A Review of EM algorithm for Missing Data

UIC BioStats, Pei-Shan Yen

May 6, 2020

Abstract

Expectation-Maximization algorithm (EM) is a numerical approximation method that has mainly been used to iteratively find the maximum likelihood estimation (MLE) of the parameter (Dempster et al., 1977). The iteration of EM begins with the execution of the Expectation step (E-step), followed by the Maximization step (M-step). The iteration continues until the convergence is attained. In missing data problems, the E-step serves to find the integral of the expected full likelihood conditional on observed data under Missing at Random (MAR) assumption. Naturally, the integral of the E-step cannot be carried out analytically. To resolve this issue, Monte Carlo EM (Wei and Tanner, 1990) and Stochastic Approximation of EM (Delyon et al. 1999), which employ simulation strategies via Markov Chain Monte Carlo methods, are proposed. In this report, we briefly review the histories and the algorithms of EM, MCEM, and SAEM.

keywords: Expectation-Maximization algorithm, MCEM, SAEM

1 Introduction

The maximum likelihood estimation (MLE) of incomplete data problems usually has to be computed through the iterative process. Many numerical approximation methods, such as the Newton-Raphson method (NR), can tackle this issue. However, NR needs extensive computational time to deal with complex likelihood. To resolve this issue, several new methods are advanced. Due to its high efficiency, Expectation-Maximization algorithm (EM) is a particular means of iteratively computing the MLE when the Missing at Random (MAR) assumption holds

Dempster et al. (1977) first proposes EM and showed its monotone behavior of the likelihood and convergence of the algorithm. In the EM framework, the iteration begins with the execution of the Expectation step (E-step), followed by the Maximization step (M-step). The iteration continues until the user-specific tolerance is satisfied (i.e., the convergence is attained). The EM algorithm has mainly been applied not only to missing data research fields but also to a variety of statistical problems, such as mixture distributions or truncated distributions.

In order to implement the E-step in the EM algorithm, the expectation of full data likelihood conditional on observed data, and the current estimate of the parameter must be calculated. Naturally, the integral of the E-step cannot be carried analytically. Because of this hurdle, several modifications and the extended versions of the algorithm have been proposed, such as Monte Carlo EM (Wei and Tanner, 1990), SAEM (Delyon et al. 1999). The extension of the EM algorithm usually employs simulation strategies via Markov Chain Monte Carlo methods (MCMC) to derive the target distribution. For example, the Metropolis-Hastings Algorithm is typically used when producing samples from a complicated distribution that includes high-dimensional target posterior distribution or hierarchical mixed-effect models.

2 Expectation-Maximization algorithm (EM)

2.1 EM Introduction

Let W be the underlying sample space, and $w \in R^n$ is a full data observation of W. Let X be the observed data sample space, and $x \in R^m$ with $m < n$ is an observation of X. Naturally, we cannot directly observe the full dataset in W, but only through the observed data x where $x = x(w)$ is a many-to-one mapping from W to X. In other words, we only observe the incomplete dataset x in X. This observed x determines a subset $W(x)$ in underlying sample space W, which is the inverse mapping of x. The illustration of many-to-one mapping is demonstrated in Figure 1.

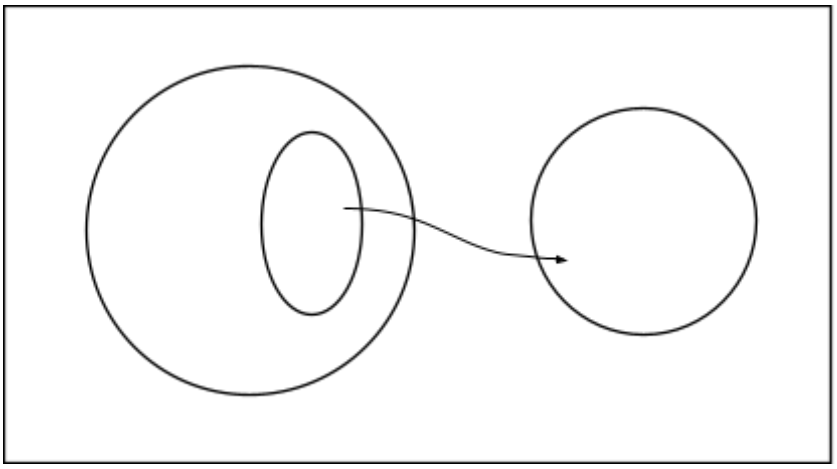

Figure 1. The illustration of many-to-one mapping

We can denote the unobservable full data log-likelihood as

$$logL_F(\theta) = log\left[p_F(\theta)\right].$$

, and the density of the observed data x is

$$p(\theta) = \int_{W(x)} p_F(\theta) dw.$$

E-Step.

To find the estimation of parameter θ, Dempster et al. (1977) suggests that the conditional expectation for the full data log-likelihood.

$$Q\left(\theta, \theta^{(k)}\right) = E\left\{logL_F(\theta)|\theta^{(k)}, x\right\} = E\left\{log\left[p_F(\theta)\right]|\theta^{(k)}, x\right\}$$

should be calculated.

M-Step.

Next, we define the observed data log-likelihood

$$logL(\theta) = \log log\,[p(\theta)] \;=\; Q\left(\theta, \theta^{(k)}\right) - E\left\{log[k(\theta, x)]|\theta^{(k)}, x\right\} = Q\left(\theta, \theta^{(k)}\right) - H\left(\theta, \theta^{(k)}\right)$$

The observed data log-likelihood is not decreased after an EM iteration due to

$$logL\left(\theta^{(k+1)}\right) - logL\left(\theta^{(k)}\right) = \left\{Q\left(\theta^{(k+1)}, \theta^{(k)}\right) - Q\left(\theta^{(k)}, \theta^{(k)}\right)\right\} - \left\{H\left(\theta^{(k+1)}, \theta^{(k)}\right) - H\left(\theta^{(k)}, \theta^{(k)}\right)\right\} \geq 0$$

The above inequality holds for two reasons. First, the first difference in Q function is nonnegative for any θ. Second, based on Jensen's Inequality and the concavity of the logarithmic function, the second difference in H function is nonpositive. Based on the behavior of the monotone observed data log-likelihood, we only need to maximize $Q\left(\theta, \theta^{(k)}\right)$ in the M-step.

The E-step and M-step are repeatedly executed until the sequence of likelihood values converges to a local maximum that depends on the initial starting value $\theta^{(0)}$. Note that the rate of convergence is proportional to the maximal fraction of the unobservable data information. This implies that the convergence speed of the EM algorithm may be slow when a vast amount of missing data is present.

2.2 EM application on Exponential Family

McLachlan (2008) indicates that the EM algorithm can be simplified if the unobservable full data density is in the exponential family. Assuming

$$p\,(w|\theta) = b(w)exp[t(w)c(\theta) - a(\theta)].$$

Where $t(w)$ is the sufficient statistic of the family which can provides all of the necessary information to estimate the parameter θ from the data. The E-step can be written as

$$Q\left(\theta, \theta^{(k)}\right) = E\left\{log[b(w)]|\theta^{(k)}, x\right\} + c(\theta)E\left[t(w)|\theta^{(k)}, x\right] - a(\theta)$$

Since $E\left\{log[b(w)]|\theta^{(k)}, x\right\}$ does not depend on θ, the E-step can be simplified to $t^{k+1} = E\left[t(w)|\theta^{(k)}, x\right]$. In the M-step, we will maximize $c(\theta)t^{k+1} - a(\theta)$.

2.3 EM application on Missing Covariate

In GLM, Iteratively Reweighted Least Squares (IRLS) is an iterative procedure for finding the MLE of the parameter. Each iteration is a weighted least-squares procedure with the weights changing with the iterations. Dempster (1980) shows that the IRLS procedure is an EM algorithm under distributional assumptions.

Ibrahim (1990) states that the EM algorithm can be expressed in terms of a weighted complete data log-likelihood for any GLM with missing covariates. Define *Y* as the outcome

and $X = (X^{obs}, X^{mis})$ which are the covariates. Assuming that the density of $Y|X$ belongs to the exponential family. The complete data log-likelihood is

$$l(\theta|y,x) = l(\beta,\alpha|y,x) = \sum_{i=1}^{n} l\left(\beta,\alpha|y_i,x_i\right) = \sum_{i=1}^{n} log\left[p(y_i|x_i,\beta)\right] + log\left[p(x_i|\alpha)\right].$$

The Q function for the i^{th} observation is

$$Q_i\left(\theta,\theta^{(k)}\right) = E\left\{log\left[p\left(y_i,x_i\right)\right]|\theta^{(k)},y_i,x_i^{obs}\right\} = \sum_{x_i^{mis}} p\left(\theta^{(k)},y_i,x_i^{obs}\right) l\left(\theta|y_i,x_i\right).$$

The Q function of the E-step for all of the observations is given by

$$Q\left(\theta,\theta^{(k)}\right) = \sum_{i=1}^{n}\left\{Q_i\left(\theta,\theta^{(k)}\right)\right\} = \sum_{i=1}^{n}\sum_{x_i^{mis}} p(x_i^{mis}|\theta^{(k)},y_i,x_i^{obs}) l\left(\theta|y_i,x_i\right) = \sum_{i=1}^{n}\sum_{x_i^{mis}}\left[w_i\right] l\left(\theta|y_i,x_i\right).$$

Therefore, the EM algorithm for GLM with missing covariates can be expressed by a weighted complete data log-likelihood. The weight w_i corresponds to the incomplete observations.

2.1.4 Covariance Estimation by EM

In the context of the variance estimation, Louis (1982) presents the observed information matrix within the EM. Assuming that the full dataset X consists of observed dataset X^{obs} and the missing dataset X^{mis}. The density of observed data is

$$p(\theta) = \int_E f(\theta)dx \text{ where } E = \left\{x : x^{mis}(x) = x^{mis}\right\}.$$

The first derivative of the observed data log-likelohood can be denoted as

$$\frac{\partial log[f(\theta)]}{\partial\theta} = \frac{\int_E f'(\theta)dx}{\int_E f(\theta)dx} = \frac{\int_E \frac{f'(\theta)}{f(\theta)} f(\theta)dx}{\int_E f(\theta)dx} = E\left(\frac{\partial log[f(\theta)]}{\partial\theta}|x^{obs},\theta\right) = S^*\left(x^{obs},\theta\right).$$

Hence, the observed data information $I(\theta)$ can be written as

$$\frac{\partial^2 log[f(\theta)]}{\partial\theta^2} = \frac{\int_E f''(\theta)dx}{\int_E f(\theta)dx} - \left[\frac{\int_E f'(\theta)dx}{\int_E f(\theta)dx}\right]\left[\frac{\int_E f'(\theta)dx}{\int_E f(\theta)dx}\right]^T = \frac{\int_E \frac{f''(\theta)}{f(\theta)} f(\theta)dx}{\int_E f(\theta)dx} - \left[S^*\left(x^{obs},\theta\right)\right]\left[S^*\left(x^{obs},\theta\right)\right]^T = E\left(\frac{f''(\theta)}{f(\theta)}|y,\theta\right) -$$

The asymptotic covariance estimation of $\hat{\theta}$ is $\left[I(\hat{\theta})\right]^{-1}$.

3 Monte Carlo EM algorithm (MCEM)

3.1 MCEM Introduction

It is expected that under some conditions, the E-step of the EM algorithm is complex and cannot be tracked analytically. Under this condition, Wei and Tanner (1990) suggest that we can employ Gibbs Sampling and Monte Carlo Integration to approximate the integral of the E-step. The procedure is described as follows.

We augment the observed data X by a latent data Z, which is the missing data. In addition, we define the E-step as the conditional expectation of the augmented log posterior. Based on the concept of Gibbs Sampling from the Method of Monte Carlo, the parameter sequence $\{\theta^{(1)}, \theta^{(2)}, ..., \theta^{(k)}\}$ will converge to $\hat{\theta}$.

Imputation Step.

Assuming that the Q function is $Q(\theta, \theta^{(k)}) = \int_Z log[p(\theta|z,x)]p(z|\theta^{(k)}, x)dz$. Given $z^{(k)} = (z^{(k1)}, z^{(k2)}, ..., z^{(km_k)})$, then calculate $log[p(\theta|z^{(k)}, x)]$.

Posterior Step.

As $m \to \infty$, we replace the conditional expectation of the full data log-likelihood in the E-step by Monte Carlo Integration. Namely,

$$Q(\theta, \theta^{(k)}) = \sum_{j=1}^{m_k} \frac{1}{m_k}\{log[p(\theta|z^{(k)}, x)]\} \to \int log[p(\theta|z^{(k)}, x)]dz = E\{log[L_F(\theta|z,x)]\}$$

The M-step is to maximize the above Q function. The iteration of this algorithm continues until the convergence of the parameter estimation is attained. In MCEM, Monte Carlo error is included in the E-step and the monotone likelihood characteristic is lost. Wei and Tanner (1990) suggest that we can choose small value of m in the initial stage and then increase m when this algorithm moves toward convergence. The plot of $\theta^{(k)}$ can help to monitor the convergence.

3.2 MCEM application on Missing Mixed Covariate for arbitrary regression model

Ibrahim (1999) employs MCEM on arbitrary parametric regression models with missing mixed (i.e., continuous or categorial) covariates. To model the relationship between the

outcome Y and the covariates $X = (X_1, X_2, ...X_p)$, the conditional density function of the model $Y|X$ is $p(\beta, x)$. Assuming that $(x_1, y_1), ..., (x_n, y_n)$ are independent observations. Each x_i has $p\times1$ random vecotr of covariates, for $i = 1, 2, ..., n$.

We define the full dataset of the covariates as $X_i = (X_i^{obs}, X_i^{mis})$ and X_i^{mis} is a $q_i\times1$ vector. Assuming that the covariates have missing data in the observed dataset. Since we cannot observe X_i^{mis}, a latent variable $Z = X_i^{mis}$ should be assumed. In addition, we define the marginal distribution of X as $p(\alpha)$. From Bayes Theorem, the likelihood for each observation can be expressed as

$$p(\beta, \alpha) = p(y_i|x_i, \beta)p(x_i|\alpha).$$

Hence, the full data log-likelihood is

$$l(\theta|y, x) = \sum_{i=1}^{n} log\Big[p(y_i|x_i, \beta)\Big] + log\Big[p(x_i|\alpha)\Big].$$

Then, we can employ MCEM for the E-step. Given $\theta^{(k)}$, we can draw sample $x_i^{mis(k)} = z^{(k)} = (z^{(k1)}, z^{(k2)}, ..., z^{(km_k)})$ from target distribution $p\left(x_i^{mis}|\theta^{(k)}, x_i^{obs}, y_i\right)$.

Defining the Q function for the i^{th} observation as

$$Q_i\left(\theta, \theta^{(k)}\right) = E\Big\{log\Big[p\left(y_i, x_i\right)\Big]|\theta^{(k)}, y_i, x_i^{obs}\Big\} = E\Big\{log\Big[p\left(y_i, x_i^{obs}, x_i^{mis}\right)\Big]|\theta^{(k)}, y_i, x_i^{obs}\Big\}$$

$$= \int log\Big[p(y_i|x_i, \beta)\Big]\, p\left(x_i^{mis}|\theta^{(k)}, x_i^{obs}, y_i\right)dx_i^{mis} + \int log\Big[p(x_i|\alpha)\Big]\, p\left(x_i^{mis}|\theta^{(k)}, x_i^{obs}, y_i\right)dx_i^{mis}$$

where we can replace $p\left(x_i^{mis}|\theta^{(k)}, x_i^{obs}, y_i\right)$ by $p\left(y_i|x_i, \beta^{(k)}\right)p(x_i|\alpha^{(k)})$.

The Q function of the E-step for all of the observations is given as

$$Q\left(\theta, \theta^{(k)}\right) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{m_k}\frac{1}{m_k}log\left[p\left(\theta|x_i^{mis^{(kj)}}, x_i^{obs}, y_i\right)\right]\right\}.$$

The M-step will maximize Q function. We continue to update the EM step until we find the convergence of the parameter estimation.

3.3 Covariance Estimation by MCEM

Based on the work of Wei and Tanner (1990), we can use Monte Carlo Integration to obtain the observed data information $I(\theta)$.

$$\sum_{i=1}^{n}\left\{-\sum_{j=1}^{m_k}\frac{1}{m_k}\frac{\partial^2 log\left[p\left(y,z^{(j)}\right)\right]}{\partial\theta^2}-\sum_{j=1}^{m_k}\frac{1}{m_k}\left\{\frac{\partial log\left[p\left(y,z^{(j)}\right)\right]}{\partial\theta}\right\}^2+\left\{\sum_{j=1}^{m_k}\frac{1}{m_k}\frac{\partial log\left[p\left(y,z^{(j)}\right)\right]}{\partial\theta}\right\}^2\right\}$$

$$\rightarrow -\int_z \frac{\partial^2 log[p(y,z)]}{\partial\theta^2}p(y,\theta)dz-\int_z\left\{\frac{\partial log[p(y,z)]}{\partial\theta}\right\}^2 p(y,\theta)dz+\left\{\int_z\frac{\partial log[p(y,z)]}{\partial\theta}p(y,\theta)dz\right\}^2$$

Therefore, the asymptotic covariance estimation of $\hat{\theta}$ is $\left[I(\hat{\theta})\right]^{-1}$.

4 Stochastic Approximation of EM (SAEM)

4.1 SAEM Introduction

However, an accurate Monte Carlo approximation in the E-step in MCEM is computational extensive. Delyon et al. (1999) propose a stochastic method for MCEM, which is known as SAEM, to overcome the convergence issue. This algorithm replaces the E-step by a stochastic approximation from drawing a single sample of z. The algorithm consists of the following steps:

Simulation Step

Given initial $\theta^{(0)}$, draw a single sample of $x_i^{mis^{(k)}} = z_i^{(k)}$ from $p\left(z_i|\theta^{(k-1)}, x_i^{obs}\right)$.

Stochastic E-Step

Defining the Q function as

$$Q\left(\theta,\theta^{(k)}\right) = Q\left(\theta,\theta^{(k-1)}\right)+\gamma_k\left\{\log log\left[p\left(\theta|z^{(k)},x^{obs}\right)\right]-Q\left(\theta,\theta^{(k-1)}\right)\right\},$$

where $\{\gamma_k\}$ is a decreasing sequence of positive numbers.

Maximization Step

We will obtain MLE of the parameter by maximizing the above Q function. The iteration of this algorithm continues until we find the convergence of the parameter estimation.

If the sample size of observed dataset is small, we can improve the convergence of SAEM by creating M Markov chains for each observation instead of single sample. That is, we have to draw M sequences of $(z^{(k1)}, z^{(k2)}, ..., z^{(kM)})$ and update the Q function by

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left\{ \frac{1}{M} \sum_{j=1}^{M} \log log \left[p\left(\theta | z^{(kj)}, x\right) \right] - Q\left(\theta, \theta^{(k-1)}\right) \right\}$$

If we let $\gamma_k = \frac{1}{k}$ for all k, the sequence $\left\{\theta^{(k)}\right\}$ converges almost surely to $\hat{\theta}$ with a very slow speed. To improve the speed of the algorithm, we may choose $\gamma_k = 1$ for all k. This setting makes the sequence $\left\{\theta^{(k)}\right\}$ converges almost surely (i.e., with probability 1) to the maximum likelihood estimate of θ with a quick convergence speed. Under this condition, SAEM is simplified and is equivalent of Stochastic EM (Celeux and Diebolt, 1985). In Stochastic EM algorithm, we first simulate $z^{(k)}$ from $p\left(z|\theta^{(k-1)}, x\right)$, then maximizing $p\left(\theta|z^{(k)}, x\right)$ to obtain MLE of θ. Likewise, when we let $\gamma_k = 1$ and $M \gg 1$ for all k, SAEM is simplified and is equivalent to MCEM. Normally, the convergence speed of MCEM is slower than that of Stochastic EM and SAEM.

Actually, the choice of the $\left\{\gamma_k\right\}$ determine the convergence speed. The convergence of the sequence $\left\{\gamma_k\right\}$ needs to satisfy $\sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$. This paper recommend that we should choose $\gamma_1 = 1$ in the first iteration and $\gamma_k = \frac{1}{k}$ in the following iterations. This setting makes the sequence $\left\{\theta^{(k)}\right\}$ converges almost surely with a quick convergence speed.

4.2 SAEM application on Exponential Family

When the posterior distribution $p\,(\theta|z, x)$ belongs to exponential family, we can simplify the implementation of SAEM. That is,

$$p\,(\theta|z, x) = b(z, x) exp[t(z, x) c(\theta) - a(\theta)]$$

where $t(z, x)$ is sufficient statistic. After drawing $z^{(k)}$ from $p\left(z|\theta^{(k-1)}, x\right)$, in the stochastic approximation step, we update the Q function by $t_k = t_{k-1} + \gamma_k \left\{ t\left(z^{(k)}, x\right) - t_{k-1} \right\}$ for any $t \in t(z, x)$. Then we maximize s_k to obtain the MLE of the parameter. The iteration continues until the convergence of the parameter estimation is reached.

4.3 SAEM application on Missing Covariate

Jiang et al. (2020) propose a new version of SAEM to estimate the parameter of the logistic model with missing covariates. Assuming that the outcome Y is a binary variable and the covariates $X = (X_1, X_2, ...X_p)$. The logistic model can be given as

$$p(x) = \frac{exp(\beta^t X)}{1 + exp(\beta^t X)}$$

Assuming that X follows a multivariate normal distribution with mean μ and covariance Σ, and let the parameter $\theta = (\beta, \mu, \Sigma)$. The full data log-likelihood is

$$l(\theta|y,x) = l(\beta,\mu,\Sigma|y,x) = \sum_{i=1}^{n} l\left(\beta,\mu,\Sigma|y_i,x_i\right) = \sum_{i=1}^{n} log\left[p(y_i|x_i,\beta)\right] + log\left[p(x_i|\mu,\Sigma)\right]$$

The target distribution is $p(x^{mis}|\theta, x_i^{obs}, y_i)$. Using Metropolis–Hastings algorithm, we choose the proposal distribution $x_i^{mis}|x_i^{obs}$ which follows a normal distribution with mean u_i and covariance Σ_i. Here, $u_i = u_i^{mis} + \Sigma_i^{mis,obs}\left(\Sigma_i^{obs,obs}\right)^{-1}\left(x_i^{obs} - u_i^{obs}\right)$ and $\Sigma_i = \Sigma_i^{mis,mis} - \Sigma_i^{mis,obs}\left(\Sigma_i^{obs,obs}\right)^{-1}\Sigma_i^{mis,obs}$.

5 Conclusion

As opposed to Newton-Raphson and Fisher's scoring method that find the approximation of MLE, the EM algorithm is easy to be implemented and is computationally efficient when the likelihood is complex. EM is a numerically stable algorithm, and the likelihood increases over the course of the iteration (i.e., EM can always find the monotone convergence of local maximum). The execution of EM begins with guessing an arbitrary initial value of $\theta^{(0)}$. Then the E-step is executed to calculate the conditional expectation for the full data log-likelihood. Lastly, the M-step is performed for maximizing the E-step. The E- and M-steps are iteratively executed until the convergence of MLE is attained.

It is expected that under some conditions, the E-step is hard to track analytically. To resolve this issue, we can use MCEM or SAEM, each of which is based on Monte Carlo approach, to approximate the integral of the E-step. Compared with MCEM, SAEM is a better option when dealing with a complex integral, and this can be summarized to two reasons. The first is that SAEM only draws a single sample from the conditional predictive distribution. The other is that SAEM includes a sequence of decreasing stochastic positive number, adjusting the Q function, which, in turn, accelerates the convergence performance.

Reference

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational statistics quarterly: CSQ, 2(1): 73-82.

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): 1-38.

Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. The Annals of Statistics, 27(1): 94-128.

Ibrahim, J. G. (1990). Incomplete Data in Generalized Linear Models. Journal of the American Statistical Association, 85(411): 765-769.

Ibrahim, J. G., Chen, M.-H. and Lipsitz, S. R. (1999). Monte Carlo EM for Missing Covariates in Parametric Regression Models. Biometrics, 55: 591-596.

Jiang, W., Josse, J., Lavielle, M. and Group, T. (2020). Logistic regression with missing covariates - Parameter estimation, model selection and prediction within a joint-modeling framework. Computational Statistics and Data Analysis, 145, 106907.

Lavielle, M. (2014). Mixed Effects Models for the Population Approach, Models Tasks, Methods and Tools. Chapman and Hall/CRC.

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): 226-233.

McLachlan, G. J. and Krishnan, T. (2008). The EM Algorithm and Extensions. Wiley Johnsons.

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): 699-704.