Convergence properties of the EM algorithm in the mixture model with missing data

N Dwidayati1,2* and Zaenuri1
1Department of Mathematics FMIPA Universitas Negeri Semarang, Indonesia
*Corresponding author: nurkaromah.mat@mail.unnes.ac.id

Abstract. The research aims to investigate the convergence properties of the EM Algorithm in the mixture model with missing data. We have the key property of the EM algorithm is that the incomplete data log-likelihood increases after each iteration of the algorithm. So, the EM sequence of likelihoods must converge if the likelihood is bounded above.

1. Introduction
In this paper, we study a mixture model for analyze cure rate with missing data. Model parameters are estimated via an appropriate EM (Expectation Maximization) algorithm that perform the ML (Maximum Likelihood) in presence of missing data. A two-step iterative approach that estimates the parameters of a model starting from an initial guess. Each iteration consists of two steps: (1) an expectation step that finds the distribution for the missing data based on the known values for the observed variables and the current estimate of the parameters and (2) a maximization step that substitutes the missing data with the expected value.

We will investigate the convergence properties of the EM Algorithm in the mixture model with missing data. We will show that the incomplete data log-likelihood increases after each iteration of the algorithm.

1.1. Mixture Model
Cure models are survival models basically developed to estimate the proportion of patients cured in a clinical trial. These models estimate the cured proportion and also the probability of survival of the uncured patients up to a given point of time. One of the model developed was to estimate the proportion of patients cured among those who were receiving treatment for cancer of mouth and throat, cervix, uteri and breast [1]. This model is called the mixture model since it can estimate the proportion of patients cured and the survival function of the uncured patients.

The mixture model is applicable and useful for both continuous and discrete (probability) distributions. This model is based on representation of the (cumulative) probability distribution function or, more simply, the (discrete or continuous) probability density function \( f(\cdot) \) from with realizations are desired in terms of mixture.

1.2. Definition
Let \( Y = \{Y_1, \ldots, Y_n\} \) denote a random sample of size \( n \), where \( Y_i \) is a vector of \( \mathbb{R}^q \) and \( y_i \) its realization. \( f(y_i) \) its density functions. In the mixture model context, the density of \( Y_i \) is supposed to be a mixture of \( P \) parametric densities such that
\[ f(y_i; \psi) = \sum_{p=1}^{P} \pi_p f_p(y_i; \theta_p) \] (1)

with the constraint \[ \sum_{p=1}^{P} \pi_p = 1, \quad P \text{ being fixed.} \]

Coefficients \( \pi_p \) can be viewed as the weights of the \( p^{th} \) component of the mixture, which is characterized by parameter \( \theta_p \) and \( \psi = (\pi_1, \ldots, \pi_{P-1}, \theta_1, \ldots, \theta_P) \) denotes the vector of parameters of the model.

In incomplete data problem, mixture model is reformulated since the assignment of the observed data is unknown. We note \( X_t = \{Y_t, Z_t\} \) the complete data vector whose only component being observed is \( Y_t \), its density function is then,

\[ g(x_i; \psi) = \prod_{p=1}^{P} [\pi_p f(y_i; \theta_p)]^{z_{i,p}} \] (2)

Given a sample of \( n \) independent observation from a mixture model defined in (1) the likelihood function is

\[ L(y; \psi) = \prod_{i=1}^{n} \left\{ \sum_{p=1}^{P} \pi_p f(y_i; \theta_p) \right\} \] (3)

The mixture model is said to be a parametric mixture cure model when standard probability distributions such as exponential, Weibull, Gompertz and generalized F are used. The mixture model used without any standard probability distribution is called a non-parametric mixture cure model.

The mixture model is one of several fundamental tools used in constructing efficient exact algorithms for generating repeated realizations of random variables on computer. Its key feature is generating with high probability a realization from a simple distribution that is close to the desired one. Mixture models have been used to analyze clinical trials with potentially cured patients [2-7].

These models are useful when a proportion of study subjects never experience the event of interest. Mixture modeling approach is commonly used to formulate a cure model, which assumes that the underlying population is a mixture of susceptible and nonsusceptible subjects. Mixture model have been used to analyze clinical trials with potentially cured patients.

Mixture model have been used to model the failure time data with the existence of the long-term survivors. The mixture model assumes that a fraction of survivors are cured from the disease of interest. The failure time distribution for uncured individuals (latency) can be modeled by either parametric models or a semi-parametric proportional hazards model.

1.3. Missing Data
Incomplete or missing data is a pervasive problem in almost all areas of research. When someone drops out of a research study randomly, if they move out of state or have an accident on the way to the test site, that is seen as data missing “randomly”. But in the case of those dropping out of a study because their condition is so severe, or because they don’t have the condition at all, that is considered “non-random”. That type of missing data, if ignored, can lead to biased conclusions in a study.

Based on terminology of [8] and [9] for referring to missing-value or nonresponse process. Let \( Y \) is data matrix and \( M \) is missing data matrix \( M = \{M_{ij}\} \)

\[ M_{ij} = \begin{cases} 1; \text{if } Y_{ij} \text{ is missing} \\ 0; \text{if } Y_{ij} \text{ is observed} \end{cases} \]
Convenient to introduce the notation \( Y = (Y_{\text{obs}}, Y_{\text{mis}}) \) where \( Y_{\text{mis}} \) are conceptual and denote the data that were not observed.

Missing data is one of the most important statistical and design problems in research. In traditional methods, researchers choose simply deleted participants with missing data from their analyses. But that could skew conclusions, particularly in studies that rely on random samples to draw conclusions about entire populations. Delete more than a few participants because of missing data, and suddenly our samples are no longer random.

The methods proposed by Schafer and Graham recommend can preserve the integrity of data set by using statistics to fill in probable values for missing information, thereby allowing researchers to make more accurate conclusions the population under study.

2. Methods
The population for this research consisted of patient who sought for help to RSUP DrSardjito Yogyakarta because of breast cancer. The sample for this research was breast cancer patient (BCP), taken from the same population based on a five-year survival.

This research was designed using deductive-analytic approach. The data for this research were secondary data collected from clinical studies and medical records of the BCP treated in the Hospital. The medical records consisted of registered numbers, the patient’ name, dates of births, dates of first visits, dates of check outs or dates of last re-visits, disease stadiums, types of treatments and post-treatment health statuses. The data were analyzed deductively based on reviews on the results of previous studies, clinical definitions, clinical assumptions and clinical theorems. Based on the estimator was obtained, this research focus was properties of convergence.

3. Result and Discussion
3.1. EM Algorithm for Analysis with Missing Data.
One of the methods that can be used for analysis with missing data is Expectation Maximization (EM) Algorithm [10]. The EM algorithm is a very general iterative algorithm for maximum likelihood (ML) estimation in incomplete-data problem [11]. Unknown parameters in this mixture model are estimated via an appropriate EM (Expectation Maximization) algorithm that perform the ML (Maximum Likelihood) in presence of missing data. The basic idea of the EM algorithm is to associate a complete data model to the incomplete structure that is observed in order to simplify the computation of maximum likelihood estimates.

The EM algorithm formalizes for handling missing data: (1) replace missing values by estimated values, (2) estimate parameters, (3) reestimate the missing values assuming the new parameter estimates are correct, and (4) reestimate parameters, and so forth, iterating until convergence.

A two-step iterative approach that estimates the parameters of a model starting from an initial guess. Each iteration consists of two steps: (1) an expectation step (E-step) that finds the distribution for the missing data based on the known values for the observed variables and the current estimate of the parameters and (2) a maximization step (M-step) that substitutes the missing data with the expected value.

In the incomplete data formulation of mixture models, let us note \( X \) the complete data sample space from which \( x \) arises, \( Y \) the observed sample space and \( Z \) the hidden sample space. It follows that \( X = Y \times Z \) and \( x = (y, z) \). The density of the observed data \( X \) can be written:

\[
g(x; \psi) = f(y; \psi)k(z | y; \psi)
\]

where \( f(y; \psi) \) is density of the observed data and \( k(z | y; \psi) \) is conditional density of missing observation given the data.

Let \( L(y; \psi) \) is the observed/incomplete-data likelihood and \( L'(x; \psi) \) is the unobserved/complete-data likelihood. These likelihoods are linked with the relationship:
\[
\log L'(x; \psi) = \log L(y; \psi) + \log k(z \mid y; \psi)
\]  
(6)

with
\[
\log L'(x; \psi) = \sum_{i=1}^{n} \log g(x_i; \psi)
\]
(7)

and
\[
\log k(z \mid y; \psi) = \sum_{i=1}^{n} \sum_{p=1}^{P} z_{yp} \log \mathbb{E}(Z_{yp} \mid Y_i = y_i)
\]
(8)

The hidden variables are not observed, so the EM machinery consists of the indirect optimization of the incomplete-data likelihood via the iterative optimization of the conditional expectation of the complete-data likelihood using the current fit for \( \psi \). We note \( \psi^{(h)} \) the value of the parameter at iteration \( h \).

\[
\log L(y; \psi) = Q(\psi; \psi^{(h)}) - H(\psi; \psi^{(b)})
\]
(9)

with conventions:
\[
Q(\psi; \psi^{(b)}) = E_{\psi^{(b)}} \{ \log L(X; \psi) \mid Y \}
\]
(10)

\[
H(\psi; \psi^{(b)}) = E_{\psi^{(b)}} \{ \log k(Z \mid Y; \psi) \mid Y \}
\]
(11)

where \( E_{\psi^{(b)}} \{ \} \) denotes the expectation operator, taking the current fit \( \psi^{(b)} \) for \( \psi \).

The EM algorithm consists of two steps, E-step calculate \( Q(\psi; \psi^{(h)}) \) and M-step choose \( \psi^{(h+1)} = \text{Arg max}_\psi \{ Q(\psi; \psi^{(h)}) \} \). The E-steps and M-steps are repeated alternatively until the difference \( |\psi^{(h+1)} - \psi^{(h)}| \) changes by an arbitrarily small amount.

**3.2. Convergence Properties of Estimator in Mixture Model**

Based on the inference done on the mixture model for cure rate analysis if there is missing data it can be derived theorem, corollary and lemma based on the study of the properties of estimator in the model as follows.

**Theorem 1**

If the likelihood is bounded above then the EM sequence of likelihoods must converge

**Proof**

The proof of this theorem is based on the definition of the M-step that ensures
\[
Q(\psi, \psi^{(h+1)}) \geq Q(\psi, \psi^{(h)})
\]
(12)

While the application of the Jensen inequality gives
\[
H(\psi, \psi^{(h+1)}) \leq H(\psi, \psi^{(h)})
\]
(13)

so with relation (9) these inequalities ensure the monotonicity of the likelihood sequence:
\[
\log L(y; \psi^{(h+1)}) \geq \log L(y; \psi^{(h)})
\]
(14)

These inequality proves that EM sequence of likelihoods must converge if the likelihood is bounded above.

Then formed a sequence of iterations \( \theta^{(0)}, \theta^{(1)}, \ldots \) and define the mapping \( \theta \rightarrow M(\theta) \) from such that \( \theta^{(k+1)} = M(\theta^{(k)}) \) (\( k = 0, 1, 2, \ldots \)).
The function M is called EM mapping. Based on theorem 1 are derived as follows corollary 1 and 2.

**Corollary 1**

If for any given $\theta^*$ on the parameter space of $\theta$

$$\log(\theta^*; Y_{\text{obs}}) \geq \log(\theta; Y_{\text{obs}})$$

for all $\theta$ then for every EM algorithm apply:

$$\log(M(\theta'|Y_{\text{obs}})) = \log(\theta'|Y_{\text{obs}})$$

$$Q(M(\theta')|\theta^*) = Q(\theta' \mid \theta^*)$$

and

$$f(Y_{\text{mis}} \mid Y_{\text{obs}}, M(\theta')) = f(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta^*)$$

almost everywhere

**Corollary 2**

If for any given $\theta^*$ on the parameter space of $\theta$

$$\log(\theta^*; Y_{\text{obs}}) > \log(\theta; Y_{\text{obs}})$$

for all $\theta$ then for every EM algorithm apply:

$$M(\theta') = \theta^*$$

Based on the previous description, the EM algorithm starts from the initial $\psi^{(0)} \in \Omega$ where $\Omega$ is parameter space $\Omega \subset R^d$ where $R^d$ is Euclid space $d$ dimensional, with $d$ declare many parameters in the mixture model, and iteration $i$ $\psi^{(k+1)} = M(\psi^{(k)})$. $\psi, M(\psi)$ is global maximized $Q(\cdot, \psi)$. So $Q(M(\psi)|\psi) > Q(\psi^{(k)}|\psi)$ for all $\psi^{(k)} \in \Omega$.

Maximizer on the M-step $g$ denoted with $M(\cdot)$ called the EM transition function and, assumed that well defined, that is every global maximize of $Q(\cdot, \psi)$ is unique or there is a deterministic procedure to choose $M(\psi)$ from some global maximizer.

Given $i$ $S_g(\psi) = \frac{\partial}{\partial \psi} \log L(\psi; Y_{\text{obs}})$. For all $\psi^{(k)} \in \Omega$, apply

$$\log L(\psi^{(k)}; Y_{\text{obs}}) = Q(\psi | \psi^{(k)}) - H(\psi | \psi^{(k)})$$

with:

$$Q(\psi | \psi^{(k)}) = \int_{-\infty}^\infty \log L(\psi; Y)f(Y_{\text{mis}} \mid Y_{\text{obs}}, \psi^{(k)})dY_{\text{mis}}$$

$$H(\psi | \psi^{(k)}) = \int_{-\infty}^\infty \log f(Y_{\text{mis}} \mid Y_{\text{obs}}, \psi)f(Y_{\text{mis}} \mid Y_{\text{obs}}, \psi^{(k)})dY_{\text{mis}}$$

The set of stationary points of log-likelihood observed data is denoted by $S_k \subset \Omega$ that is $\psi^* \in S_k \Rightarrow S_g(\psi^*) = 0$. If sequence $\left\{\psi^{(k)}\right\}$ convergent to $\psi^*$ then called the convergence point for EM.

Here are some conditions that are assumed to guarantee $M(\psi)$.

A1: $\Omega$ open set in $R^d$, where $R^d$ is Euclid space with dimension $d$

A2: log-likelihood observed data differentiable with continuous derivatives $S_g(\cdot)$

A3: level set $\Omega_{\psi} = \left\{\psi^{(k)} \in \Omega: \log L(\psi^{(k)} | Y_{\text{obs}}) \geq \log L(\psi | Y_{\text{obs}})\right\}$ compact in $R^d$, that is closed and bounded

A4: The distribution of missing data has the same support for all $\psi \in \Omega$

A5: $Q(\psi^{(k)} | \psi)$ continuous inside $\psi^{(k)}$ and $\psi$ differentiable $\psi^{(k)}$
Lemma 1
Let $\psi \in \Omega$. If $Q(\psi)$ have maximize unique global then $M(.)$ continue at $\psi$.

4. Conclusion
The purpose of the EM algorithm is the iterative computation of maximum likelihood estimators when observation can be viewed as incomplete data. The basic idea of the EM algorithm is to associate a complete data model to the incomplete structure that is observed in order to simplify the computation of maximum likelihood estimates.

It has been in previous section that the EM algorithm generates a sequence $(\psi^h)_{h \geq 0}$. The EM sequence of likelihoods must converge if the likelihood is bounded above. So, we have the key property of the EM algorithm is that the incomplete data log-likelihood increases after each iteration of the algorithm.

References
[1] Boag J W 1949 *J. Royal Stat. Soc.* 11 15
[2] Farewell V T 1986 *J. Stat.* 14 57
[3] Cantor A B and Shuster J J 1992 *Stat. Med.* 11 31
[4] Kuk A Y and Chen C 1992 *Biometrika* 79 531
[5] Maller R A and Zhou S 1992 *Biometrika* 79 731
[6] Sy J P and Taylor J M G 2000 *Biometric* 56 27
[7] Peng Y and Dear K G B 2000 *Biometric* 56 37
[8] Rubin D B 1976 *Biometrika* 63 581
[9] Little R J and Rubin D B 2014 *Statistical analysis with missing data* (New Jersey: John Wiley & Sons)
[10] Kazemil 2005 *Methods for Missing Data* (UK: Lancastar University)
[11] Dempster A P, Laird N M and Rubin D B 1977 *J. Royal Stat. Soc.* 39 1