However, as the data set we use is not very large, the sizes of the bootstrap samples are too small to get good estimates when using the EM-algorithm. Therefore, instead of using bootstrap sampling with the real data set, a simulation will be performed, where data will be simulated using the parameter estimates by the EM-algorithm on the real data set. Apart from the advantage that the true parameters are known in this case, we also know for sure that the data can described by a mixture of two normal distributions. This is important, as the negative values of the variances obtained by Louis' method may be caused by the fact that the chosen distribution does not describe the data well enough. Therefore, the parameter estimates obtained by the EM-algorithm might be wrong and this might cause Louis' method to perform poorly.

5 Simulation with real parameter values

In this section, the set-up of the simulation experiment will be discussed in subsection 5.1, after which the outcomes of the experiment will be discussed in subsection 5.2.

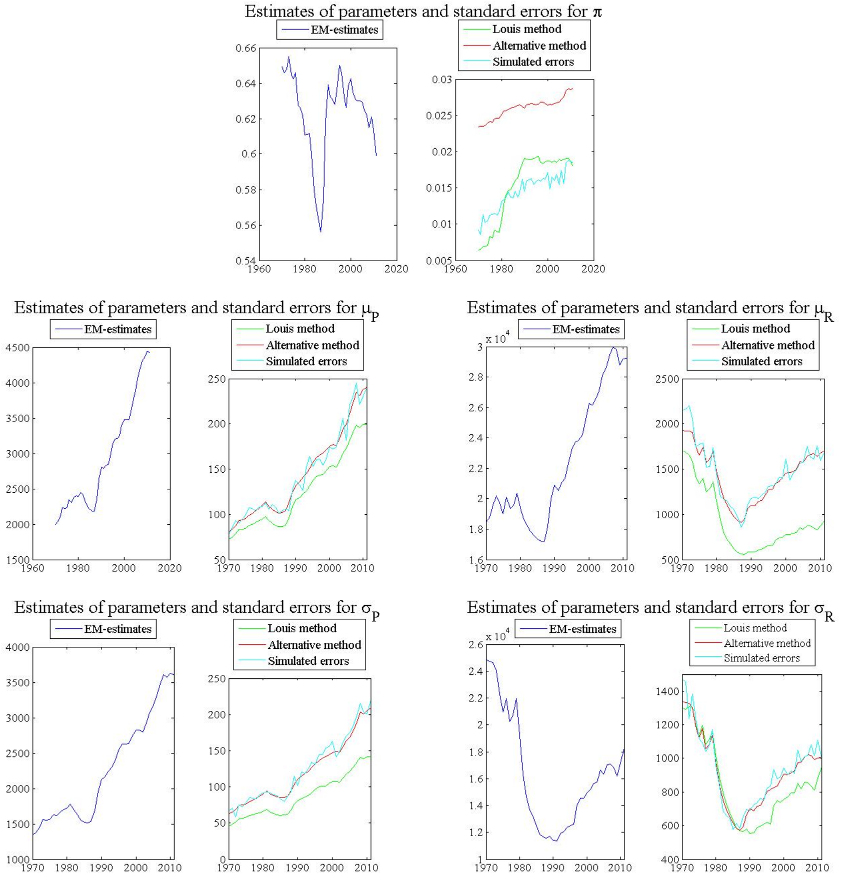
5.1 Set-up

As mentioned before, for the simulation experiment, the outcomes of the parameters obtained by using the EM-algorithm on the real GDP data set are used. That is, for each year, we use the parameters estimates of π, μ_P, μ_R, σ_P and σ_R as the true parameters of the simulation. With the given parameters, 300 data sets are simulated, each containing 500 data points. The data sets are thus larger than the real data set was. By using larger data sets, we are more certain that the EM-algorithm performs well. The simulation of 300 data sets with size 500 is repeated with every years' parameter estimates. For every year and every data set that is constructed using the parameters for that year, the EM-algorithm is applied to estimate the parameters and Louis' and the alternative method are used to get standard error estimates. As a final estimate of parameter estimates, we use the mean of the parameter estimates of all data sets for a given year. For Louis' and the alternative method, we work with the mean of the standard error estimates. Since there are 300 data sets, it is also possible to compute the empirical variance of the parameter- and standard error estimates. The square roots of the empirical variances of the parameter estimates can be used as another estimate of standard errors. These standard errors will be referred to as simulated standard errors or simply simulated errors. Although these simulated errors are only asymptotically correct, we will treat them as the true standard errors in the remaining of this section. The standard errors obtained by Louis' method and the alternative method will be compared with the simulated errors. The empirical variances of the standard error estimates will be interpreted as the variances of standard error estimates.

With the information obtained by the simulation, one can check several things. For example, one can check whether the EM-algorithm performs well in all cases. That is, it can be checked whether the parameter estimates are significantly different from the true parameters, but that is not of major importance in this research. What is more interesting in this context, is whether the standard error estimates by both methods are similar to the estimates that were obtained using the real data set. If that is not the case, there might be something wrong with the assumption that the data can be described by a mixture of two normal distributions. Another useful insight the simulation can give, is whether the alternative methods performs better than Louis' method or not. To check this, one can compare the standard errors of Louis' and the alternative method with the simulated standard errors. Furthermore, it can also be checked whether the standard errors obtained by Louis' and the alternative method are significantly different from each other. It is also possible to check whether the standard error estimates obtained by Louis' and the alternative method are significantly different from the simulated standard errors.

5.2 Results

The results of the simulation are summarized in Fig. 2. Clearly, the parameter estimates look very similar to the ones in Fig. 1, which makes sense, since we used these parameters for simulating the data sets. It indicates that the EM-algorithm is performing well. However, the standard error estimates look very different, especially the ones obtained with Louis' method. These standard errors are always positive now, so that we have an indication that the mixture of two normal distributions did not describe the real data set so well, which caused Louis' method to perform badly. In Fig. 2, one can also find the simulated standard errors. In all but for the standard error estimates of π, the estimates obtained by the alternative method are more similar to the simulated errors than the estimates obtained by Louis' method. It can be seen that the line representing the simulated errors fluctuates over the line that represents the standard errors by the alternative method for the parameters μ_P and μ_R. Louis'



Estimates of parameters and standard errors for π

Estimates of parameters and standard errors for μ_P

Estimates of parameters and standard errors for μ_R

Estimates of parameters and standard errors for σ_P

Estimates of parameters and standard errors for σ_R

Fig. 2. Plots of the parameter and standard error estimates when using the EM-algorithm and the two methods for standard error estimation on the simulated data sets with parameters for each year (1970-2011) equal to the parameters obtained by the EM-algorithm on the real GDP data set.

standard errors are somewhat lower. For the parameters σ_P and σ_R, one can see that the simulated errors are generally slightly higher than the standard errors estimated by the alternative method, while the estimates by Louis' method are even lower. For the parameter π, the results are different. In this case, it seems that Louis' estimates are closer to the simulated errors. These standard errors are also more fluctuating over time, while the standard error estimates obtained by the alternative method are more constant.

With the variances of standard errors, one can easily compute the 95% confidence intervals of the standard error estimates. When these confidence intervals are computed, one needs to conclude that the simulated errors are outside the confidence interval for all parameters and almost for all years. For some years the simulated errors are within the confidence interval of the standard errors by the alternative method for the parameters μ_P, μ_R, σ_P and σ_R, but that situation is rather exceptional. Since the standard errors estimated with the alternative method are generally closer to the simulated errors than the standard errors obtained by Louis' method for the parameters μ_P, μ_R, σ_P and σ_R, we can say that the alternative method performs better than Louis' method in estimating standard errors of the parameters of the underlying distributions. The results are different for the standard errors of the parameter π, as in this case, the estimates obtained by Louis' method are more similar to the simulated errors. This observation indicates that there is a trade-off between Louis' method and the alternative method. The first performs relatively well in estimating the standard errors of π, while the second performs better in estimating the standard errors of the parameters of the underlying distributions.

Although the estimated standard errors obtained by Louis' method and, respectively, the alternative method seem very different, it remains to be tested if they significantly differ from each other. For this matter, we will use the central limit theorem. Since the variances of the standard errors are already estimated and the data sets we have are large, we can perform a t-test on the hypothesis that $\widehat{SE_{L,i,p}} - \widehat{SE_{A,i,p}} = 0$, where $SE_{L,i,p}$ is the mean standard error of Louis method in year i and for parameter p. $SE_{A,i,p}$ represents the mean standard error of the alternative method in year i for parameter p. A t-test rejects the hypothesis for all parameters and all years at a significance level of 0.001, indicating that Louis' method and the alternative method provide significantly different estimates of the standard errors.

6 Another simulation experiment

The simulation as performed in Section 5 provides useful insights in the performance of Louis' method and the alternative method in the context of Gaussian mixture models. However, the range of parameter values that are used as input for the simulation of data sets is limited. When doing a simulation experiment, one wants to know if the results are stable across different situations. For that reason, another simulation experiment is performed. 500 data points are drawn from two different normal distributions. The proportion of data points per distribution and the parameters of the distributions are to be varied. The data points are combined into one data set, which now has a Gaussian mixture distribution with known parameters. On the data set, the EM-algorithm can be applied, after which the two methods to obtain standard error estimates can be used. The simulation of a data set and the appliance of the EM-algorithm and standard error algorithms is repeated 200 times with the same parameter settings.

The results of the simulations can be found in Table 1. In Table 1, it can be seen that when the distinction between the distributions is very clear, the results of Louis' method and the alternative method are quite similar. Furthermore, the standard error estimations by both methods are close to the values of the simu-

Table 1. The results of the simulations with different parameter settings with 500 data points per data set and 200 sets, with in parenthesis the simulated errors of the parameter and standard error estimates by the different methods. An asterisk is used to indicate whether the simulated error is within the 95% confidence interval of the standard error estimate.

		True values	**EM-estimates**	**Louis method**	**Alternative method**
1	π	0.5	0.516(0.160)	0.016(0.006)	0.134(0.048)
	μ_1	0.0	-0.011(0.334)	0.065(0.015)	0.284(0.083)
	μ_2	2.0	2.080(0.342)	0.057(0.007)	0.288(0.098)
	σ_1	1.0	0.975(0.147)	0.059(0.019)	0.129(0.026)
	σ_2	1.0	0.939(0.160)	0.050(0.008)	0.132(0.033)
2	π	0.25	0.252(0.035)	0.003(0.002)	0.039(0.010)
	μ_1	0.0	0.010(0.220)	0.107(0.019)	0.213(0.053)*
	μ_2	2.0	1.998(0.038)	0.025(0.001)	0.035(0.003)
	σ_1	1.0	0.987(0.125)	0.100(0.020)	0.126(0.025)*
	σ_2	0.5	0.496(0.030)	0.018(0.001)	0.027(0.004)
3	π	0.5	0.501(0.001)	0.022(0.000)	0.022(0.000)
	μ_1	0.0	0.003(0.060)	0.063(0.003)	0.064(0.003)
	μ_2	6.0	6.006(0.064)	0.063(0.003)	0.063(0.003)
	σ_1	1.0	1.000(0.051)	0.046(0.003)	0.047(0.003)
	σ_2	1.0	0.987(0.042)	0.045(0.002)	0.046(0.003)
4	π	0.5	0.497(0.018)	0.013(0.008)	0.029(0.003)
	μ_1	0.0	-0.003(0.084)	0.061(0.003)	0.078(0.006)
	μ_2	6.0	5.988(0.242)	0.146(0.023)	0.231(0.033)
	σ_1	1.0	0.992(0.059)	0.044(0.003)	0.060(0.006)
	σ_2	2.5	2.506(0.171)	0.101(0.040)	0.173(0.020)*
5	π	0.75	0.747(0.027)	0.017(0.001)	0.032(0.004)
	μ_1	0.0	-0.023(0.245)	0.133(0.008)	0.239(0.027)
	μ_2	6.0	5.996(0.133)	0.081(0.008)	0.132(0.022)*
	σ_1	3.0	2.971(0.173)	0.092(0.012)	0.167(0.018)
	σ_2	1.0	0.989(0.120)	0.061(0.006)	0.112(0.021)

lated standard errors. This observation implies that when the EM-algorithm is used on a data set where observations clearly come from different distributions, it does not really matter a lot which method you use for standard error estimation. When it is less clear to which underlying distribution a data point belongs and it is thus harder to cluster the data to their underlying distribution, the performances of the methods are very different with respect to each other. In most cases, it holds that the standard errors obtained by the alternative method are higher than the standard errors obtained by Louis' method, as was the case in Section 5.

When one would focus solely on the observation that the standard errors by the alternative method have higher values than those by Louis' method, it is hard to tell which of the methods performs better, as in which method gives standard errors that are similar or at least closer to the true standard errors of parameter estimates. Thus, when determining which standard error estimates are more convenient, one needs to have the true estimates. Although the simulated standard errors are only asymptotically correct, I will consider these to be the true standard error estimates, as was done in Section 5 as well. In Table 1, it can be seen that the simulated errors are within the 95% confidence interval of the standard errors estimated with the alternative method in some exceptional cases. Furthermore, it can be seen that when the distinction between the distributions is less clear, the standard errors of the parameters generated by the alternative method are generally closer to the simulated standard errors. However, the errors obtained by this method are in most cases still lower than the simulated errors, suggesting that both methods underestimate the true parameter values when it is harder to determine to which group a data point belongs. This result is remarkable, since in Kang et al. (2013) it is stated that Louis method overestimates standard errors. It seems that the distribution, thus the application of the EM-algorithm, makes a difference when evaluating the performance of Louis' method. It remains to be investigated whether this is also the case with the alternative method.

7 Conclusion

In this research, we investigated the performance of two methods that are to obtain standard error estimates when using the EM-algorithm. One of the methods is Louis' method, which is currently the most popular method in the field. For this method, only the gradient and second derivative matrix of the complete-data log likelihood are needed. As an alternative method, we used the negative of the second derivative matrix of the incomplete-data problem as an approximation of the information matrix. To test the performance of the two methods, a data set containing real GDP per capita was used. Earlier research has shown that the data set can be described by a bimodal distribution and in this research, a mixture of two normal distributions was used.

The results on the real GDP data set show that the standard error estimates obtained by the alternative method are generally higher than the estimates obtained by Louis' method. The estimates obtained by Louis' method are sometimes invalid, as the variance estimates are below zero in some cases. With a long existing and commonly used method such as Louis' method, this shouldn't be possible. An explanation could be that the data cannot be described well enough by a mixture of two normal distributions. That is one of the reasons why a simulation with the parameters that were estimated by the EM-algorithm on the real data set was performed. Findings of the simulation are that the standard errors obtained by both method are significantly different from each other and from the simulated errors, but the estimates by the alternative method are closer to the simulated standard error estimates for all parameters but for π.

To obtain even more insights in the performance of Louis' method and the alternative method in the context of Gaussian mixture models, a second simulation experiment was performed. The experiment shows us that when it is relatively easy to determine to which group a data point belongs, the results of both methods are quite similar and are close to the simulated errors as well. When it is harder to determine to which group an observation belongs, the standard errors by Louis' method are generally below those obtained the alternative method. However, both methods tend to underestimate the standard errors of parameters when the clustering of data points is more difficult.

A limitation of this research is that the performance of the two methods is only tested on Gaussian mixture models with two components. It remains to be investigated whether the alternative method performs better than Louis' method when other distributions are used as well. Furthermore, it might in some cases not be easy to obtain derivatives of the incomplete-data log likelihood. However, it is not always easy to obtain the derivatives of the complete-data problem as well. Further research needs to be done to investigate whether obtaining derivatives of the incomplete-data problem is harder than getting the derivatives of the complete-data problem.

References

Baker, S. G. (1992). A Simple Method for Computing the Observed Information Matrix When Using the EM Algorithm with Categorical Data. *Journal of Computational and Graphical Statistics*, 1(1):pp. 63–76.

Cappé, O. (2011). Online EM Algorithm for Hidden Markov Models. *Journal of Computational and Graphical Statistics*, 20(3):pp. 728–749.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):pp. 1–38.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426):pp. 463–475.

Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2013). The Next Generation of the Penn World Table.

Hasselblad, V. (1966). Estimation of Parameters for a Mixture of Normal Distributions. *Technometrics*, 8(3):pp. 431–444.

Kang, L., Carter, R., Darcy, K., Kauderer, J., and Liao, S.-Y. (2013). A fast Monte Carlo expectation–maximization algorithm for estimation in latent class model analysis with an application to assess diagnostic accuracy for cervical neoplasia in women with atypical glandular cells. *Journal of Applied Statistics*, 40(12):2699–2719.

Kim, T. and Lee, H.-Y. (2011). External Validity of Market Segmentation Methods: A study of buyers of prestige cosmetic brands. *European Journal of Marketing*, 45(1/2):pp.153 – 169.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):pp. 226–233.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley series in probability and statistics. Wiley, 2. ed edition.

Melnykov, V. and Melnykov, I. (2012). Initializing the EM Algorithm in Gaussian Mixture Models with an Unknown Number of Components. *Computational Statistics & Data Analysis*, 56(6):pp. 1381 – 1395.

Oakes, D. (1999). Direct Calculation of the Information Matrix via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(2):pp. 479–482.

Paap, R. and van Dijk, H. (2009). *Distribution and Mobility of Wealth of Nations*. S.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):pp. 699–704.

Xu, C., Baines, P. D., and Wang, J.-L. (2014). Standard Error Estimation Using the EM Algorithm for the Joint Modeling of Survival and Longitudinal Data. *Biostatistics*.

A The second derivatives of the incomplete data problem

The second derivative matrix of the incomplete data problem is symmetric and has the following form:

$$B(\boldsymbol{y};\boldsymbol{\psi}) = \begin{bmatrix} B(1,1) & B(1,2) & B(1,3) & B(1,4) & B(1,5) \\ B(1,2) & B(2,2) & B(2,3) & B(2,4) & B(2,5) \\ B(1,3) & B(2,3) & B(3,3) & B(3,4) & B(3,5) \\ B(1,4) & B(2,4) & B(1,3) & B(4,4) & B(4,5) \\ B(1,5) & B(2,5) & B(3,5) & B(4,5) & B(5,5) \end{bmatrix} \tag{8}$$

All the elements in $B(\boldsymbol{y};\boldsymbol{\psi})$ are of the form $\sum_{i=1}^{n} \frac{g'h-gh'}{h^2}$, where h is always the same, namely $h = \pi\phi_1 + (1-\pi)\phi_2$. Herein, ϕ_1 represents the density function of the normal distribution with parameters $\boldsymbol{\theta}_1$ and ϕ_2 represents the density function of the normal distribution with parameters $\boldsymbol{\theta}_2$. The derivatives for g, g' and h' are different for the different elements of $B(\boldsymbol{y};\boldsymbol{\psi})$ and are now specified, starting with the diagonal elements.

$B(1,1):$
$g = (\phi_1 - \phi_2)$
$g' = 0$
$h' = (\phi_1 - \phi_2)$

$B(2,2):$
$g = \pi\phi_1(\frac{y_i-\mu_1}{\sigma_1^2})$
$g' = \frac{\pi}{\sigma_1^2}\phi_1(-1 + (\frac{y_i-\mu_1}{\sigma_1})^2)$
$h' = g$

$B(3,3):$
$g = (1-\pi)\phi_2(\frac{y_i-\mu_2}{\sigma_2^2})$
$g' = \frac{1-\pi}{\sigma_2^2}\phi_2(-1 + (\frac{y_i-\mu_2}{\sigma_2})^2)$
$h' = g$

$B(4,4):$
$g = \frac{\pi}{\sigma_1}\phi_1(-1 + (\frac{y_i-\mu_1}{\sigma_1})^2)$
$g' = \frac{\pi}{\sigma_1^2}(2 - 5(\frac{y_i-\mu_1}{\sigma_1})^2 + (\frac{y_i-\mu_1}{\sigma_1})^4)$
$h' = g$

$B(5,5):$
$g = \frac{1-\pi}{\sigma_2}\phi_2(-1 + (\frac{y_i-\mu_2}{\sigma_2})^2)$
$g' = \frac{1-\pi}{\sigma_2^2}(2 - 5(\frac{y_i-\mu_2}{\sigma_2})^2 + (\frac{y_i-\mu_2}{\sigma_2})^4)$
$h' = g$

$B(1,2):$
$g = (\phi_1 - \phi_2)$
$g' = \phi_1(\frac{y_i-\mu_1}{\sigma_1^2})$
$h' = \pi\phi_1(\frac{y_i-\mu_1}{\sigma_1^2})$

$B(1,3):$
$g = (\phi_1 - \phi_2)$
$g' = -\phi_2(\frac{y_i-\mu_2}{\sigma_2^2})$
$h' = (1-\pi)\phi_1(\frac{y_i-\mu_2}{\sigma_2^2})$

$B(1,4):$
$g = (\phi_1 - \phi_2)$
$g' = (\frac{1}{\sigma_1}\phi_1(-1 + (\frac{y_i-\mu_1}{\sigma_1})^2))$
$h' = \frac{\pi}{\sigma_1}\phi_1(-1 + (\frac{y_i-\mu_1}{\sigma_1})^2)$

$B(1,5):$
$g = (\phi_1 - \phi_2)$
$g' = (-\frac{1}{\sigma_2}\phi_2(-1 + (\frac{y_i-\mu_2}{\sigma_2})^2))$
$h' = \frac{1-\pi}{\sigma_2}\phi_2(-1 + (\frac{y_i-\mu_2}{\sigma_2})^2)$

$B(2,3):$
$g = \pi\phi_1(\frac{y_i - \mu_1}{\sigma_1^2})$
$g' = 0$
$h' = (1-\pi)\phi_2(\frac{y_i - \mu_2}{\sigma_2^2})$

$B(2,4):$
$g = \pi\phi_1(\frac{y_i - \mu_1}{\sigma_1^2})$
$g' = \frac{\pi(y_i - \mu_1)}{\sigma_1^3}\phi_1(-3 + (\frac{y_i - \mu_1}{\sigma_1})^2)$
$h' = \frac{\pi}{\sigma_1}\phi_1(-1 + (\frac{y_i - \mu_1}{\sigma_1})^2)$

$B(2,5):$
$g = \pi\phi_1(\frac{y_i - m_1}{\sigma_1^2})$
$g' = 0$
$h' = \frac{1-\pi}{\sigma_2}\phi_2(-1 + (\frac{y_i - \mu_2}{\sigma_2})^2)$

$B(3,4):$
$g = (1-\pi)\phi_2(\frac{y_i - \mu_2}{\sigma_2^2})$
$g' = 0$
$h' = \frac{\pi}{\sigma_1}\phi_1(-1 + (\frac{y_i - \mu_1}{\sigma_1})^2)$

$B(3,5):$
$g = (1-\pi)\phi_2(\frac{y_i - \mu_2}{\sigma_2^2})$
$g' = \frac{(1-\pi)(y_i - \mu_2)}{\sigma_2^3}\phi_2(-3 + (\frac{y_i - \mu_2}{\sigma_2})^2)$
$h' = \frac{1-\pi}{\sigma_2}\phi_2(-1 + (\frac{y_i - \mu_2}{\sigma_2})^2)$

$B(4,5):$
$g = \frac{\pi}{\sigma_1}\phi_1(-1 + (\frac{y_i - \mu_1}{\sigma_1})^2)$
$g' = 0$
$h' = \frac{1-\pi}{\sigma_2}\phi_2(-1 + (\frac{y_i - \mu_2}{\sigma_2})^2)$