



A Stochastic Version of the EM Algorithm to Analyze Multivariate Skew-Normal Data with Missing Responses

M. Khounsiavash¹, M. Ganjali^{*2} and T. Baghfalaki²

¹Department of Statistics
Science and Research Branch
Islamic Azad University
Tehran, Iran

²Department of Statistics
Shahid Beheshti University
Tehran, Iran
m-ganjali@sbu.ac.ir

*Correspondence author

Received: January 21, 2011; Accepted: September 12, 2011

Abstract

In this paper an algorithm called SEM, which is a stochastic version of the EM algorithm, is used to analyze multivariate skew-normal data with intermittent missing values. Also, a multivariate selection model framework for modeling of both missing and response mechanisms is formulated. By the SEM algorithm missing values of responses are inputted by the conditional distribution of missing values given observed data and then the log-likelihood of the pseudo-complete data is maximized. The algorithm is iterated until convergence of parameter estimates. Results of an application are also reported where a Bootstrap approach is used to compute the standard error of the parameter estimates.

Keywords: Skew-normal distribution; Stochastic EM algorithm; Intermittent missingness; Selection model; Bootstrap.

MSC 2010: 62P99

1. Introduction

The skew-normal is a class of distribution that includes the normal distribution as a special case. In this distribution an extra parameter, λ , measures the skewness. A systematic treatment of the skew-normal distribution has been given in Azzalini (1985) and Henze (1986). Azzalini and Della-Valle (1996), and Azzalini and Capitanò (1999), generalized this distribution to the multivariate case. Arellano-Valle et al. (2002) show that many of the properties of the multivariate skew-normal distribution hold for a general class of skewed distribution. Such classes obtained from asymmetric distribution, defined in terms of independence conditions on signs and absolute values and give a general formula to obtain skewed pdf's. From these results, Arellano-Valle and Genton (2005), introduced the class of fundamental skewed distributions, and gave a unified approach to obtain multivariate skew distributions starting with symmetric distributions.

In this study, we use a version of multivariate skew-normal distribution, which was introduced by Azzalini and Dalla-Valle (1996), and is a special case of the fundamental skew-normal distribution proposed by Arellano-Valle and Genton (2005).

In this way we consider a $p \times 1$ random vector \mathbf{Y} as a multivariate SN random variable with $p \times 1$ location vector $\boldsymbol{\mu}$, and $p \times p$ positive definite dispersion matrix $\boldsymbol{\Sigma}$ and $p \times 1$ skewness parameter vector $\boldsymbol{\lambda}$, and write $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. The standard multivariate skew-normal distribution will be denoted by $SN_p(\boldsymbol{\lambda})$.

Longitudinal data are measurements of individual subjects over a period of time; these kinds of measurements are frequently used in medical, public health and social sciences. The response variable may be continuous, categorical or ordinal. One of the main interests of these studies, is to investigate the change in the response variable over time.

In this study, missing data occur whenever one or more of, measurement sequences are incomplete. Rubin (1976) and Little and Rubin (1987) provided a framework for the incomplete data by introducing a taxonomy of missing data mechanisms, consisting of missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In the MCAR mechanism, the missing values are independent of both observed and unobserved data, in the MAR mechanism conditioning on the observed data, the missing mechanism is independent of missing values, and otherwise the missing process is named as MNAR, informative or non-ignorable mechanisms, and ignoring the missing values with such data would lead to biased conclusions. Another important feature is whether the missing values pattern is dropout (monotone) or intermittent (non-monotone). In dropout pattern some subjects may withdraw permanently, i.e. a missing value is never followed by an observed value. In the intermittent pattern an observed value is available even after a missing value occurs. Diggle and Kenward (1994) defined the dropout process to be completely random dropout (CRD), random dropout (RD) and nonrandom dropout (NRD) with the same concepts as those mentioned for MCAR, MAR and MNAR, respectively.

Diggle and Kenward (1994) use a modeling framework for longitudinal data which decomposes the joint distribution of missing mechanism and responses into a marginal distribution for longitudinal continuous responses and a conditional distribution for missing mechanism given responses. Let M_i denote the associated vector of missingness indicator which is related to the standard multivariate skew-normal vector \mathbf{Y}_i , such that $M_{ij} = 1$ if Y_{ij} (the j^{th} response of the i^{th} subject) is missing and otherwise $M_{ij} = 0$.

When the missing mechanism is MNAR, three modeling frameworks may be used to model the missing mechanism and responses jointly. These are the selection, the pattern-mixture and the shared parameter models. These models are defined by the conditional factorization of joint distribution of Y and M . The selection model factorization is as follows:

$$f(\mathbf{y}_i, M_i | \mathcal{G}, \boldsymbol{\psi}) = f(\mathbf{y}_i | \mathcal{G}) f(M_i | \mathbf{y}_i, \boldsymbol{\psi}), \quad (1)$$

where \mathcal{G} and $\boldsymbol{\psi}$ denote distinct parameter vectors of the measurements and missingness mechanisms, respectively. The first factor on the right of the equation (1) is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the outcomes.

Another factorization so called pattern-mixture model (Little 1993, 1994), is as follow:

$$f(\mathbf{y}_i, M_i | \mathcal{G}, \boldsymbol{\psi}) = f(\mathbf{y}_i | M_i, \mathcal{G}) f(M_i | \boldsymbol{\psi}). \quad (2)$$

The third model referred to as shared-parameter model is:

$$f(\mathbf{y}_i, M_i | \mathcal{G}, \boldsymbol{\psi}, \mathbf{b}_i) = f(\mathbf{y}_i | M_i, \mathcal{G}, \mathbf{b}_i) f(M_i | \boldsymbol{\psi}, \mathbf{b}_i), \quad (3)$$

where we explicitly include a vector of unit-specific latent (or random) effects \mathbf{b}_i of which one or more components are shared between both components in the joint distribution, some references to such modeling approach include Wu and Carroll (1988), and Crouchley and Ganjali (2002).

In this study we consider the selection model framework for multivariate skew-normal with a probit regression as the missingness mechanism.

The EM algorithm, Dempster et al. (1977), is a very useful tool for the iterative computation of maximum likelihood estimates, in missing or incomplete data problems, where algorithms such as the Newton-Raphson method may turn out to be more complicated. In each iteration of the EM algorithm, there are two steps called the expectation step or the E-step and the maximization step or the M-step. Because of this, the algorithm is called the EM algorithm.

The main problem of the EM algorithm is that the expectation step may be infeasible, especially when this expectation is a high dimensional integral or a large sum or an integral over an irregular region, thus it can not be calculated explicitly. Many authors have tried to introduce

new variants of the EM algorithm that can overcome the complexity of the problem. A possible solution for intractable E-step is to use a stochastic version of the EM algorithm, (Celux and Diebolt, 1985; Delyon et al. 1999; Diebolt and Ip, 1996; Zhu and Lee, 2002).

A brief history of the EM algorithm can be found in McLachlan and Krishnan (1997), and references therein. The stochastic EM (SEM) algorithm is a stochastic version of the EM algorithm which was introduced by Celux and Diebolt (1985), and Diebolt and Ip (1996), as a way for executing the E-step using simulation.

When some subjects leave the study temporarily and subsequently return, i.e. an observed value is available even after a missing value occurs, then the missing data pattern is defined as intermittent or non-monotone. Gad and Ahmed (2006) proposed the SEM algorithm to handle intermittent missing data patterns, in selection models for multivariate normal responses. In Section 2 we will present our extended model which may be used for a vector of response with multivariate skew-normal distribution. In Section 3 we will discuss the EM algorithm and explain the SEM algorithm for analyzing multivariate skew-normal responses. In Section 4 simulation study and in Section 5 an application of the model will be presented. The conclusion will be discussed in Section 6.

2. Selection Model for Longitudinal Data with Intermittent Missing Responses Using Multivariate Skew-Normal Distribution

A random variable Z , has a skew-normal distribution if its density function is given by

$$f(z; \lambda) = 2\phi(z) \Phi(\lambda z) \quad z, \lambda \in R, \quad (4)$$

where in brief we write $Z \sim SN(0,1)$, here ϕ and Φ , respectively denote density and distribution functions of the standard normal distribution. It is clear that when $\lambda = 0$, Z has a standard normal distribution, and the sign of λ gives the direction of the skewness.

If we use location and scale parameters μ and σ , for more flexibility, then the density of the new random variable Y , in brief written as $Y \sim SN(\mu, \sigma, \lambda)$, is

$$f(y; \mu, \sigma, \lambda) = \frac{2}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) \Phi\left(\lambda \frac{y-\mu}{\sigma}\right). \quad (5)$$

If $Y \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, its probability density function is given by:

$$f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi\left(\boldsymbol{\lambda}' \boldsymbol{\Sigma}^{-\frac{1}{2}} (\mathbf{y} - \boldsymbol{\mu})\right), \quad (6)$$

where $\phi_p(\cdot | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the pdf of a p -variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Note that if in expression (6) we let $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = I_p$, we have a standard multivariate skew-normal distribution with skewness parameter vector $\boldsymbol{\lambda}$.

Suppose that we have n independent subjects with repeated measurements. Each subject i is introduced by a skew outcome Y_{ij} designed to be measured at times j ($j = 1, 2, \dots, T$). Assume that the observed and missing components of \mathbf{Y}_i are denoted as $\mathbf{Y}_{i,obs}$ and $\mathbf{Y}_{i,mis}$, respectively. Let M_i be a vector of the missingness indicator, such that for a particular realization of (\mathbf{Y}_i, M_i) , each element of M_i (M_{ij}) gets one or zero if its corresponding element of Y_i is missing or observed.

In a selection model, the joint distribution of \mathbf{Y}_i and M_i is factorized as product of the marginal distribution of \mathbf{Y}_i and the conditional distribution of M_i given \mathbf{Y}_i as in (1), which in our study we assumed $\mathbf{Y}_i \sim SN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ and for the conditional distribution of M_i given \mathbf{Y}_i , a probit model is considered as

$$P(M_{ij} = 1 | \mathbf{Y}_i, \boldsymbol{\Psi}) = \Phi(\psi_0 + \psi_1 Y_{ij-1} + \psi_2 Y_{ij})$$

where $\boldsymbol{\Psi} = (\psi_0, \psi_1, \psi_2)'$. The special case of the above model corresponding to MAR and MCAR are obtained from setting $\psi_2 = 0$ and $\psi_1 = \psi_2 = 0$, respectively.

Let $\mathcal{G} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, the joint likelihood of the observed data $(Y_{i,obs}, M_i)$ is:

$$\begin{aligned} L(\mathcal{G}, \boldsymbol{\Psi} | \mathbf{Y}_{obs}, M) &= \prod_{i=1}^n f(\mathbf{y}_{i,obs}, M_i | \mathcal{G}, \boldsymbol{\Psi}) = \prod_{i=1}^n f(\mathbf{y}_{i,obs} | \mathcal{G}) \times f(\mathbf{M}_i | \mathbf{y}_i, \boldsymbol{\Psi}) \\ &= \prod_{i=1}^n \int f(\mathbf{y}_i | \mathcal{G}) \times f(M_i | \mathbf{y}_i, \boldsymbol{\Psi}) d\mathbf{y}_{i,mis} \\ &= \prod_{i=1}^n \int 2\phi_p(\mathbf{y}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi(\boldsymbol{\lambda}' \boldsymbol{\Sigma}^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu})) \times f(M_i | \mathbf{y}_i, \boldsymbol{\Psi}) d\mathbf{y}_{i,mis} \end{aligned}$$

Parameter estimates of this observed likelihood can be found using numerical method such as Newton-Raphson, but because of flatness of this likelihood function and complexity in obtaining the parameter estimates, we shall propose the use of the SEM method.

If there are some explanatory variables for individual i , $i = 1, 2, \dots, n$, then $\mathbf{Y}_i \sim SN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, where $\boldsymbol{\mu}_i = \mathbf{x}_i' \boldsymbol{\beta}$, \mathbf{x}_i is a $q \times p$ explanatory matrix and $\boldsymbol{\beta}$ is a $q \times 1$ regression coefficient vector. In this situation one has to estimate $\boldsymbol{\beta}$ instead of $\boldsymbol{\mu}$ in (6).

The following proposition is fundamental in our study.

Proposition 1: Let $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, such that \mathbf{Y} is partitioned into two sub-vectors of interest, \mathbf{Y}_1 and \mathbf{Y}_2 , where \mathbf{Y}_1 is $p_1 \times 1$ and \mathbf{Y}_2 is $(p - p_1) \times 1$, then

$$\mathbf{Y}_1 \sim SN_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\lambda}_1^*),$$

where

$$\boldsymbol{\lambda}_1^* = \boldsymbol{\Sigma}_{11}^{-1/2} \left(\frac{\boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_1}{\sqrt{1 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22.1} \boldsymbol{\gamma}_2}} \right), \quad \boldsymbol{\gamma} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\lambda}$$

has dimension $p_1 \times 1$, also $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}$ are partitioned to conform with $(\mathbf{Y}_1', \mathbf{Y}_2')$ where $\boldsymbol{\mu}_1, \boldsymbol{\lambda}_1, \boldsymbol{\gamma}_1$ are $p_1 \times 1$, $\boldsymbol{\Sigma}_{11}$ is $p_1 \times p_1$, $\boldsymbol{\mu}_2, \boldsymbol{\lambda}_2, \boldsymbol{\gamma}_2$ are $(p - p_1) \times 1$, $\boldsymbol{\Sigma}_{22}$ is $(p - p_1) \times (p - p_1)$, $\boldsymbol{\Sigma}_{12}$ is $p_1 \times (p - p_1)$ and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}'$, also $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$.

$$f(\mathbf{y}_2 | \mathbf{y}_1) = \phi_{p-p_1}(\mathbf{y}_2 | \boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}) \frac{\Phi(\boldsymbol{\lambda}' \boldsymbol{\Sigma}_{11}^{-1/2} (\mathbf{y}_1 - \boldsymbol{\mu}_1))}{\Phi(\boldsymbol{\lambda}_1^* \boldsymbol{\Sigma}_{11}^{-1/2} (\mathbf{y}_1 - \boldsymbol{\mu}_1))},$$

where

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{y}_1 - \boldsymbol{\mu}_1).$$

The distribution of \mathbf{Y}_1 is obtained by integrating out \mathbf{Y}_2 , under expressions given in proposition, we have

$$\begin{aligned} f_{\mathbf{Y}_1}(\mathbf{y}_1) &= \int 2\phi(\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi(\boldsymbol{\lambda}' \boldsymbol{\Sigma}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})) d\mathbf{y}_2 \\ &= 2\phi(\mathbf{y}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \int \phi(\mathbf{y}_2 | \boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}) \Phi(\boldsymbol{\lambda}' \boldsymbol{\Sigma}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})) d\mathbf{y}_2 \\ &= 2\phi(\mathbf{y}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \int \phi(\mathbf{y}_2 | \boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}) \Phi(\boldsymbol{\gamma}_1 \mathbf{y}_1 - \boldsymbol{\gamma}' \boldsymbol{\mu} + \boldsymbol{\gamma}_2 \mathbf{y}_2) d\mathbf{y}_2 \\ &= 2\phi(\mathbf{y}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \int \phi(\mathbf{z} | \mathbf{0}, \boldsymbol{\Sigma}_{22.1}) \Phi(\boldsymbol{\gamma}_1 \mathbf{y}_1 - \boldsymbol{\gamma}' \boldsymbol{\mu} + \boldsymbol{\gamma}_2 (\mathbf{z} + \boldsymbol{\mu}_{2.1})) d\mathbf{z} \\ &= 2\phi(\mathbf{y}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \Phi\left(\frac{\boldsymbol{\gamma}_1 \mathbf{y}_1 - \boldsymbol{\gamma}' \boldsymbol{\mu} + \boldsymbol{\gamma}_2 \boldsymbol{\mu}_{2.1}}{\sqrt{1 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22.1} \boldsymbol{\gamma}_2}}\right) \\ &= 2\phi(\mathbf{y}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \Phi(\boldsymbol{\lambda}_1^* \boldsymbol{\Sigma}_{11}^{-1/2} (\mathbf{y}_1 - \boldsymbol{\mu}_1)); \quad \boldsymbol{\lambda}_1^* = \boldsymbol{\Sigma}_{11}^{1/2} \left(\frac{\boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_1}{\sqrt{1 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22.1} \boldsymbol{\gamma}_2}} \right). \end{aligned}$$

$$\begin{aligned} f(\mathbf{y}_2|\mathbf{y}_1) &= \frac{f(\mathbf{y})}{f(\mathbf{y}_1)} = \frac{\phi(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi(\boldsymbol{\lambda}' \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}))}{\phi(\mathbf{y}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \Phi(\boldsymbol{\lambda}_1' \boldsymbol{\Sigma}_{11}^{-1/2}(\mathbf{y}_1 - \boldsymbol{\mu}_1))} \\ &= \phi(\mathbf{y}_2|\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}) \frac{\Phi(\boldsymbol{\lambda}' \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}))}{\Phi(\boldsymbol{\lambda}_1' \boldsymbol{\Sigma}_{11}^{-1/2}(\mathbf{y}_1 - \boldsymbol{\mu}_1))}. \end{aligned}$$

In our study to generate skew-normal random number, we will use a stochastic representation of the multivariate skew-normal as

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\delta |T_0| + (\mathbf{I}_p - \delta \delta')^{\frac{1}{2}} \mathbf{T}_1 \right) \quad (7)$$

with

$$\boldsymbol{\delta} = \frac{\boldsymbol{\lambda}}{\sqrt{1 + \boldsymbol{\lambda}' \boldsymbol{\lambda}}}$$

where $|T_0|$ denotes the absolute value of T_0 , $T_0 \sim N(0,1)$, and $\mathbf{T}_1 \sim N_p(\mathbf{0}, \mathbf{I}_p)$. For more details on this approach, see Arellano-Valle and Genton (2005) and Arellano-Valle et al. (2005).

3. The EM Algorithm and its Stochastic Version

At first we briefly review the basic idea of the EM algorithm (Dempster et al., 1977). The EM algorithm is an iterative procedure to find the maximum of likelihood function in incomplete data problems. In each iteration, the EM algorithm performs an expectation and a maximization step. Let $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \boldsymbol{\psi})$ and $\boldsymbol{\theta}^{(t-1)}$ denote the current parameter value. Then in the t^{th} iteration of algorithm, given the observed data and current parameter value, the E-step computes the conditional expectation of the complete data log-likelihood:

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t-1)}) = E(\log f(\mathbf{Y}, M; \boldsymbol{\theta}) | \mathbf{Y}_{obs}, \boldsymbol{\theta}^{(t-1)}, M).$$

Then in the M step by maximizing the $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t-1)})$, $\boldsymbol{\theta}^{(t)}$, is computed. Given an initial value $\boldsymbol{\theta}^{(0)}$, the EM algorithm generates a sequence $\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots\}$ that under regularity condition (Wu, 1983), converges. Since the conditional expectation plays an important role in the EM algorithm, it is often referred to as the Q-function. The EM algorithm has a basic property that, in every iteration guarantees an increase in the likelihood function. But when there are several stationary points (local maxima and minima, saddle points), the EM does not necessarily converge to a significant maxima. In addition, when the likelihood surface is littered with saddle points and sub-optimal maxima, the limiting position of the EM greatly depends on its initial points.

In order to go around the above problems of EM algorithm, we describe how the SEM algorithm works. It has been shown that this algorithm is computationally less burdensome and more

appropriate than the EM algorithm for missing data (Ip, 1994). In addition this algorithm can cover the problem of likelihood multimodality surface (Ip, 1994).

3.1. Stochastic EM Algorithm

The stochastic version of the EM algorithm has three steps:

- Simulation and approximation: in SEM the E-step of the EM algorithm is replaced by a single draw from conditional distribution of the missing data given the observed data.
- Maximization: in the maximization step, after filling the empty cells (inputting missing data), the log-likelihood function will be maximized using the usual maximization procedures, for instance Newton-Raphson.
- Iteration and convergence: this step decides how long the algorithm is to be run and determines the stopping rule.

If the iterations converge, the final output of the SEM algorithm is a sample from the stationary distribution of the parameter, the mean of this sample, after turning in the first early point, is considered an SEM estimate for θ .

A relatively recent overview of simulation types was given in Jank (2005) and references there in. In order to simulate from conditional distribution, we use the most flexible and generally applicable approach, Gibbs sampler (see Robert and Casella, 2002).

Assume that the missing components of \mathbf{Y}_i are denoted as $\mathbf{Y}_{i,mis}$ and assume that this vector is of dimension $a_1 \times r$, i.e., $\mathbf{Y}_{i,mis} = (\mathbf{Y}_{i,mis_1}, \dots, \mathbf{Y}_{i,mis_r})$. To implement the SEM algorithm, a sample is drawn from the conditional distribution of the missing data, $\mathbf{Y}_{i,mis} = (\mathbf{Y}_{i,mis_1}, \dots, \mathbf{Y}_{i,mis_r})$, given the observed data, $(\mathbf{Y}_{i,obs}, M_i)$. At the $(t+1)^{th}$ iteration $\mathbf{Y}_{i,mis}^{(t+1)} = (\mathbf{Y}_{i,mis_1}^{(t+1)}, \dots, \mathbf{Y}_{i,mis_r}^{(t+1)})$ is simulated from the full conditional distributions. This iteration is executed in the r sub-step. First, $\mathbf{Y}_{i,mis_1}^{(t+1)}$ is simulated from the conditional distribution $f(\mathbf{Y}_{i,mis_1} | \mathbf{Y}_{i,mis_2}^{(t)}, \dots, \mathbf{Y}_{i,mis_r}^{(t)}, \mathbf{Y}_{i,obs}, M_i, \theta^{(t)})$. Then, in the second sub-step, $\mathbf{Y}_{i,mis_2}^{(t+1)}$ is simulated from the conditional distribution

$$f(\mathbf{Y}_{i,mis_2} | \mathbf{Y}_{i,mis_1}^{(t+1)}, \dots, \mathbf{Y}_{i,mis_r}^{(t)}, \mathbf{Y}_{i,obs}, M_i, \theta^{(t)}).$$

In the third sub-step, $\mathbf{Y}_{i,mis_3}^{(t+1)}$ is simulated from the distribution

$$f(\mathbf{Y}_{i,mis_3} | \mathbf{Y}_{i,mis_1}^{(t+1)}, \mathbf{Y}_{i,mis_2}^{(t+1)}, \dots, \mathbf{Y}_{i,mis_r}^{(t)}, \mathbf{Y}_{i,obs}, M_i, \theta^{(t)})$$

In the last sub-step, the last missing value $Y_{i,mis_r}^{(t+1)}$ is simulated from the conditional distribution

$$f\left(\mathbf{Y}_{i,mis_r} \mid \mathbf{Y}_{i,mis_1}^{(t+1)}, \mathbf{Y}_{i,mis_2}^{(t+1)}, \dots, \mathbf{Y}_{i,mis_{r-1}}^{(t+1)}, \mathbf{Y}_{i,obs}, M_i, \boldsymbol{\theta}^{(t)}\right)$$

Now, the steps of the SEM algorithm can be developed in the current setting as follows:

S-Step: At the $(t+1)^{th}$ iteration, a sample is drawn from the conditional distribution of the missing value, $\mathbf{Y}_{i,mis} = (\mathbf{Y}_{i,mis_1}, \dots, \mathbf{Y}_{i,mis_r})$, given the observed data $\mathbf{Y}_{i,obs}, M_i$, and the current parameter estimate, $\boldsymbol{\theta}^{(t)}$. The full conditional distribution does not have a standard form, hence it is not possible to simulate directly from it. An accept-reject procedure is proposed for generating the missing values. The procedure is as follows:

1. Generate a candidate value, \mathbf{y}^* , from the conditional distribution

$$f\left(\mathbf{Y}_{i,mis_j} \mid \mathbf{Y}_{i,obs}, \mathbf{Y}_{i,mis_1}^{(t+1)}, \dots, \mathbf{Y}_{i,mis_{j-1}}^{(t+1)}, \mathbf{Y}_{i,mis_{j+1}}^{(t)}, \dots, \mathbf{Y}_{i,mis_r}^{(t)}, \boldsymbol{g}^{(t)}\right)$$

for

$$j = 1, 2, \dots, r.$$

2. Calculate the probability of missingness for the candidate value, \mathbf{y}^* , according to the missing data mechanism, (which in our study is a probit model), with the parameter $\boldsymbol{\psi}$ fixed at the current values $\boldsymbol{\psi}^{(t)}$. Let us denote the resulting value as P_i . The probability of missingness will be assumed to depend only on the current and the previous response values.

3. Simulate a random variate U from the uniform distribution on the interval $[0,1]$ then take $\mathbf{Y}_{i,mis_j} = \mathbf{y}^*$ if $U \leq P_i$; otherwise go to step 1.

M-Step: with the pseudo-complete data which is defined as \mathbf{Y}^{ps} , a likelihood maximization routine is then used to obtain updated parameters $\boldsymbol{\theta}^{(t+1)}$. The likelihood of the pseudo-complete data for each subject can be written as

$$f\left(\mathbf{Y}_i^{ps}, M_i \mid \boldsymbol{\theta}^{(t+1)}\right) = P\left(M_i \mid \mathbf{Y}_i^{ps}, \boldsymbol{\theta}^{(t+1)}\right) f\left(\mathbf{Y}_i^{ps} \mid \boldsymbol{\theta}^{(t+1)}\right) \quad (8)$$

where $f\left(\mathbf{Y}_i^{ps} \mid \boldsymbol{\theta}^{(t+1)}\right)$ is a multivariate skew-normal distribution by which the ML estimates are obtained using an appropriate approach, and for missingness mechanism, i.e. $P\left(M_i \mid \mathbf{Y}_i^{ps}, \boldsymbol{\theta}^{(t+1)}\right)$, the ML estimates can also be obtained by a GLM with the probit link procedure. When we use the SEM algorithm, it is needed to check the convergence of the resulting chain. Several methods

have been proposed in the literature. We will use the Gelman-Rubin method (Gelman and Rubin, 1992). Based on this method, multiple, $k \geq 2$, chains are generated in parallel for $n = 2pt$ iterations. For each chain, this method suggests starting from different points for which the starting distribution is over-dispersed compared to the target distribution. This method separately monitors the convergence of each scalar parameter of interest by evaluating the Potential Scale Reduction Factor, (PSRF), $\sqrt{\hat{R}}$ as

$$\sqrt{\hat{R}} = \sqrt{\frac{n-1}{n} + \frac{1}{n} \frac{B}{W}} \quad (9)$$

where B/n is the between sequence variance and W is the average of within sequence variances. The convergence is achieved if the PSRF is close to one.

4. Simulation Study

In this section, the usefulness of the proposed methodology has been evaluated using a simulation study where we compare its performance with that of the ordinary EM algorithm. As we will generate data by a MNAR mechanism and the EM algorithm can only be used on the assumption of MAR, one expects to see the lack of fit of the ordinary EM algorithm. For this purpose, a sample of size 250 was used, generated from a bivariate skew-normal distribution with the following parameters:

$$\boldsymbol{\lambda} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 4.5 \\ 4.5 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 4.5 \\ 4.5 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$$

We consider the following missing mechanism for generating incomplete data set:

$$p(M_{i2} = 1 | \mathbf{y}_i, \boldsymbol{\Psi}) = \Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}),$$

where $\psi_0 = 0.4$, $\psi_1 = -0.9$ and $\psi_2 = 1.5$ are chosen such that a mechanism with an expected missing rate of 30% can be produced. Results of using SEM and EM algorithms (see Baghfalaki and Ganjali, 2011, for using the EM algorithm) are given in tables 1 and 2, respectively. In our simulation, we estimate the above parameters through the SEM algorithm for 500 times, and considered the mean of the parameter estimates as the final parameter estimates. The standard errors are computed using Bootstrap.

Table 1: Results of simulation study using SEM algorithm for a sample size of 250 where data are simulated from a bivariate skew-normal distribution under MNAR, (Abs. B: absolute bias of the estimate).

Parameter	True	Estimate	Std. E	Abs. B
μ_1	0.000	0.015	0.084	0.015
μ_2	0.000	0.009	0.078	0.009
λ_1	4.500	4.738	1.304	0.238
λ_2	4.500	4.849	1.386	0.349
σ_{11}	1.000	0.976	0.143	0.024
σ_{12}	0.600	0.571	0.113	0.029
σ_{22}	1.000	0.984	0.144	0.016

Table 2: Results of simulation study using the EM algorithm for a sample size of 250 where data are simulated from a bivariate skew-normal distribution under MNAR, (Abs. B: absolute bias of the estimate).

Parameter	True	Estimate	Std. E	Abs. B
μ_1	0.000	0.258	0.047	0.258
μ_2	0.000	0.690	0.060	0.690
λ_1	4.500	3.081	0.136	1.419
λ_2	4.500	2.161	0.082	2.339
σ_{11}	1.000	0.967	0.051	0.033
σ_{12}	0.600	0.131	0.027	0.469
σ_{22}	0.000	0.258	0.047	0.258

The criterion used for comparison is the absolute bias of the estimates. A closer examination of absolute bias shows that the SEM estimates result in smaller absolute bias as compared to the EM algorithm. As results of table 2 show, the EM algorithm is not the best approach to be used for data with MNAR.

5. Application

As an application, we shall make use of a the well-known data set called the Mastitis data. These data, concerning the occurrence of the infectious disease called Mastitis in dairy cows, was introduced in Diggle and Kenward (1994). Data were available of the milk yields of 107 dairy cows from a single herd in two consecutive years. In the first year all animals were safe, in the next year 27 became infected. Mastitis typically reduces milk yield and these are considered as missing data. In addition, Molenberghs et al. (2001) and Crouchley and Ganjali (2002) found 3 outliers (cows 4, 5 and 66) in these data. Using bivariate skew-normal and bivariate normal

distributions, we analyze the Mastitis data in two situations, the first with full data, and the second without outliers. Parameter estimates and their standard errors (computed by Bootstrap method) are given in tables 3 and 4. The $\sqrt{\hat{R}}$'s, (PSRF's), have been calculated for all parameters. Minimum and maximum of these values are .986 and 1.073 respectively, which means the generated sequences have been converged properly. Also we give the contour plots of bivariate skew-normal and bivariate normal distributions in both situations of whole data and data without outliers in figures 1 and 2, respectively. It is obvious in both with and without outliers that the fit of Skew-Normal case is better than that of Normal case. Also we test $\lambda = 0$ *ver.* $\lambda \neq 0$ to check the usefulness of SN distribution for this application.

From these results, we see that using a bivariate skew-normal distribution for these data, missingness is ignorable and the mechanism which is obtained from our study is MCAR. This is not obtained by using normal one (Diggle and Kenward, 1994). These different results can be a consequence of the existence of a significant skewness parameter. This result (ignorable missingness under bivariate skew-normal model) is obtained in both situations where outliers are, or are not considered as a part of the data, but when we deleted the outliers from the study, the estimates of ψ_1 and ψ_2 are closer to zero by the skew-normal model.

Table 3: Parameter estimates and their standard errors under MNAR mechanism for analyzing the whole mastitis data using skew-normal and normal assumptions.

Parameter	Bivariate skew-normal		Bivariate normal model	
	Estimate	S.D.	Estimate	S.D.
μ_1	5.971	0.248	5.765	0.000
μ_2	5.532	0.320	6.146	0.008
λ_1	-0.901	0.425	0.000	-
λ_2	1.706	0.774	0.000	-
σ_{11}	0.971	0.109	0.867	0.003
σ_{12}	0.502	0.177	0.525	0.031
σ_{22}	2.060	0.545	1.447	0.144
ψ_0	-0.494	1.227	0.775	0.715
ψ_1	0.840	0.556	1.052	0.294
ψ_2	-0.875	0.736	-1.206	0.407
$-2\log L$	618.598		626.748	

Table 4: Parameter estimates and their standard errors under MNAR mechanism for analyzing the mastitis data without outliers using skew-normal and normal assumptions.

Parameter	Bivariate skew-normal		Bivariate normal model	
	Estimate	S.D	Estimate	S.D
μ_1	5.181	0.074	5.798	0.000
μ_2	6.566	0.183	6.339	0.053
λ_1	2.726	0.638	0.000	-
λ_2	-1.419	0.525	0.000	-
σ_{11}	1.147	0.088	0.761	0.000
σ_{12}	0.548	0.161	0.576	0.030
σ_{22}	1.280	0.332	0.949	0.036
ψ_0	-1.374	1.290	-0.980	0.535
ψ_1	-0.067	0.754	0.320	0.183
ψ_2	0.002	0.888	-0.430	0.246
$-2\log L$	611.110		617.448	

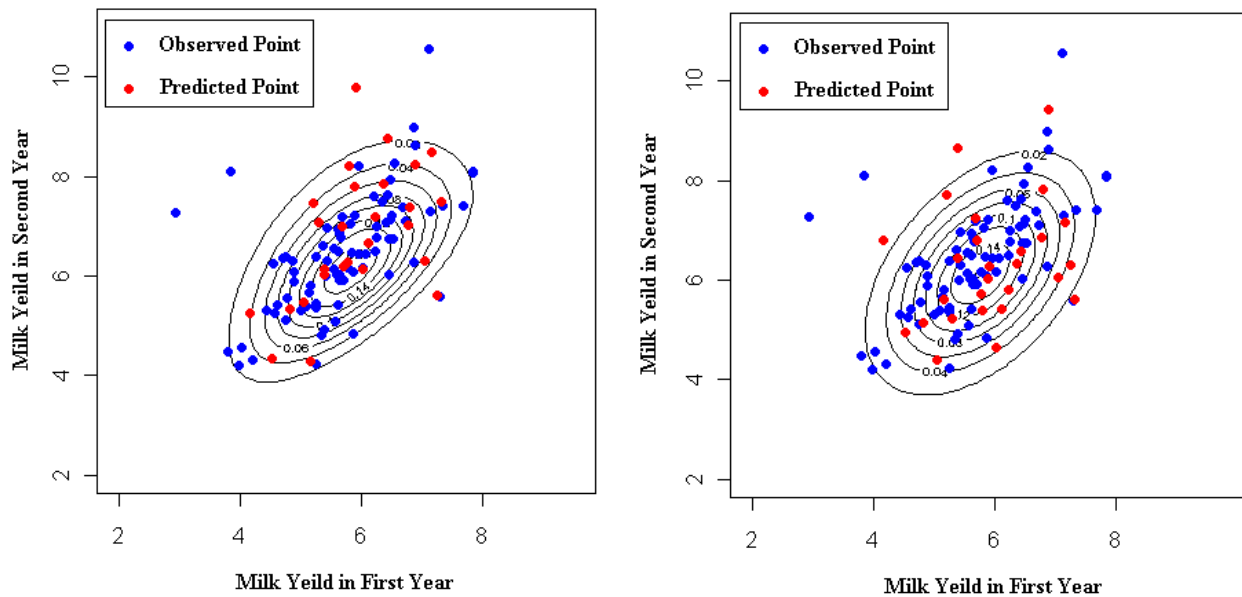


Figure 1: Superimposed “scatter plot of the milk yield in first year versus that of the second year” and “contour plot” of the whole Mastitis data. The left panel is due to bivariate skew-normal and the right panel is due to bivariate Normal model.

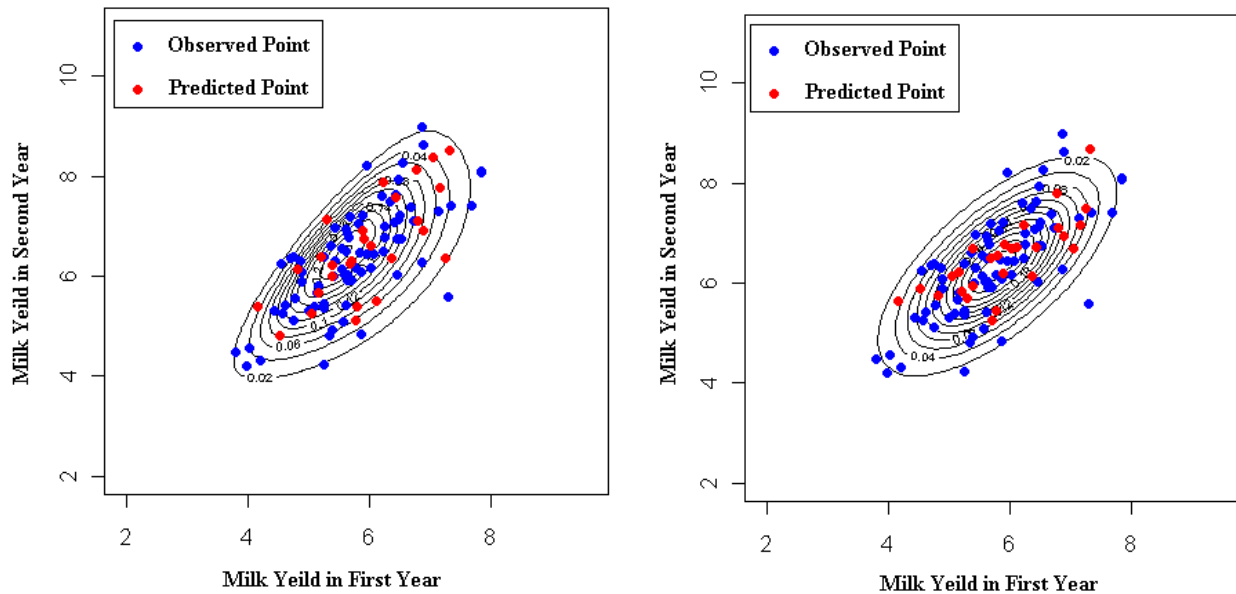


Figure 2: Superimposed “scatter plot of the milk yield in first year versus that of the second year” and “contour plot” of the whole Mastitis data. The left panel is due to bivariate skew-normal and the right panel is due to bivariate Normal model.

Comparing two tables and results, (with and without outliers), we see that the signs of skewness parameters have been changed. The outliers points are (2.93, 7.28), (3.84, 8.10) and (7.11, 10.57). If we focus on outlier values, we see, the first elements are lower than mean value of the first year, and the second ones are greater than the mean value of the second year, which, after deleting them from analysis, can change the sign of skewness parameter, from negative to positive in first year, and positive to negative in second year. Also the expectation of \mathbf{Y} , if $\mathbf{Y} \sim SN_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, is given by

$$E(\mathbf{Y}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \times \boldsymbol{\delta} \sqrt{\frac{2}{\pi}}$$

(Arellano- Valle and Genton, 2005). So, when we use the whole data, $E(\mathbf{Y}) = (5.737, 6.208)'$, and when we use data without outliers, $E(\mathbf{Y}) = (5.755, 6.210)'$. As missing data are CAR based on joint model of bivariate skew-normal and missing mechanism, one may ignore missing mechanism and use a complete case (80 cases) analysis to find unbiased estimates of parameters. Table 5 gives results of such analysis. These results also show a significant skew parameters which change sign in analyzing the whole data and the data without outliers. Due to using fewer data points, the standard errors of the estimates are larger than those using the whole data.

Table 5: Parameter estimates using complete case analyzing of the mastitis data under skew-normal assumption using data with and without outliers.

Parameter	Whole data		Without outliers	
	Estimate	S.E.	Estimate	S.E.
μ_1	5.927	0.280	5.069	0.104
μ_2	5.726	0.309	6.375	0.218
λ_1	-1.051	0.556	2.722	0.840
λ_2	1.724	0.653	-0.941	0.655
σ_{11}	0.960	0.107	1.237	0.111
σ_{12}	0.491	0.156	0.660	0.217
σ_{22}	1.833	0.491	1.096	0.340

6. Conclusion

In this paper a stochastic version of the EM algorithm (SEM) was used to analyze for intermittent missing response data. SEM, previously was used for longitudinal response data. In this paper, we extended the use of SEM for analyzing data with multivariate skew-normal responses. We conducted a simulation study. We also used a selection model framework to reanalyze mastitis data, using a bivariate skew-normal response.

For these data in both cases (with and without outliers), we rejected symmetry ($\lambda = \mathbf{0}$ versus $\lambda \neq \mathbf{0}$ with p -value = 0.016 for data including outliers and p -value = 0.042 for data without outliers). This emphasized the importance of using skew-normal distribution. Considering the skewness nature of the process generating the data we found an ignorable missing data mechanism for the mastitis data.

REFERENCES

- Arellano-Valle, R. B., Del Pino, G. San Martin, E. (2002). Definition and probabilistic properties of skew-distributions. *Statist. Probab. Lett.* 58: 111-121.
- Arellano-Valle, R. B., Gentone, M.G. (2005). On fundamental skew distributions. *J. Multivariate Anal.* 96: 93-116.
- Arellno-Valle, R. B., Bolfaine, H., and Lachos, V. H. (2005). Skew-normal linear mixed models. *J. Data Sci.* 3: 415-438.
- Azzalini, A (1985). A class of distribution which includes the normal ones. *Scand. J statist.* 12: 171-178.
- Azzalini, A., Capitano, A. (1999). Statistical applications of the multivariate skew-normal distributions. *J. Roy. Statist. Sco. B:* 579-602.
- Azzalini, A., Dalla-Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika.* 83: 715-726.
- Baghfalaki, T and Ganjali, M. (2011). An EM estimation approach for analyzing bivariate skew-normal data with non-monotone missing values, *Communications in Statistics-Theory Methods*, 40(9), 1671 - 1686.

- Celux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problems. *Computational Statistics*. 2: 73-82.
- Crouchley, R. and Ganjali, M. (2002). The common structure of several recent statistical models for dropout in repeated continuous responses, *Stat. Modelling*. 2: 39-62.
- Demirtas, H. (2007). Practical advice on how to impute continuous data when the ultimate interest centers on dichotomized outcomes through pre-specified thresholds. *Communications in Statistics-Simulation and Computation*. 36: 871-889.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*. 39: 1-38.
- Delyon, b., Lavielle, M., Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* 27: 94-128.
- Diebolt, J., Ip, E. H. S. (1996). Stochastic EM: method and application. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.(Chapter 15).
- Diggle, P. J. and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Appl. Stat.* 43: 49-94.
- Gelman, A and Rubin, DB. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*. 7: 457-511.
- Gad, A. M. and Ahmed, A.S. (2006). Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm. *Computational Statistics and Data Analysis*. 50: 2702-2714.
- Henze, N.A. (1986). A probabilistic representation of skew-normal distribution. *Scand. J. stat.* 13: 271-275.
- Ip, E.H.S. (1994). A stochastic EM estimator in the presence of missing data: Theory and applications. PhD thesis, Division of biostatistics, Stanford University, California, USA.
- Jank, W. (2005). Stochastic variants of EM: Monte Carlo, quasi Monte Carlo and More. In the 2005 proceedings of the American Statistical Association, August 6-11, Min-neapolis, Minnesota.
- Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *J. Amer. Stat. Associate*. 88: 125-134.
- Little, R.J.A. (1994). A class of pattern-mixture model for normal incomplete data. *Biometrika*. 81: 471-483.
- Little, R.J.A., Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley, New York.
- Melachlan, J, Karishnan. T (1997). *The EM Algorihm and Extensions*. New York: Wiley.
- Molenberghs G., Verbeke, G., Thijs, H., Lesaffre, E. and Kenward, M. G. (2001). Mastitis in dairy cattle: influence analysis to assess sensitivity of the dropout process. *Comput. Stat.*, 37: 93-113.
- Robert, C. and Casella, G. (2002). *Monte Carlo statistical methods*. Springer-Verlag.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*. 63: 581-592.
- Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.* 11: 95-103.
- Wu, M.C. and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44: 175-88.
- Zhu, H.T., Lee, S.Y. (2002). Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte Carlo method. *Statist. Comput.* 12: 175-183.