Handling missing data with expectation maximization algorithm

Loc Nguyen
Independent scholar, Vietnam
Email: ng_phloc@yahoo.com
Homepage: www.locnguyen.net

Abstract
Expectation maximization (EM) algorithm is a powerful mathematical tool for estimating parameter of statistical models in case of incomplete data or hidden data. EM assumes that there is a relationship between hidden data and observed data, which can be a joint distribution or a mapping function. Therefore, this implies another implicit relationship between parameter estimation and data imputation. If missing data which contains missing values is considered as hidden data, it is very natural to handle missing data by EM algorithm. Handling missing data is not a new research but this report focuses on the theoretical base with detailed mathematical proofs for fulfilling missing values with EM. Besides, multinormal distribution and multinomial distribution are the two sample statistical models which are concerned to hold missing values.
Keyworks: expectation maximization (EM), missing data, multinormal distribution, multinomial distribution.

1. Introduction to expectation maximization algorithm
Literature of expectation maximization (EM) algorithm in this report is mainly extracted from the preeminent article "Maximum Likelihood from Incomplete Data via the EM Algorithm" by Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin (Dempster, Laird, & Rubin, 1977). For convenience, let DLR be reference to such three authors. The preprint "Tutorial on EM algorithm" (Nguyen, 2020) by Loc Nguyen is also referred in this report.

Now we skim through an introduction of EM algorithm. Suppose there are two spaces $\boldsymbol{X}$ and $\boldsymbol{Y}$, in which $\boldsymbol{X}$ is *hidden space* whereas $\boldsymbol{Y}$ is *observed space*. We do not know $\boldsymbol{X}$ but there is a mapping from $\boldsymbol{X}$ to $\boldsymbol{Y}$ so that we can survey $\boldsymbol{X}$ by observing $\boldsymbol{Y}$. The mapping is many-one function φ: $\boldsymbol{X} \to \boldsymbol{Y}$ and we denote $\varphi^{-1}(Y) = \{X \in \boldsymbol{X}: \varphi(X) = Y\}$ as all $X \in \boldsymbol{X}$ such that $\varphi(X) = Y$. We also denote $\boldsymbol{X}(Y) = \varphi^{-1}(Y)$. Let $f(X \mid \Theta)$ be the probability density function (PDF) of random variable $X \in \boldsymbol{X}$ and let $g(Y \mid \Theta)$ be the PDF of random variable $Y \in \boldsymbol{Y}$. Note, Y is also called observation. Equation 1.1 specifies $g(Y \mid \Theta)$ as integral of $f(X \mid \Theta)$ over $\varphi^{-1}(Y)$.

$$g(Y|\Theta) = \int_{\varphi^{-1}(Y)} f(X|\Theta)\mathrm{d}X \tag{1.1}$$

Where Θ is probabilistic parameter represented as a column vector, $\Theta = (\theta_1, \theta_2, \ldots, \theta_r)^T$ in which each θ_i is a particular parameter. If X and Y are discrete, equation 1.1 is re-written as follows:

$$g(Y|\Theta) = \sum_{X \in \varphi^{-1}(Y)} f(X|\Theta)$$

According to viewpoint of Bayesian statistics, Θ is also random variable. As a convention, let Ω be the domain of Θ such that $\Theta \in \Omega$ and the dimension of Ω is r. For example, normal distribution has two particular parameters such as mean μ and variance σ^2 and so we have $\Theta = (\mu, \sigma^2)^T$. Note that, Θ can degrades into a scalar as $\Theta = \theta$. The conditional PDF of X given Y, denoted $k(X \mid Y, \Theta)$, is specified by equation 1.2.

$$k(X|Y,\Theta) = \frac{f(X|\Theta)}{g(Y|\Theta)} \tag{1.2}$$

According to DLR (Dempster, Laird, & Rubin, 1977, p. 1), $\boldsymbol{X}$ is called *complete data* and the term "incomplete data" implies existence of $\boldsymbol{X}$ and $\boldsymbol{Y}$ where $\boldsymbol{X}$ is not observed directly and $\boldsymbol{X}$ is only known by the many-one mapping φ: $\boldsymbol{X} \rightarrow \boldsymbol{Y}$. In general, we only know $\boldsymbol{Y}$, $f(X \mid \Theta)$, and $k(X \mid Y, \Theta)$ and so our purpose is to estimate Θ based on such $\boldsymbol{Y}$, $f(X \mid \Theta)$, and $k(X \mid Y, \Theta)$. Like MLE approach, EM algorithm also maximizes the likelihood function to estimate Θ but the likelihood function in EM concerns $\boldsymbol{Y}$ and there are also some different aspects in EM which will be described later. Pioneers in EM algorithm firstly assumed that $f(X \mid \Theta)$ belongs to exponential family with note that many popular distributions such as normal, multinomial, and Poisson belong to exponential family. Although DLR (Dempster, Laird, & Rubin, 1977) proposed a generality of EM algorithm in which $f(X \mid \Theta)$ distributes arbitrarily, we should concern exponential family a little bit. Exponential family (Wikipedia, Exponential family, 2016) refers to a set of probabilistic distributions whose PDF (s) have the same exponential form according to equation 1.3 (Dempster, Laird, & Rubin, 1977, p. 3):

$$f(X|\Theta) = b(X)\exp\left(\Theta^T\tau(X)\right)/a(\Theta) \tag{1.3}$$

Where $b(X)$ is a function of X, which is called base measure and $\tau(X)$ is a vector function of X, which is sufficient statistic. For example, the sufficient statistic of normal distribution is $\tau(X) = (X, XX^T)^T$. Equation 1.3 expresses the canonical form of exponential family. Recall that Ω is the domain of Θ such that $\Theta \in \Omega$. Suppose that Ω is a convex set. If Θ is restricted only to Ω then, $f(X \mid \Theta)$ specifies a *regular exponential family*. If Θ lies in a curved sub-manifold Ω_0 of Ω then, $f(X \mid \Theta)$ specifies a *curved exponential family*. The $a(\Theta)$ is *partition function* for variable X, which is used for normalization.

$$a(\Theta) = \int_X b(X)\exp\left(\Theta^T\tau(X)\right)\mathrm{d}X$$

As usual, a PDF is known as a popular form but its exponential family form (canonical form of exponential family) specified by equation 1.3 looks unlike popular form although they are the same. Therefore, parameter in popular form is different from parameter in exponential family form.

For example, multinormal distribution with theoretical mean μ and covariance matrix Σ of random variable $X = (x_1, x_2, \ldots, x_n)^T$ has PDF in popular form is:

$$f(X|\mu,\Sigma) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)$$

Hence, parameter in popular form is $\Theta = (\mu, \Sigma)^T$. Exponential family form of such PDF is:

$$f(X|\theta_1,\theta_2) = (2\pi)^{-\frac{n}{2}} * \exp\left((\theta_1,\theta_2)\begin{pmatrix}X\\XX^T\end{pmatrix}\right)\Big/\exp\left(-\frac{1}{4}\theta_1^T\theta_2^{-1}\theta_1 - \frac{1}{2}\log|-2\theta_2|\right)$$

Where,

$$\Theta = \begin{pmatrix}\theta_1\\\theta_2\end{pmatrix}$$

$$\theta_1 = \Sigma^{-1}\mu$$

$$\theta_2 = -\frac{1}{2}\Sigma^{-1}$$

$$b(X) = (2\pi)^{-\frac{n}{2}}$$

$$\tau(X) = \begin{pmatrix}X\\XX^T\end{pmatrix}$$

$$a(\Theta) = \exp\left(-\frac{1}{4}\theta_1^T\theta_2^{-1}\theta_1 - \frac{1}{2}\log|-2\theta_2|\right)$$

The exponential family form is used to represents all distributions belonging to exponential family as canonical form. Parameter in exponential family form is called exponential family parameter. As a convention, parameter Θ mentioned in EM algorithm is often exponential family parameter if PDF belongs to exponential family and there is no additional information.

Expectation maximization (EM) algorithm has many iterations and each iteration has two steps in which expectation step (E-step) calculates sufficient statistic of hidden data based on observed data and current parameter whereas maximization step (M-step) re-estimates parameter. When DLR proposed EM algorithm (Dempster, Laird, & Rubin, 1977), they firstly concerned that the PDF $f(X \mid \Theta)$ of hidden space belongs to exponential family. E-step and M-step at the t^{th} iteration are described in table 1.1 (Dempster, Laird, & Rubin, 1977, p. 4), in which the current estimate is $\Theta^{(t)}$, with note that $f(X \mid \Theta)$ belongs to regular exponential family.

E-step:

We calculate current value $\tau^{(t)}$ of the sufficient statistic $\tau(X)$ from observed Y and current parameter $\Theta^{(t)}$ according to following equation:

$$\tau^{(t)} = E\left(\tau(X)\middle|Y, \Theta^{(t)}\right) = \int_{\varphi^{-1}(Y)} k\left(X\middle|Y, \Theta^{(t)}\right)\tau(X)\mathrm{d}X$$

M-step:

Basing on $\tau^{(t)}$, we determine the next parameter $\Theta^{(t+1)}$ as solution of following equation:

$$E(\tau(X)|\Theta) = \int_{X} f(X|\Theta)\tau(X)\mathrm{d}X = \tau^{(t)}$$

Note, $\Theta^{(t+1)}$ will become current parameter at the next iteration ($(t+1)^{th}$ iteration).

Table 1.1. E-step and M-step of EM algorithm given regular exponential PDF $f(X|\Theta)$

EM algorithm stops if two successive estimates are equal, $\Theta^* = \Theta^{(t)} = \Theta^{(t+1)}$, at some t^{th} iteration. At that time we conclude that Θ^* is the optimal estimate of EM process. As a convention, the estimate of parameter Θ resulted from EM process is denoted Θ^* instead of $\widehat{\Theta}$ in order to emphasize that Θ^* is solution of optimization problem.

For further research, DLR gave a preeminent generality of EM algorithm (Dempster, Laird, & Rubin, 1977, pp. 6-11) in which $f(X \mid \Theta)$ specifies arbitrary distribution. In other words, there is no requirement of exponential family. They define the conditional expectation $Q(\Theta' \mid \Theta)$ according to equation 1.4 (Dempster, Laird, & Rubin, 1977, p. 6).

$$Q(\Theta'|\Theta) = E\left(\log\left(f(X|\Theta')\right)\middle|Y, \Theta\right) = \int_{\varphi^{-1}(Y)} k(X|Y, \Theta)\log\left(f(X|\Theta')\right)\mathrm{d}X \tag{1.4}$$

If X and Y are discrete, equation 2.4 can be re-written as follows:

$$Q(\Theta'|\Theta) = E\left(\log\left(f(X|\Theta')\right)\middle|Y, \Theta\right) = \sum_{X \in \varphi^{-1}(Y)} k(X|Y, \Theta)\log\left(f(X|\Theta')\right)$$

The two steps of generalized EM (*GEM*) algorithm aim to maximize $Q(\Theta \mid \Theta^{(t)})$ at some t^{th} iteration as seen in table 1.2 (Dempster, Laird, & Rubin, 1977, p. 6).

E-step:

The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current parameter $\Theta^{(t)}$, according to equation 1.4. Actually, $Q(\Theta \mid \Theta^{(t)})$ is formulated as function of Θ.

M-step:

The next parameter $\Theta^{(t+1)}$ is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with subject to Θ. Note that $\Theta^{(t+1)}$ will become current parameter at the next iteration (the $(t+1)^{th}$ iteration).

Table 1.2. E-step and M-step of GEM algorithm

DLR proved that GEM algorithm converges at some t^{th} iteration. At that time, $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ is the optimal estimate of EM process, which is an optimizer of $L(\Theta)$.

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}}\, L(\Theta)$$

It is deduced from E-step and M-step that $Q(\Theta \mid \Theta^{(t)})$ is increased after every iteration. How to maximize $Q(\Theta|\Theta^{(t)})$ is the optimization problem which is dependent on applications. For example, the estimate $\Theta^{(t+1)}$ can be solution of the equation created by setting the first-order derivative of $Q(\Theta|\Theta^{(t)})$ regarding Θ to be zero, $DQ(\Theta|\Theta^{(t)}) = \mathbf{0}^T$. If solving such equation is too

complex or impossible, some popular methods to solve optimization problem are Newton-Raphson (Burden & Faires, 2011, pp. 67-71), gradient descent (Ta, 2014), and Lagrange duality (Wikipedia, Karush–Kuhn–Tucker conditions, 2014).

In practice, if Y is observed as particular N observations $Y_1, Y_2,\ldots, Y_N$. Let $\mathcal{Y} = \{Y_1, Y_2,\ldots, Y_N\}$ be the observed sample of size N with note that all Y_i (s) are mutually independent and identically distributed (iid). Given an observation Y_i, there is an associated random variable X_i. All X_i (s) are iid and they are not existent in fact. Each $X_i \in \boldsymbol{X}$ is a random variable like X. Of course, the domain of each X_i is $\boldsymbol{X}$. Let $\mathcal{X} = \{X_1, X_2,\ldots, X_N\}$ be the set of associated random variables. Because all X_i (s) are iid, the joint PDF of $\mathcal{X}$ is determined as follows:

$$f(\mathcal{X}|\Theta) = f(X_1, X_2, \ldots, X_N|\Theta) = \prod_{i=1}^{N} f(X_i|\Theta)$$

Because all X_i (s) are iid and each Y_i is associated with X_i, the conditional joint PDF of $\mathcal{X}$ given $\mathcal{Y}$ is determined as follows:

$$k(\mathcal{X}|\mathcal{Y}, \Theta) = k(X_1, X_2, \ldots, X_N|Y_1, Y_2, \ldots, Y_N, \Theta) = \prod_{i=1}^{N} k(X_i|Y_1, Y_2, \ldots, Y_N, \Theta) = \prod_{i=1}^{N} k(X_i|Y_i, \Theta)$$

The conditional expectation $Q(\Theta' \mid \Theta)$ given samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ is re-written according to equation 1.5.

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N} \int_{\varphi^{-1}(Y_i)} k(X|Y_i, \Theta)\log\big(f(X|\Theta')\big)\mathrm{d}X \tag{1.5}$$

Equation 1.5 is proved in (Nguyen, 2020, pp. 45-47). In case that $f(X \mid \Theta)$ and $k(X \mid Y_i, \Theta)$ belong to exponential family, equation 1.5 becomes equation 1.6 with an observed sample $\mathcal{Y} = \{Y_1, Y_2,\ldots, Y_N\}$.

$$Q(\Theta'|\Theta) = \left(\sum_{i=1}^{N} E\big(\log\big(b(X)\big)\big|Y_i, \Theta\big)\right) + \left((\Theta')^T \sum_{i=1}^{N} \tau_{\Theta,Y_i}\right) - N\log\big(a(\Theta')\big) \tag{1.6}$$

Where,

$$E\big(\log\big(b(X)\big)\big|Y_i, \Theta\big) = \int_{\varphi^{-1}(Y_i)} k(X|Y_i, \Theta)\log\big(b(X)\big)\mathrm{d}X$$

$$\tau_{\Theta,Y_i} = E(\tau(X)|Y_i, \Theta) = \int_{\varphi^{-1}(Y_i)} k(X|Y_i, \Theta)\tau(X)\mathrm{d}X$$

DLR (Dempster, Laird, & Rubin, 1977, p. 1) called $\boldsymbol{X}$ as *complete data* because the mapping φ: $\boldsymbol{X} \to \boldsymbol{Y}$ is many-one function. There is another case that the complete space $\boldsymbol{Z}$ consists of hidden space $\boldsymbol{X}$ and observed space $\boldsymbol{Y}$ with note that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are separated. There is no explicit mapping φ from $\boldsymbol{X}$ and $\boldsymbol{Y}$ but there exists a PDF of $Z \in \boldsymbol{Z}$ as the joint PDF of $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$.

$$f(Z|\Theta) = f(X, Y|\Theta)$$

The PDF of Y becomes:

$$f(Y|\Theta) = \int_{X} f(X, Y|\Theta)\mathrm{d}X$$

The PDF $f(Y|\Theta)$ is equivalent to the PDF $g(Y|\Theta)$ mentioned in equation 1.1. Although there is no explicit mapping from $\boldsymbol{X}$ to $\boldsymbol{Y}$, the PDF of Y above implies an implicit mapping from $\boldsymbol{Z}$ to $\boldsymbol{Y}$. The conditional PDF of X given Z is specified according to Bayes' rule as follows:

$$f(Z|Y, \Theta) = f(X, Y|Y, \Theta) = f(X|Y)f(Y|Y) = f(X|Y, \Theta) = \frac{f(X, Y|\Theta)}{f(Y|\Theta)} = \frac{f(X, Y|\Theta)}{\int_X f(X, Y|\Theta)\mathrm{d}X}$$

The conditional PDF $f(X|Y, \Theta)$ is equivalent to the conditional PDF $k(X|Y, \Theta)$ mentioned in equation 1.2. Of course, given Y, we always have:

$$\int_X f(X|Y,\Theta)\mathrm{d}X = 1$$

Equation 1.7 specifies the conditional expectation $Q(\Theta' \mid \Theta)$ in case that there is no explicit mapping from $\boldsymbol{X}$ to $\boldsymbol{Y}$ but there exists the joint PDF of X and Y.

$$Q(\Theta'|\Theta) = \int_X f(Z|Y,\Theta)\log\big(f(Z|\Theta')\big)\mathrm{d}X = \int_X f(X|Y,\Theta)\log\big(f(X,Y|\Theta')\big)\mathrm{d}X \tag{1.7}$$

Where,

$$f(X|Y,\Theta) = \frac{f(X,Y|\Theta)}{f(Y|\Theta)} = \frac{f(X,Y|\Theta)}{\int_X f(X,Y|\Theta)\mathrm{d}X}$$

Note, $\boldsymbol{X}$ is separated from $\boldsymbol{Y}$ and the complete data $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$ is composed of $\boldsymbol{X}$ and $\boldsymbol{Y}$. For equation 1.7, the existence of the joint PDF $f(X, Y \mid \Theta)$ can be replaced by the existence of the conditional PDF $f(Y|X, \Theta)$ and the prior PDF $f(X|\Theta)$ due to:

$$f(X,Y|\Theta) = f(Y|X,\Theta)f(X|\Theta)$$

In applied statistics, equation 1.4 is often replaced by equation 1.7 because specifying the joint PDF $f(X, Y \mid \Theta)$ is more practical than specifying the mapping φ: $\boldsymbol{X} \rightarrow \boldsymbol{Y}$. However, equation 1.4 is more general equation 1.7 because the requirement of the joint PDF for equation 1.7 is stricter than the requirement of the explicit mapping for equation 1.4. In case that X and Y are discrete, equation 1.7 becomes:

$$Q(\Theta'|\Theta) = \sum_X P(X|Y,\Theta)\log\big(P(X,Y|\Theta')\big)$$

In practice, suppose Y is observed as a sample $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$ of size N with note that all Y_i (s) are mutually independent and identically distributed (iid). The observed sample $\mathcal{Y}$ is associated with a a hidden set (latent set) $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ of size N. All X_i (s) are iid and they are not existent in fact. Let $X \in \boldsymbol{X}$ be the random variable representing every X_i. Of course, the domain of X is $\boldsymbol{X}$. Equation 1.8 specifies the conditional expectation $Q(\Theta' \mid \Theta)$ given such $\mathcal{Y}$.

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N}\int_X f(X|Y_i,\Theta)\log\big(f(X,Y_i|\Theta')\big)\mathrm{d}X \tag{1.8}$$

Equation 1.8 is a variant of equation 1.5 in case that there is no explicit mapping between X_i and Y_i but there exists the same joint PDF between X_i and Y_i. If both X and Y are discrete, equation 1.8 becomes:

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N}\sum_X P(X|Y_i,\Theta)\log\big(P(X,Y_i|\Theta')\big) \tag{1.9}$$

If X is discrete and Y is continuous such that $f(X, Y \mid \Theta) = P(X|\Theta)f(Y \mid X, \Theta)$ then, according to the total probability rule, we have:

$$f(Y|\Theta) = \sum_X P(X|\Theta)f(Y|X,\Theta)$$

Note, when only X is discrete, its PDF $f(X|\Theta)$ becomes the probability $P(X|\Theta)$. Therefore, equation 1.10 is a variant of equation 1.8, as follows:

$$Q(\Theta'|\Theta) = \sum_{i=1}^{N}\sum_X P(X|Y_i,\Theta)\log\big(P(X|\Theta')f(Y_i|X,\Theta')\big) \tag{1.10}$$

Where $P(X \mid Y_i, \Theta)$ is determined by Bayes' rule, as follows:

$$P(X|Y_i,\Theta) = \frac{P(X|\Theta)f(Y_i|X,\Theta)}{\sum_X P(X|\Theta)f(Y_i|X,\Theta)}$$

Equation 1.10 is the base for estimating the probabilistic mixture model by EM algorithm, which is not main subject of this report. Now we consider how to apply EM into handling missing data in which equation 1.8 is most concerned. The goal of maximum likelihood estimation (MLE), maximum a posteriori (MAP), and EM is to estimate statistical based on sample. Whereas MLE and MAP require complete data, EM accepts hidden data or incomplete data. Therefore, EM is appropriate to handle missing data which contains missing values. Indeed, estimating parameter with missing data is very natural for EM but it is necessary to have a new viewpoint in which missing data is considered as hidden data (X). Moreover, the GEM version with joint probability (without mapping function, please see equation 1.7 and equation 1.8) is used and some changes are required. Handling missing data, which is the main subject of this report is described in next section.

2. Handling missing data

Let $X = (x_1, x_2, \ldots, x_n)^T$ be n-dimension random variable whose n elements are partial random variables x_j (s). Suppose X is composed of two parts such as observed part X_{obs} and missing part X_{mis} such that $X = \{X_{obs}, X_{mis}\}$. Note, X_{obs} and X_{mis} are considered as random variables.

$$X = \{X_{obs}, X_{mis}\} = (x_1, x_2, \ldots, x_n)^T \quad (2.1)$$

When X is observed, X_{obs} and X_{mis} are determined. For example, given $X = (x_1, x_2, x_3, x_4)^T$, when X is observed as $X = (x_1=1, x_2=?, x_3=4, x_4=?, x_5=9)^T$ where question mask "?" denotes missing value, X_{obs} and X_{mis} are determined as $X_{obs} = (x_1=1, x_3=4, x_5=9)^T$ and $X_{mis} = (x_2=?, x_4=?)^T$. When X is observed as $X = (x_1=?, x_2=3, x_3=4, x_4=?, x_5=?)^T$ then, X_{obs} and X_{mis} are determined as $X_{obs} = (x_2=3, x_3=4)^T$ and $X_{mis} = (x_1=?, x_4=?, x_5=?)^T$. Let M be a set of indices that x_j (s) are missing when X is observed. M is called missing index set.

$$M = \{j: x_j \text{ missing}\} \text{ where } j = \overline{1,n} \quad (2.2)$$

Suppose

$$M = \{m_1, m_2, \ldots, m_{|M|}\} \quad (2.3)$$

Where,

$$m_i = \overline{1,n}$$
$$m_i \neq m_j$$

Let $\bar{M}$ is complementary set of the set M given the set $\{1, 2, \ldots, n\}$. $\bar{M}$ is called existent index set.

$$\bar{M} = \{j: x_j \text{ existent}\} \text{ where } j = \overline{1,n} \quad (2.4)$$

M or $\bar{M}$ can be empty. They are mutual because $\bar{M}$ can be defined based on M and vice versa.

$$M \cup \bar{M} = \{1,2, \ldots, n\}$$
$$M \cap \bar{M} = \emptyset$$

Suppose

$$\bar{M} = \{\bar{m}_1, \bar{m}_2, \ldots, \bar{m}_{|\bar{M}|}\} \quad (2.5)$$

Where,

$$\bar{m}_i = \overline{1,n}$$
$$\bar{m}_i \neq \bar{m}_j$$
$$|M| + |\bar{M}| = n$$

We have:

$$X_{mis} = (x_j: j \in M)^T = \left(x_{m_1}, x_{m_2}, \ldots, x_{m_{|M|}}\right)^T \quad (2.6)$$

$$X_{obs} = (x_j: j \in \bar{M})^T = \left(x_{\bar{m}_1}, x_{\bar{m}_2}, \ldots, x_{\bar{m}_{|\bar{M}|}}\right)^T \quad (2.7)$$

Obviously, dimension of X_{mis} is $|M|$ and dimension of X_{obs} is $|\bar{M}| = n-|M|$. Note, when composing X from X_{obs} and X_{mis} as $X = \{X_{obs}, X_{mis}\}$, it is required a right re-arrangement of elements in both X_{obs} and X_{mis}.

Let $Z = (z_1, z_2,\ldots, z_n)^T$ be n-dimension random variable whose each element z_j is binary random variable indicating if x_j is missing. Random variable Z is also called missingness variable.

$$z_j = \begin{cases} 1 \text{ if } x_j \text{ missing} \\ 0 \text{ if } x_j \text{ existent} \end{cases} \tag{2.8}$$

For example, given $X = (x_1, x_2, x_3, x_4)^T$, when X is observed as $X = (x_1=1, x_2=?, x_3=4, x_4=?, x_5=9)^T$, we have $X_{obs} = (x_1=1, x_3=4, x_5=9)^T$, $X_{mis} = (x_2=?, x_4=?)^T$, and $Z = (z_1=0, z_2=1, z_3=0, z_4=1, z_5=0)^T$.

Generally, when X is replaced by a sample $\mathcal{X} = \{X_1, X_2,\ldots, X_N\}$ whose X_i (s) are iid, let $\mathcal{Z} = \{Z_1, Z_2,\ldots, Z_N\}$ be a set of missingness variables associated with $\mathcal{X}$. All Z_i (s) are iid too. $\mathcal{X}$ and $\mathcal{Z}$ can be represented as matrices. Given X_i, its associative quantities are Z_i, M_i, and $\bar{M}_i$. Let $X = \{X_{obs}, X_{mis}\}$ be random variable representing every X_i. Let Z be random variable representing every Z_i. As a convention, $X_{obs}(i)$ and $X_{mis}(i)$ refer to X_{obs} part and X_{mis} part of X_i. We have:

$$\begin{aligned} X_i &= \{X_{obs}(i), X_{mis}(i)\} = (x_{i1}, x_{i2}, \ldots, x_{in})^T \\ X_{mis}(i) &= \left(x_{im_1}, x_{im_2}, \ldots, x_{im_{|M_i|}}\right)^T \\ X_{obs}(i) &= \left(x_{i\bar{m}_{i1}}, x_{i\bar{m}_{i2}}, \ldots, x_{i\bar{m}_{i|\bar{M}_i|}}\right)^T \\ M_i &= \left\{m_{i1}, m_{i2}, \ldots, m_{i|M_i|}\right\} \\ \bar{M}_i &= \left\{\bar{m}_{i1}, \bar{m}_{i2}, \ldots, \bar{m}_{i|\bar{M}_i|}\right\} \\ Z_i &= (z_{i1}, z_{i2}, \ldots, z_{in})^T \end{aligned} \tag{2.9}$$

For example, given sample of size 4, $\mathcal{X} = \{X_1, X_2, X_3, X_4\}$ in which $X_1 = (x_{11}=1, x_{12}=?, x_{13}=3, x_{14}=?)^T$, $X_2 = (x_{21}=?, x_{22}=2, x_{23}=?, x_{24}=4)^T$, $X_3 = (x_{31}=1, x_{32}=2, x_{33}=?, x_{34}=?)^T$, and $X_4 = (x_{41}=?, x_{42}=?, x_{43}=3, x_{44}=4)^T$ are iid. Therefore, we also have $Z_1 = (z_{11}=0, z_{12}=1, z_{13}=0, z_{14}=1)^T$, $Z_2 = (z_{21}=1, z_{22}=0, z_{23}=1, z_{24}=0)^T$, $Z_3 = (z_{31}=0, z_{32}=0, z_{33}=1, z_{34}=1)^T$, and $Z_4 = (z_{41}=1, z_{42}=1, z_{43}=0, z_{44}=0)^T$. All Z_i (s) are iid too.

	x_1	x_2	x_3	x_4
X_1	1	?	3	?
X_2	?	2	?	4
X_3	1	2	?	?
X_4	?	?	3	4

	z_1	z_2	z_3	z_4
Z_1	0	1	0	1
Z_2	1	0	1	0
Z_3	0	0	1	1
Z_4	1	1	0	0

Of course, we have $X_{obs}(1) = (x_{11}=1, x_{13}=3)^T$, $X_{mis}(1) = (x_{12}=?, x_{14}=?)^T$, $X_{obs}(2) = (x_{22}=2, x_{24}=4)^T$, $X_{mis}(2) = (x_{21}=?, x_{23}=?)^T$, $X_{obs}(3) = (x_{31}=1, x_{32}=2)^T$, $X_{mis}(3) = (x_{33}=?, x_{34}=?)^T$, $X_{obs}(4) = (x_{43}=3, x_{44}=4)^T$, and $X_{mis}(4) = (x_{41}=?, x_{42}=?)^T$. We also have $M_1 = \{m_{11}=2, m_{12}=4\}$, $\bar{M}_1 = \{\bar{m}_{11}=1, \bar{m}_{12}=3\}$, $M_2 = \{m_{21}=1, m_{22}=3\}$, $\bar{M}_2 = \{\bar{m}_{21}=2, \bar{m}_{22}=4\}$, $M_3 = \{m_{31}=3, m_{32}=4\}$, $\bar{M}_3 = \{\bar{m}_{31}=1, \bar{m}_{32}=2\}$, $M_4 = \{m_{41}=1, m_{42}=2\}$, and $\bar{M}_4 = \{\bar{m}_{41}=3, \bar{m}_{42}=4\}$.

Both X and Z are associated with their own PDFs, as follows:

$$\begin{aligned} &f(X|\Theta) = f(X_{obs}, X_{mis}|\Theta) \\ &f(Z|X_{obs}, X_{mis}, \Phi) \end{aligned} \tag{2.10}$$

Where Θ and Φ are parameters of PDFs of $X = \{X_{obs}, X_{mis}\}$ and Z, respectively. The goal of handling missing data is to estimate Θ and Φ given X. Sufficient statistic of $X = \{X_{obs}, X_{mis}\}$ is composed of sufficient statistic of X_{obs} and sufficient statistic of X_{mis}.

$$\tau(X) = \tau(X_{obs}, X_{mis}) = \{\tau(X_{obs}), \tau(X_{mis})\} \tag{2.11}$$

How to compose $\tau(X)$ from $\tau(X_{obs})$ and $\tau(X_{mis})$ is dependent on distribution type of the PDF $f(X|\Theta)$.

The joint PDF of X and Z is main object of handling missing data, which is defined as follows:

$$f(X, Z|\Theta, \Phi) = f(X_{obs}, X_{mis}, Z|\Theta, \Phi) = f(Z|X_{obs}, X_{mis}, \Phi) f(X_{obs}, X_{mis}|\Theta) \quad (2.12)$$

The PDF of X_{obs} is defined as integral of $f(X|\Theta)$ over X_{mis}:

$$f(X_{obs}|\Theta) = \int_{X_{mis}} f(X_{obs}, X_{mis}|\Theta) \mathrm{d}X_{mis} \quad (2.13)$$

The PDF of X_{mis} is conditional PDF of X_{mis} given X_{obs} is:

$$f(X_{mis}|X_{obs}, \Theta_M) = f(X_{mis}|X_{obs}, \Theta) = \frac{f(X|\Theta)}{f(X_{obs}|\Theta)} = \frac{f(X_{obs}, X_{mis}|\Theta)}{f(X_{obs}|\Theta)} \quad (2.14)$$

The notation Θ_M implies that the parameter Θ_M of the PDF $f(X_{mis} \mid X_{obs}, \Theta_M)$ is derived from the parameter Θ of the PDF $f(X|\Theta)$, which is function of Θ and X_{obs} as $\Theta_M = u(\Theta, X_{obs})$. Thus, Θ_M is not a new parameter and it is dependent on distribution type.

$$\Theta_M = u(\Theta, X_{obs}) \quad (2.15)$$

How to determine $u(\Theta, X_{obs})$ is dependent on distribution type of the PDF $f(X|\Theta)$.

There are three types of missing data, which depends on relationship between X_{obs}, X_{mis}, and Z (Josse, Jiang, Sportisse, & Robin, 2018):

- Missing data (X or $\mathcal{X}$) is Missing Completely At Random (MCAR) if the probability of Z depends on both X_{obs} and X_{mis} such that $f(Z \mid X_{obs}, X_{mis}, \Phi) = f(Z \mid \Phi)$.
- Missing data (X or $\mathcal{X}$) is Missing At Random (MAR) if the probability of Z depends on only X_{obs} such that $f(Z \mid X_{obs}, X_{mis}, \Phi) = f(Z \mid X_{obs}, \Phi)$.
- Missing data (X or $\mathcal{X}$) is Missing Not At Random (MNAR) in all other cases.

There are two main approaches for handling missing data (Josse, Jiang, Sportisse, & Robin, 2018):

- Using some statistical models such as EM to estimate parameter with missing data.
- Inputting plausible values for missing values to obtain some complete samples (copies) from the missing data. Later on, every complete sample is used to produce an estimate of parameter by some estimation methods, for example, MLE and MAP. Finally, all estimates are synthesized to produce the best estimate.

Here we focus on the first approach with EM to estimate parameter with missing data. Without loss of generality, given sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ in which all X_i (s) are iid, by applying equation 1.8 for GEM with the joint PDF $f(X_{obs}, X_{mis}, Z \mid \Theta, \Phi)$, we consider $\{X_{obs}, Z\}$ as observed part and X_{mis} as hidden part. Let $X = \{X_{obs}, X_{mis}\}$ be random variable representing all X_i (s). Let $X_{obs}(i)$ denote observed part X_{obs} of X_i and let Z_i be missingness variable corresponding to X_i, by following equation 1.8, the expectation $Q(\Theta', \Phi' \mid \Theta, \Phi)$ becomes:

$$Q(\Theta', \Phi'|\Theta, \Phi) = \sum_{i=1}^{N} \int_{X_{mis}} f(X_{mis}|X_{obs}(i), Z_i, \Theta, \Phi) * \log\big(f(X_{obs}(i), X_{mis}, Z_i|\Theta', \Phi')\big) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f(X_{mis}|X_{obs}(i), \Theta) * \log\big(f(X_{obs}(i), X_{mis}, Z_i|\Theta', \Phi')\big) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\big(X_{mis}\big|X_{obs}(i), \Theta_{M_i}\big) * \log\big(f(X_{obs}(i), X_{mis}, Z_i|\Theta', \Phi')\big) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) * \log\left(f(X_{obs}(i), X_{mis}|\Theta', \Phi') * f(Z_i|X_{obs}(i), X_{mis}, \Theta', \Phi')\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) * \log\left(f(X_{obs}(i), X_{mis}|\Theta') * f(Z_i|X_{obs}(i), X_{mis}, \Phi')\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) * \left(\log\left(f(X_{obs}(i), X_{mis}|\Theta')\right) + \log\left(f(Z_i|X_{obs}(i), X_{mis}, \Phi')\right)\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) \log\left(f(X_{obs}(i), X_{mis}|\Theta')\right) \mathrm{d}X_{mis} + \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) \log\left(f(Z_i|X_{obs}(i), X_{mis}, \Phi')\right) \mathrm{d}X_{mis}$$

In short, $Q(\Theta', \Phi' | \Theta, \Phi)$ is specified as follows:

$$Q(\Theta', \Phi'|\Theta, \Phi) = Q_1(\Theta'|\Theta) + Q_2(\Phi'|\Theta) \tag{2.16}$$

Where,

$$Q_1(\Theta'|\Theta) = \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) \log\left(f(X_{obs}(i), X_{mis}|\Theta')\right) \mathrm{d}X_{mis}$$

$$Q_2(\Phi'|\Theta) = \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) \log\left(f(Z_i|X_{obs}(i), X_{mis}, \Phi')\right) \mathrm{d}X_{mis}$$

Note, unknowns of $Q(\Theta', \Phi' | \Theta, \Phi)$ are Θ' and Φ'. Because it is not easy to maximize $Q(\Theta', \Phi' | \Theta, \Phi)$ with regard to Θ' and Φ', we assume that the PDF $f(X|\Theta)$ belongs to exponential family.

$$f(X|\Theta) = f(X_{obs}, X_{mis}|\Theta) = b(X_{obs}, X_{mis}) * \exp\left((\Theta)^T \tau(X_{obs}, X_{mis})\right)/a(\Theta) \tag{2.17}$$

Note,

$$b(X) = b(X_{obs}, X_{mis})$$
$$\tau(X) = \tau(X_{obs}, X_{mis}) = \{\tau(X_{obs}), \tau(X_{mis})\}$$

It is easy to deduce that

$$f(X_{mis}|X_{obs}, \Theta_M) = b(X_{mis}) \exp\left((\Theta_M)^T \tau(X_{mis})\right)/a(\Theta_M) \tag{2.18}$$

Therefore,

$$f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) = b(X_{mis}) \exp\left(\left(\Theta_{M_i}\right)^T \tau(X_{mis})\right)/a\left(\Theta_{M_i}\right)$$

We have:

$$Q_1(\Theta'|\Theta) = \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) \log\left(f(X_{obs}(i), X_{mis}|\Theta')\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right) * \log\left(b(X_{obs}(i), X_{mis}) \exp\left((\Theta')^T \tau(X_{obs}(i), X_{mis})\right)/a(\Theta')\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) * \left(\log\left(b(X_{obs}(i), X_{mis})\right) + (\Theta')^T \tau(X_{obs}(i), X_{mis}) - \log\left(a(\Theta')\right)\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \log\left(b(X_{obs}(i), X_{mis})\right) \mathrm{d}X_{mis} + \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) (\Theta')^T \tau(X_{obs}(i), X_{mis}) \mathrm{d}X_{mis} - \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \log\left(a(\Theta')\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \log\left(b(X_{obs}(i), X_{mis})\right) \mathrm{d}X_{mis} + (\Theta')^T \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \tau(X_{obs}(i), X_{mis}) \mathrm{d}X_{mis} - \log\left(a(\Theta')\right) \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \log\left(b(X_{obs}(i), X_{mis})\right) \mathrm{d}X_{mis} + (\Theta')^T \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \tau(X_{obs}(i), X_{mis}) \mathrm{d}X_{mis} - N\log\left(a(\Theta')\right)$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \log\left(b(X_{obs}(i), X_{mis})\right) \mathrm{d}X_{mis} + (\Theta')^T \sum_{i=1}^{N} \begin{Bmatrix} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \tau\left(X_{obs}(i)\right) \mathrm{d}X_{mis}, \\ \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \tau(X_{mis}) \mathrm{d}X_{mis} \end{Bmatrix} - N\log\left(a(\Theta')\right)$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \log\left(b(X_{obs}(i), X_{mis})\right) \mathrm{d}X_{mis} + (\Theta')^T \sum_{i=1}^{N} \begin{Bmatrix} \tau\left(X_{obs}(i)\right) \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \mathrm{d}X_{mis}, \\ \int_{X_{mis}} f\left(X_{mis} \middle| X_{obs}(i), \Theta_{M_i}\right) \tau(X_{mis}) \mathrm{d}X_{mis} \end{Bmatrix} - N\log\left(a(\Theta')\right)$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right)\log\left(b(X_{obs}(i), X_{mis})\right)\mathrm{d}X_{mis}$$
$$+ (\Theta')^T \sum_{i=1}^{N} \left\{ \begin{matrix} \tau\left(X_{obs}(i)\right), \\ \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right)\tau(X_{mis})\mathrm{d}X_{mis} \end{matrix} \right\} - N\log\left(a(\Theta')\right)$$

Therefore, equation 2.19 specifies $Q_1(\Theta'|\Theta)$ given $f(X|\Theta)$ belongs to exponential family.

$$Q_1(\Theta'|\Theta) = \sum_{i=1}^{N} E\left(\log\left(b(X_{obs}(i), X_{mis})\right)\middle|\Theta_{M_i}\right) + (\Theta')^T \sum_{i=1}^{N} \left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}\right)\right\} - N\log\left(a(\Theta')\right) \quad (2.19)$$

Where,

$$E\left(\log\left(b(X_{obs}(i), X_{mis})\right)\middle|\Theta_{M_i}\right) = \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right)\log\left(b(X_{obs}(i), X_{mis})\right)\mathrm{d}X_{mis} \quad (2.20)$$

$$E\left(\tau(X_{mis})\middle|\Theta_{M_i}\right) = \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}\right)\tau(X_{mis})\mathrm{d}X_{mis} \quad (2.21)$$

At M-step of some t^{th} iteration, the next parameter $\Theta^{(t+1)}$ is solution of the equation created by setting the first-order derivative of $Q_1(\Theta'|\Theta)$ to be zero. The first-order derivative of $Q_1(\Theta'|\Theta)$ is:

$$\frac{\partial Q_1(\Theta'|\Theta)}{\partial \Theta'} = \sum_{i=1}^{N} \left(E\left(\tau(X_{obs}(i), X_{mis})\middle|\Theta_{M_i}\right)\right)^T - N\log'\left(a(\Theta')\right)$$
$$= \sum_{i=1}^{N} \left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}\right)\right\}^T - N\log'\left(a(\Theta')\right)$$

By referring table 1.2, we have:

$$\log'\left(a(\Theta')\right) = \left(E(\tau(X)|\Theta')\right)^T = \int_X f(X|\Theta)\left(\tau(X)\right)^T\mathrm{d}X$$

Where,

$$f(X|\Theta) = f(X_{obs}, X_{mis}|\Theta) = b(X_{obs}, X_{mis}) * \exp\left((\Theta)^T\tau(X_{obs}, X_{mis})\right)/a(\Theta)$$
$$b(X) = b(X_{obs}, X_{mis})$$
$$\tau(X) = \tau(X_{obs}, X_{mis}) = \{\tau(X_{obs}), \tau(X_{mis})\}$$

Thus, the next parameter $\Theta^{(t+1)}$ is solution of the following equation:

$$\frac{\partial Q_1(\Theta'|\Theta)}{\partial \Theta'} = \sum_{i=1}^{N} \left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}\right)\right\}^T - N\left(E(\tau(X)|\Theta')\right)^T = \mathbf{0}^T$$

This implies the next parameter $\Theta^{(t+1)}$ is solution of the following equation:

$$E(\tau(X)|\Theta') = \frac{1}{N}\sum_{i=1}^{N} \left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}\right)\right\}$$

As a result, at E-step of some t^{th} iteration, given current parameter $\Theta^{(t)}$, the sufficient statistic of X is calculated as follows:

$$\tau^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}^{(t)}\right)\right\} \quad (2.22)$$

Where,

$$\Theta_{M_i}^{(t)} = u\left(\Theta^{(t)}, M_i\right)$$

$$E\left(\tau(X_{mis})\middle|\Theta_{M_i}^{(t)}\right) = \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}^{(t)}\right)\tau(X_{mis})\mathrm{d}X_{mis}$$

Equation 2.22 is variant of equation 2.11 when $f(X|\Theta)$ belongs to exponential family but how to compose $\tau(X)$ from $\tau(X_{obs})$ and $\tau(X_{mis})$ is not determined exactly yet.

As a result, at M-step of some t^{th} iteration, given $\tau^{(t)}$ and $\Theta^{(t)}$, the next parameter $\Theta^{(t+1)}$ is a solution of the following equation:

$$E(\tau(X)|\Theta) = \tau^{(t)} \quad (2.23)$$

Moreover, at M-step of some t^{th} iteration, the next parameter $\Phi^{(t+1)}$ is a maximizer of $Q_2(\Phi \mid \Theta^{(t)})$ given $\Theta^{(t)}$ as follows:

$$\Phi^{(t+1)} = \underset{\Phi}{\operatorname{argmin}}\, Q_2\left(\Phi\middle|\Theta^{(t)}\right) \quad (2.24)$$

Where,

$$Q_2\left(\Phi\middle|\Theta^{(t)}\right) = \sum_{i=1}^{N}\int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}^{(t)}\right)\log\left(f(Z_i|X_{obs}(i), X_{mis}, \Phi)\right)\mathrm{d}X_{mis} \quad (2.25)$$

How to maximize $Q_2(\Phi \mid \Theta^{(t)})$ depends on distribution type of Z_i which is also formulation of the PDF $f(Z \mid X_{obs}, X_{mis}, \Phi)$. For some reasons, such as accelerating estimation speed or ignoring missingness variable Z then, the next parameter $\Phi^{(t+1)}$ will not be estimated.

In general, the two steps of GEM algorithm for handling missing data at some t^{th} iteration are summarized in table 2.1 with assumption that the PDF of missing data $f(X|\Theta)$ belongs to exponential family.

E-step:

Given current parameter $\Theta^{(t)}$, the sufficient statistic $\tau^{(t)}$ is calculated according to equation 2.22.

$$\tau^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}^{(t)}\right)\right\}$$

Where,

$$\Theta_{M_i}^{(t)} = u\left(\Theta^{(t)}, M_i\right)$$

$$E\left(\tau(X_{mis})\middle|\Theta_{M_i}^{(t)}\right) = \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}^{(t)}\right)\tau(X_{mis})\mathrm{d}X_{mis}$$

M-step:

Given $\tau^{(t)}$ and $\Theta^{(t)}$, the next parameter $\Theta^{(t+1)}$ is a solution of equation 2.23.

$$E(\tau(X)|\Theta) = \tau^{(t)}$$

Given $\Theta^{(t)}$, the next parameter $\Phi^{(t+1)}$ is a maximizer of $Q_2(\Phi \mid \Theta^{(t)})$ according to equation 2.24.

$$\Phi^{(t+1)} = \underset{\Phi}{\operatorname{argmin}}\, Q_2\left(\Phi\middle|\Theta^{(t)}\right)$$

Where,

$$Q_2\left(\Phi\middle|\Theta^{(t)}\right) = \sum_{i=1}^{N}\int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}(i), \Theta_{M_i}^{(t)}\right)\log\left(f(Z_i|X_{obs}(i), X_{mis}, \Phi)\right)\mathrm{d}X_{mis}$$

Table 2.1. E-step and M-step of GEM algorithm for handling missing data given exponential PDF

GEM algorithm converges at some t^{th} iteration. At that time, $\Theta^* = \Theta^{(t+1)} = \Theta^{(t)}$ and $\Phi^* = \Phi^{(t+1)} = \Phi^{(t)}$ are optimal estimates. If missingness variable Z is ignored for some reasons, parameter Φ is not estimated. Because X_{mis} is a part of X and $f(X_{mis} \mid X_{obs}, \Theta_M)$ is derived directly from $f(X|\Theta)$, in practice we can stop GEM after its first iteration was done, which is reasonable enough to handle missing data.

An interesting application of handling missing data is to fill in or predict missing values. For instance, suppose the estimate resulted from GEM is Θ^*, missing values represented by $\tau(X_{mis})$ are fulfilled by expectation of $\tau(X_{mis})$ as follows:

$$\tau(X_{mis}) = E(\tau(X_{mis})|\Theta_M^*) \tag{2.26}$$

Where,

$$\Theta_M^* = u(\Theta^*, X_{obs})$$

Now we survey a popular case that sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ whose X_i (s) are iid is MCAR data and $f(X|\Theta)$ is multinormal PDF whereas missingness variable Z follows binomial distribution of n trials. Let $X = \{X_{obs}, X_{mis}\}$ be random variable representing every X_i. Suppose dimension of X is n. Let Z be random variable representing every Z_i. According to equation 2.9, recall that

$$X_i = \{X_{obs}(i), X_{mis}(i)\} = (x_{i1}, x_{i2}, \ldots, x_{in})^T$$
$$X_{mis}(i) = \left(x_{im_1}, x_{im_2}, \ldots, x_{im_{|M_i|}}\right)^T$$
$$X_{obs}(i) = \left(x_{i\bar{m}_{i1}}, x_{i\bar{m}_{i2}}, \ldots, x_{i\bar{m}_{i|\bar{M}_i|}}\right)^T$$
$$M_i = \left\{m_{i1}, m_{i2}, \ldots, m_{i|M_i|}\right\}$$
$$\bar{M}_i = \left\{\bar{m}_{i1}, \bar{m}_{i2}, \ldots, \bar{m}_{i|\bar{M}_i|}\right\}$$
$$Z_i = (z_{i1}, z_{i2}, \ldots, z_{in})^T$$

The PDF of X is:

$$f(X|\Theta) = f(X_{obs}, X_{mis}|\Theta) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right) \tag{2.27}$$

Therefore,

$$f(X_i|\Theta) = f(X_{obs}(i), X_{mis}(i)|\Theta) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right)$$

The PDF of Z is:

$$f(Z|\Phi) = p^{c(Z)}(1-p)^{n-c(Z)} \tag{2.28}$$

Therefore,

$$f(Z_i|\Phi) = p^{c(Z_i)}(1-p)^{n-c(Z_i)}$$

Where $\Theta = (\mu, \Sigma)^T$ and $\Phi = p$.

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T$$
$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix} \tag{2.29}$$

Suppose the probability of missingness at every partial random variable x_j is p and it is independent from X_{obs} and X_{mis}. The quantity $c(Z)$ is the number of z_j (s) in Z that equal 1. For example, if $Z = (1, 0, 1, 0)^T$ then, $c(Z) = 2$. The most important task here is to define equation 2.11 and equation 2.15 in order to compose $\tau(X)$ from $\tau(X_{obs})$, $\tau(X_{mis})$ and to extract Θ_M from Θ when $f(X|\Theta)$ distributes normally.

The conditional PDF of X_{mis} given X_{obs} is also multinormal PDF.

$$f(X_{mis}|\Theta_M) = f(X_{mis}|X_{obs},\Theta_M) = f(X_{mis}|X_{obs},\Theta) = (2\pi)^{-\frac{|M|}{2}}|\Sigma_M|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X_{mis}-\mu_M)^T\Sigma_M^{-1}(X_{mis}-\mu_M)\right) \quad (2.30)$$

Therefore,

$$f\left(X_{mis}(i)\middle|\Theta_{M_i}\right) = f\left(X_{mis}(i)\middle|X_{obs}(i),\Theta_{M_i}\right) = f(X_{mis}(i)|X_{obs}(i),\Theta) = (2\pi)^{-\frac{|M_i|}{2}}\left|\Sigma_{M_i}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(X_{mis}(i)-\mu_{M_i}\right)^T\Sigma_{M_i}^{-1}(X_{mis}(i)-\mu_{M_i})\right)$$

Where $\Theta_{M_i} = \left(\mu_{M_i},\Sigma_{M_i}\right)^T$. We denote

$$f\left(X_{mis}(i)\middle|\Theta_{M_i}\right) = f\left(X_{mis}(i)\middle|X_{obs}(i),\Theta_{M_i}\right)$$

Because $f\left(X_{mis}(i)\middle|X_{obs}(i),\Theta_{M_i}\right)$ only depends on Θ_{M_i} within normal PDF whereas Θ_{M_i} depends on $X_{obs}(i)$. Determining the function $\Theta_{M_i} = u(\Theta, X_{obs}(i))$ is now necessary to extract the parameter Θ_{M_i} from Θ given $X_{obs}(i)$ when $f(X_i|\Theta)$ is normal distribution.

Let $\Theta_{mis} = (\mu_{mis}, \Sigma_{mis})^T$ be parameter of marginal PDF of X_{mis}, we have:

$$f(X_{mis}|\Theta_{mis}) = (2\pi)^{-\frac{|M|}{2}}|\Sigma_{mis}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X_{mis}-\mu_{mis})^T(\Sigma_{mis})^{-1}(X_{mis}-\mu_{mis})\right) \quad (2.31)$$

Therefore,

$$f\left(X_{mis}(i)\middle|\Theta_{mis}(i)\right) = (2\pi)^{-\frac{|M_i|}{2}}|\Sigma_{mis}(i)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(X_{mis}(i)-\mu_{mis}(i)\right)^T\left(\Sigma_{mis}(i)\right)^{-1}(X_{mis}(i)-\mu_{mis}(i))\right)$$

Where,

$$\mu_{mis}(i) = \left(\mu_{m_{i1}},\mu_{m_{i2}},\dots,\mu_{m_{i|M_i|}}\right)^T$$

$$\Sigma_{mis}(i) = \begin{pmatrix} \sigma_{m_{i1}m_{i1}} & \sigma_{m_{i1}m_{i2}} & \cdots & \sigma_{m_{i1}m_{i|M_i|}} \\ \sigma_{m_{i2}m_{i1}} & \sigma_{m_{i2}m_{i2}} & \cdots & \sigma_{m_{i2}m_{i|M_i|}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m_{i|M_i|}m_{i1}} & \sigma_{m_{i|M_i|}m_{i2}} & \cdots & \sigma_{m_{i|M_i|}m_{i|M_i|}} \end{pmatrix} \quad (2.32)$$

Obviously, $\Theta_{mis}(i)$ is extracted from Θ given indicator M_i.

Let $\Theta_{obs} = (\mu_{obs}, \Sigma_{obs})^T$ be parameter of marginal PDF of X_{obs}, we have:

$$f(X_{obs}|\Theta_{obs}) = (2\pi)^{-\frac{|\bar{M}|}{2}}|\Sigma_{obs}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(X_{obs}-\mu_{obs})^T(\Sigma_{obs})^{-1}(X_{obs}-\mu_{obs})\right) \quad (2.33)$$

Therefore,

$$f\left(X_{obs}(i)\middle|\Theta_{obs}(i)\right) = (2\pi)^{-\frac{|\bar{M}_i|}{2}}|\Sigma_{obs}(i)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(X_{obs}(i)-\mu_{obs}(i)\right)^T\left(\Sigma_{obs}(i)\right)^{-1}(X_{obs}(i)-\mu_{obs}(i))\right)$$

Where,

$$\mu_{obs}(i) = \left(\mu_{\bar{m}_{i1}}, \mu_{\bar{m}_{i2}}, \dots, \mu_{\bar{m}_{i|\bar{M}_i|}}\right)^T$$

$$\Sigma_{obs}(i) = \begin{pmatrix} \sigma_{\bar{m}_{i1}\bar{m}_{i1}} & \sigma_{\bar{m}_{i1}\bar{m}_{i2}} & \cdots & \sigma_{\bar{m}_{i1}\bar{m}_{i|\bar{M}_i|}} \\ \sigma_{\bar{m}_{i2}\bar{m}_{i1}} & \sigma_{\bar{m}_{i2}\bar{m}_{i2}} & \cdots & \sigma_{\bar{m}_{i2}\bar{m}_{i|\bar{M}_i|}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\bar{m}_{i|\bar{M}_i|}\bar{m}_{i1}} & \sigma_{\bar{m}_{i|\bar{M}_i|}\bar{m}_{i2}} & \cdots & \sigma_{\bar{m}_{i|\bar{M}_i|}\bar{m}_{i|\bar{M}_i|}} \end{pmatrix} \quad (2.34)$$

Obviously, $\Theta_{obs}(i)$ is extracted from Θ given indicator $\bar{M}_i$ or M_i. We have:

$$f\left(X_{mis}(i)\middle|\Theta_{mis}(i)\right) = \int_{X_{obs}(i)} f(X_{obs}(i), X_{mis}(i)|\Theta)\mathrm{d}X_{obs}(i)$$

$$f\left(X_{obs}(i)\middle|\Theta_{obs}(i)\right) = \int_{X_{mis}(i)} f(X_{obs}(i), X_{mis}(i)|\Theta)\mathrm{d}X_{mis}(i)$$

$$f\left(X_{mis}(i)\middle|\Theta_{M_i}\right) = f(X_{mis}(i)|X_{obs}(i), \Theta) = \frac{f(X_{obs}(i), X_{mis}(i)|\Theta)}{f\left(X_{obs}(i)\middle|\Theta_{obs}(i)\right)}$$

Therefore, it is easy to form the parameter $\Theta_{M_i} = \left(\mu_{M_i}, \Sigma_{M_i}\right)^T$ from $\Theta_{mis}(i) = (\mu_{mis}(i), \Sigma_{mis}(i))^T$ and $\Theta_{obs}(i) = (\mu_{obs}(i), \Sigma_{obs}(i))^T$ as follows (Hardle & Simar, 2013, pp. 156-157):

$$\Theta_{M_i} = u\left(\Theta, X_{obs}(i)\right) = \begin{cases} \mu_{M_i} = \mu_{mis}(i) + \left(V_{obs}^{mis}(i)\right)\left(\Sigma_{obs}(i)\right)^{-1}\left(X_{obs}(i) - \mu_{obs}(i)\right) \\ \Sigma_{M_i} = \Sigma_{mis}(i) - \left(V_{obs}^{mis}(i)\right)\left(\Sigma_{obs}(i)\right)^{-1}\left(V_{mis}^{obs}\right) \end{cases} \quad (2.35)$$

Where from $\Theta_{mis}(i) = (\mu_{mis}(i), \Sigma_{mis}(i))^T$ and $\Theta_{obs}(i) = (\mu_{obs}(i), \Sigma_{obs}(i))^T$ are specified by equation 2.32 and equation 2.34. Moreover the $k_x l$ matrix $V_{obs}^{mis}(i)$ which implies correlation between X_{mis} and X_{obs} is defined as follows:

$$V_{obs}^{mis}(i) = \begin{pmatrix} \sigma_{m_{i1}\bar{m}_{i1}} & \sigma_{m_{i1}\bar{m}_{i2}} & \cdots & \sigma_{m_{i1}\bar{m}_{i|\bar{M}_i|}} \\ \sigma_{m_{i2}\bar{m}_{i1}} & \sigma_{m_{i2}\bar{m}_{i2}} & \cdots & \sigma_{m_{i2}\bar{m}_{i|\bar{M}_i|}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m_{i|M_i|}\bar{m}_{i1}} & \sigma_{m_{i|M_i|}\bar{m}_{i2}} & \cdots & \sigma_{m_{i|M_i|}\bar{m}_{i|\bar{M}_i|}} \end{pmatrix} \quad (2.36)$$

The $k_x l$ matrix $V_{mis}^{obs}(i)$ which implies correlation between X_{obs} and X_{mis} is defined as follows:

$$V_{mis}^{obs}(i) = \begin{pmatrix} \sigma_{\bar{m}_{i1}m_{i1}} & \sigma_{\bar{m}_{i1}m_{i2}} & \cdots & \sigma_{\bar{m}_{i1}m_{i|M_i|}} \\ \sigma_{\bar{m}_{i2}m_{i1}} & \sigma_{\bar{m}_{i2}m_{i2}} & \cdots & \sigma_{\bar{m}_{i2}m_{i|M_i|}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\bar{m}_{i|\bar{M}_i|}m_{i1}} & \sigma_{\bar{m}_{i|\bar{M}_i|}m_{i2}} & \cdots & \sigma_{\bar{m}_{i|\bar{M}_i|}m_{i|M_i|}} \end{pmatrix} \quad (2.37)$$

Therefore, equation 2.35 to extract Θ_{M_i} from Θ given $X_{obs}(i)$ is an instance of equation 2.15. For convenience let,

$$\mu_{M_i} = \left(\mu_{M_i}(m_{i1}), \mu_{M_i}(m_{i2}), \dots, \mu_{M_i}\left(m_{i|M_i|}\right)\right)^T$$

$$\Sigma_{M_i} = \begin{pmatrix} \Sigma_{M_i}(m_{i1}, m_{i1}) & \Sigma_{M_i}(m_{i1}, m_{i2}) & \cdots & \Sigma_{M_i}\left(m_{i1}, m_{i|M_i|}\right) \\ \Sigma_{M_i}(m_{i2}, m_{i1}) & \Sigma_{M_i}(m_{i2}, m_{i2}) & \cdots & \Sigma_{M_i}\left(m_{i2}, m_{i|M_i|}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{M_i}\left(m_{i|M_i|}, m_{i1}\right) & \Sigma_{M_i}\left(m_{i|M_i|}, m_{i2}\right) & \cdots & \Sigma_{M_i}\left(m_{i|M_i|}, m_{i|M_i|}\right) \end{pmatrix} \quad (2.38)$$

Equation 2.38 is result of equation 2.35. Given $X_{mis}(i) = \left(x_{m_{i1}}, x_{m_{i2}}, \dots, x_{m_{i|M_i|}}\right)^T$ then, $\mu_{M_i}\left(m_{ij}\right)$ is estimated partial mean of $x_{m_{ij}}$ and $\Sigma_{M_i}(m_{iu}, m_{iv})$ is estimated partial covariance of $x_{m_{iu}}$ and $x_{m_{iv}}$ given the conditional PDF $f(X_{mis} \mid \Theta_{M_i})$.

At E-step of some t^{th} iteration, given current parameter $\Theta^{(t)}$, the sufficient statistic of X is calculated according to equation 2.22. Let,

$$\tau^{(t)} = \left(\tau_1^{(t)}, \tau_2^{(t)}\right)^T = \frac{1}{N}\sum_{i=1}^{N}\left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}^{(t)}\right)\right\}$$

It is necessary to calculate the sufficient with normal PDF $f(X_i|\Theta)$, which means that we need to define what $\tau_1^{(t)}$ and $\tau_2^{(t)}$ are. The sufficient statistic of $X_{obs}(i)$ is:

$$\tau\left(X_{obs}(i)\right) = \left(X_{obs}(i), X_{obs}(i)\left(X_{obs}(i)\right)^T\right)^T$$

The sufficient statistic of $X_{mis}(i)$ is:

$$\tau\left(X_{mis}(i)\right) = \left(X_{mis}(i), X_{mis}(i)\left(X_{mis}(i)\right)^T\right)^T$$

We also have:

$$E\left(\tau(X_{mis})\middle|\Theta_{M_i}^{(t)}\right) = \int_{X_{mis}} f\left(X_{mis}\middle|\Theta_{M_i}^{(t)}\right)\tau(X_{mis})\mathrm{d}X_{mis} = \begin{pmatrix}\mu_{M_i}^{(t)} \\ \Sigma_{M_i}^{(t)} + \mu_{M_i}^{(t)}\left(\mu_{M_i}^{(t)}\right)^T\end{pmatrix}$$

Due to

$$E\left(X_{mis}(i)\left(X_{mis}(i)\right)^T\middle|\Theta_{M_i}^{(t)}\right) = \Sigma_{M_i}^{(t)} + \mu_{M_i}^{(t)}\left(\mu_{M_i}^{(t)}\right)^T$$

Where $\mu_{M_i}^{(t)}$ and $\Sigma_{M_i}^{(t)}$ are μ_{M_i} and Σ_{M_i} at current iteration, respectively. By referring to equation 2.38, we have

$$\mu_{M_i}^{(t)} = \left(\mu_{M_i}^{(t)}(m_{i1}), \mu_{M_i}^{(t)}(m_{i2}), \dots, \mu_{M_i}^{(t)}(m_{i|M_i|})\right)^T$$

And

$$\Sigma_{M_i}^{(t)} + \mu_{M_i}^{(t)}\left(\mu_{M_i}^{(t)}\right)^T = \begin{pmatrix}\tilde{\sigma}_{11}^{(t)}(i) & \tilde{\sigma}_{12}^{(t)}(i) & \cdots & \tilde{\sigma}_{1|M_i|}^{(t)}(i) \\ \tilde{\sigma}_{21}^{(t)}(i) & \tilde{\sigma}_{22}^{(t)}(i) & \cdots & \tilde{\sigma}_{2|M_i|}^{(t)}(i) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\sigma}_{|M_i|1}^{(t)}(i) & \tilde{\sigma}_{|M_i|2}^{(t)}(i) & \cdots & \tilde{\sigma}_{|M_i||M_i|}^{(t)}(i)\end{pmatrix}$$

Where,

$$\tilde{\sigma}_{uv}^{(t)}(i) = \Sigma_{M_i}^{(t)}(m_{iu}, m_{iv}) + \mu_{M_i}^{(t)}(m_{iu})\mu_{M_i}^{(t)}(m_{iv})$$

Therefore, $\tau_1^{(t)}$ is vector and $\tau_2^{(t)}$ is matrix and then, the sufficient statistic of X at E-step of some t^{th} iteration, given current parameter $\Theta^{(t)}$ is defined as follows:

$$\begin{aligned}\tau^{(t)} &= \left(\tau_1^{(t)}, \tau_2^{(t)}\right)^T \\ \tau_1^{(t)} &= \left(\bar{x}_1^{(t)}, \bar{x}_2^{(t)}, \dots, \bar{x}_n^{(t)}\right)^T \\ \tau_2^{(t)} &= \begin{pmatrix}s_{11}^{(t)} & s_{12}^{(t)} & \cdots & s_{1n}^{(t)} \\ s_{21}^{(t)} & s_{22}^{(t)} & \cdots & s_{2n}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1}^{(t)} & s_{n2}^{(t)} & \cdots & s_{nn}^{(t)}\end{pmatrix}\end{aligned} \tag{2.39}$$

Each $\bar{x}_j^{(t)}$ is calculated as follows:

$$\bar{x}_j^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}x_{ij} \text{ if } j \notin M_i \\ \mu_{M_i}^{(t)}(j) \text{ if } j \in M_i\end{cases} \tag{2.40}$$

Please see equation 2.35 and equation 2.38 to know $\mu_{M_i}^{(t)}(j)$. Each $s_{uv}^{(t)}$ is calculated as follows:

$$s_{uv}^{(t)} = s_{vu}^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} x_{iu}x_{iv} & \text{if } u \notin M_i \text{ and } v \notin M_i \\ x_{iu}\mu_{M_i}^{(t)}(m_{iv}) & \text{if } u \notin M_i \text{ and } v \in M_i \\ \mu_{M_i}^{(t)}(m_{iu})x_{iv} & \text{if } u \in M_i \text{ and } v \notin M_i \\ \Sigma_{M_i}^{(t)}(m_{iu}, m_{iv}) + \mu_{M_i}^{(t)}(m_{iu})\mu_{M_i}^{(t)}(m_{iv}) & \text{if } u \in M_i \text{ and } v \in M_i \end{cases} \tag{2.41}$$

Equation 2.39 is an instance of equation 2.11, which compose $\tau(X)$ from $\tau(X_{obs})$ and $\tau(X_{mis})$ when $f(X|\Theta)$ distributes normally. Following is the proof of equation 2.41.

If $u \notin M_i$ and $v \notin M_i$ then, the partial statistic $x_{iu}x_{iv}$ is kept intact because x_{iu} and x_{iv} are in X_{obs} are constant with regard to $f(X_{mis} \mid \Theta_{M_i}^{(t)})$ If $u \notin M_i$ and $v \in M_i$ then, the partial statistic $x_{iu}x_{iv}$ is replaced by the expectation $E\left(x_{iu}x_{iv}\middle|\Theta_{M_i}^{(t)}\right)$ as follows:

$$E\left(x_{iu}x_{iv}\middle|\Theta_{M_i}^{(t)}\right) = \int_{X_{mis}} f\left(X_{mis}\middle|\Theta_{M_i}^{(t)}\right)x_{iu}x_{iv}\mathrm{d}X_{mis} = x_{iu}\int_{X_{mis}} f\left(X_{mis}\middle|\Theta_{M_i}^{(t)}\right)x_{iv}\mathrm{d}X_{mis}$$
$$= x_{iu}\mu_{M_i}^{(t)}(m_{iv})$$

If $u \in M_i$ and $v \notin M_i$ then, the partial statistic $x_{iu}x_{iv}$ is replaced by the expectation $E\left(x_{iu}x_{iv}\middle|\Theta_{M_i}^{(t)}\right)$ as follows:

$$E\left(x_{iu}x_{iv}\middle|\Theta_{M_i}^{(t)}\right) = \int_{X_{mis}} f\left(X_{mis}\middle|\Theta_{M_i}^{(t)}\right)x_{iu}x_{iv}\mathrm{d}X_{mis} = x_{iv}\int_{X_{mis}} f\left(X_{mis}\middle|\Theta_{M_i}^{(t)}\right)x_{iu}\mathrm{d}X_{mis}$$
$$= \mu_{M_i}^{(t)}(m_{iu})x_{iv}$$

If $u \in M_i$ and $v \notin M_i$ then, the partial statistic $x_{iu}x_{iv}$ is replaced by the expectation $E\left(x_{iu}x_{iv}\middle|\Theta_{M_i}^{(t)}\right)$ as follows:

$$E\left(x_{iu}x_{iv}\middle|\Theta_{M_i}^{(t)}\right) = \int_{X_{mis}} f\left(X_{mis}\middle|\Theta_{M_i}^{(t)}\right)x_{iu}x_{iv}\mathrm{d}X_{mis} = \Sigma_{M_i}^{(t)}(m_{iu}, m_{iv}) + \mu_{M_i}^{(t)}(m_{iu})\mu_{M_i}^{(t)}(m_{iv})\blacksquare$$

At M-step of some t^{th} iteration, given $\tau^{(t)}$ and $\Theta^{(t)}$, the next parameter $\Theta^{(t+1)} = (\mu^{(t+1)}, \Sigma^{(t+1)})^T$ is a solution of equation 2.23.

$$E(\tau(X)|\Theta) = \tau^{(t)}$$

Due to

$$E(\tau(X)|\Theta) = \begin{pmatrix}\mu \\ \Sigma\end{pmatrix}$$

Equation 2.23 becomes:

$$\begin{cases}\mu = \tau_1^{(t)} \\ \Sigma = \tau_2^{(t)}\end{cases}$$

Which means that

$$\begin{cases}\mu_j^{(t+1)} = \bar{x}_j^{(t)} \\ \sigma_{uv}^{(t+1)} = \sigma_{vu}^{(t+1)} = s_{uv}^{(t)} - \bar{x}_u^{(t)}\bar{x}_v^{(t)}\end{cases} \forall j, u, v \tag{2.42}$$

Please see equation 2.40 and equation 2.41 to know $\bar{x}_j^{(t)}$ and $s_{uv}^{(t)}$.

Moreover, at M-step of some t^{th} iteration, the next parameter $\Phi^{(t+1)} = p^{(t+1)}$ is a maximizer of $Q_2(\Phi \mid \Theta^{(t)})$ given $\Theta^{(t)}$ according to equation 2.24.

$$\Phi^{(t+1)} = \underset{\Phi}{\operatorname{argmin}}\, Q_2\left(\Phi \middle| \Theta^{(t)}\right)$$

Because the PDF of Z_i is:

$$f(Z_i|\Phi) = p^{c(Z_i)}(1-p)^{n-c(Z_i)}$$

The $Q_2(\Phi|\Theta^{(t)})$ becomes:

$$Q_2\left(\Phi \middle| \Theta^{(t)}\right) = \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| \Theta_{M_i}^{(t)}\right) \log\left(f(Z_i|X_{obs}(i), X_{mis}, \Phi)\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \int_{X_{mis}} f\left(X_{mis} \middle| \Theta_{M_i}^{(t)}\right) \log\left(f(Z_i|\Phi)\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \log\left(f(Z_i|\Phi)\right) \int_{X_{mis}} f\left(X_{mis} \middle| \Theta_{M_i}^{(t)}\right) \mathrm{d}X_{mis}$$

$$= \sum_{i=1}^{N} \log\left(f(Z_i|\Phi)\right) = \sum_{i=1}^{N} \log\left(p^{c(Z_i)}(1-p)^{n-c(Z_i)}\right)$$

$$= \sum_{i=1}^{N} \left(c(Z_i)\log(p) + \left(n - c(Z_i)\right)\log(1-p)\right)$$

The next parameter $\Phi^{(t+1)} = p^{(t+1)}$ is solution of the equation created by setting the first-order derivative of $Q_2(\Phi|\Theta^{(t)})$ to be zero, which means that:

$$\frac{\partial Q_2\left(\Phi \middle| \Theta^{(t)}\right)}{\partial p} = \sum_{i=1}^{N} \left(\frac{c(Z_i)}{p} - \frac{n - c(Z_i)}{1-p}\right) = \frac{1}{p(1-p)} \left(\left(\sum_{i=1}^{N} c(Z_i)\right) - npN\right) = 0$$

It is easy to deduce that the next parameter $p^{(t+1)}$ is:

$$p^{(t+1)} = \frac{\sum_{i=1}^{N} c(Z_i)}{nN} \tag{2.43}$$

In general, given sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ whose X_i (s) are iid is MCAR data and $f(X|\Theta)$ is multinormal PDF whereas missingness variable Z follows binomial distribution of n trials, GEM for handling missing data is summarized in table 2.2.

E-step:

Given current parameter $\Theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})^T$, the sufficient statistic $\tau^{(t)}$ is calculated according to equation 2.39, equation 2.40, and equation 2.41.

$$\tau^{(t)} = \left(\tau_1^{(t)}, \tau_2^{(t)}\right)^T$$

$$\tau_1^{(t)} = \left(\bar{x}_1^{(t)}, \bar{x}_2^{(t)}, \ldots, \bar{x}_n^{(t)}\right)^T$$

$$\tau_2^{(t)} = \begin{pmatrix} s_{11}^{(t)} & s_{12}^{(t)} & \cdots & s_{1n}^{(t)} \\ s_{21}^{(t)} & s_{22}^{(t)} & \cdots & s_{2n}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1}^{(t)} & s_{n2}^{(t)} & \cdots & s_{nn}^{(t)} \end{pmatrix}$$

$$s_{uv}^{(t)} = s_{vu}^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} x_{iu}x_{iv} & \text{if } u \notin M_i \text{ and } v \notin M_i \\ x_{iu}\mu_{M_i}^{(t)}(m_{iv}) & \text{if } u \notin M_i \text{ and } v \in M_i \\ \mu_{M_i}^{(t)}(m_{iu})x_{iv} & \text{if } u \in M_i \text{ and } v \notin M_i \\ \Sigma_{M_i}^{(t)}(m_{iu}, m_{iv}) + \mu_{M_i}^{(t)}(m_{iu})\mu_{M_i}^{(t)}(m_{iv}) & \text{if } u \in M_i \text{ and } v \in M_i \end{cases}$$

Where μ_{M_i} and Σ_{M_i} are specified in equation 2.35 and equation 2.38.

M-step:

Given $\tau^{(t)}$ and $\Theta^{(t)}$, the next parameter $\Theta^{(t+1)} = (\mu^{(t+1)}, \Sigma^{(t+1)})^T$ is specified by equation 2.42.

$$\begin{cases} \mu_j^{(t+1)} = \bar{x}_j^{(t)} \\ \sigma_{uv}^{(t+1)} = \sigma_{vu}^{(t+1)} = s_{uv}^{(t)} - x_u^{(t)}x_v^{(t)} \end{cases} \forall j, u, v$$

Given $\Theta^{(t)}$, the next parameter $\Phi^{(t+1)} = p^{(t+1)}$ is specified by equation 2.43.

$$p^{(t+1)} = \frac{\sum_{i=1}^{N} c(Z_i)}{nN}$$

Where $c(Z_i)$ is the number of z_{ij} (s) in Z_i that equal 1.

Table 2.2. E-step and M-step of GEM algorithm for handling missing data given normal PDF

As aforementioned, an interesting application of handling missing data is to fill in or predict missing values. For instance, suppose the estimate resulted from GEM is $\Theta^* = (\mu^*, \Sigma^*)^T$, missing part $X_{mis} = \left(x_{m_1}, x_{m_2}, \ldots, x_{m_{|M_i|}}\right)^T$ is replaced by μ_M^* as follows:

$$x_{m_j} = \mu_M^*(m_j), \forall m_j \in M \tag{2.44}$$

Note, μ_M^* which is extracted from μ^* is estimated mean of the conditional PDF of X_{mis} (given X_{obs}) according to equation 2.35. Moreover, $\mu_M^*(m_j)$ is estimated partial mean of x_{m_j} given the conditional PDF $f(X_{mis} \mid \Theta_M^*)$, please see equation 2.38 for more details about μ_M^*. As aforementioned, in practice we can stop GEM after its first iteration was done, which is reasonable enough to handle missing data.

Now we survey another interesting case that sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ whose X_i (s) are iid is MCAR data and $f(X|\Theta)$ is multinomial PDF of K trials. We ignore missingness variable Z here because it is included in the case of multinormal PDF. Let $X = \{X_{obs}, X_{mis}\}$ be random variable representing every X_i. Suppose dimension of X is n. According to equation 2.9, recall that

$$X_i = \{X_{obs}(i), X_{mis}(i)\} = (x_{i1}, x_{i2}, \ldots, x_{in})^T$$
$$X_{mis}(i) = \left(x_{im_1}, x_{im_2}, \ldots, x_{im_{|M_i|}}\right)^T$$
$$X_{obs}(i) = \left(x_{i\bar{m}_{i1}}, x_{i\bar{m}_{i2}}, \ldots, x_{i\bar{m}_{i|\bar{M}_i|}}\right)^T$$
$$M_i = \left\{m_{i1}, m_{i2}, \ldots, m_{i|M_i|}\right\}$$
$$\bar{M}_i = \left\{\bar{m}_{i1}, \bar{m}_{i2}, \ldots, \bar{m}_{i|\bar{M}_i|}\right\}$$

The PDF of X is:

$$f(X|\Theta) = f(X_{obs}, X_{mis}|\Theta) = \frac{K!}{\prod_{j=1}^{n}(x_j!)}\prod_{j=1}^{n} p_j^{x_j} \tag{2.45}$$

Where x_j are integers and $\Theta = (p_1, p_2,\ldots, p_n)^T$ is the set of probabilities such that

$$\sum_{j=1}^{n} p_j = 1$$

$$\sum_{j=1}^{n} x_j = K$$

$$x_j \in \{0,1,\ldots,K\}$$

Note, x_j is the number of trials generating nominal value j. Therefore,

$$f(X_i|\Theta) = f(X_{obs}(i), X_{mis}(i)|\Theta) = \frac{K!}{\prod_{j=1}^{n}(x_{ij}!)}\prod_{j=1}^{n} p_j^{x_{ij}}$$

Where,

$$\sum_{j=1}^{n} x_{ij} = K$$

$$x_{ij} \in \{0,1,\ldots,K\}$$

The most important task here is to define equation 2.11 and equation 2.15 in order to compose $\tau(X)$ from $\tau(X_{obs})$, $\tau(X_{mis})$ and to extract Θ_M from Θ when $f(X|\Theta)$ is multinomial PDF.

Let Θ_{mis} be parameter of marginal PDF of X_{mis}, we have:

$$f(X_{mis}|\Theta_{mis}) = \frac{K_{mis}!}{\prod_{m_j \in M}\left(x_{m_j}!\right)}\prod_{j=1}^{|M|}\left(\frac{p_{m_j}}{P_{mis}}\right)^{x_{m_j}} \tag{2.46}$$

Therefore,

$$f\left(X_{mis}(i)\middle|\Theta_{mis}(i)\right) = \frac{K_{mis}(i)!}{\prod_{m_j \in M_i}\left(x_{im_j}!\right)}\prod_{j=1}^{|M_i|}\left(\frac{p_{m_{ij}}}{P_{mis}(i)}\right)^{x_{im_j}}$$

Where,

$$\Theta_{mis}(i) = \left(\frac{p_{m_{i1}}}{P_{mis}(i)}, \frac{p_{m_{i2}}}{P_{mis}(i)}, \ldots, \frac{p_{m_{i|M_i|}}}{P_{mis}(i)}\right)^T$$

$$P_{mis}(i) = \sum_{j=1}^{|M_i|} p_{m_{ij}} \tag{2.47}$$

$$K_{mis}(i) = \sum_{j=1}^{|M_i|} x_{m_{ij}}$$

Obviously, $\Theta_{mis}(i)$ is extracted from Θ given indicator M_i.

Let Θ_{obs} be parameter of marginal PDF of X_{obs}, we have:

$$f(X_{obs}|\Theta_{obs}) = \frac{K_{obs}!}{\prod_{\bar{m}_j \in \bar{M}}\left(x_{\bar{m}_j}!\right)}\prod_{j=1}^{|\bar{M}|}\left(\frac{p_{\bar{m}_j}}{P_{obs}}\right)^{x_{\bar{m}_j}} \tag{2.48}$$

Therefore,

$$f\left(X_{obs}(i)\middle|\Theta_{obs}(i)\right)=\frac{K_{obs}(i)!}{\prod_{\bar{m}_j\in\bar{M}_i}\left(x_{i\bar{m}_j}!\right)}\prod_{j=1}^{|\bar{M}_i|}\left(\frac{p_{\bar{m}_{ij}}}{P_{obs}(i)}\right)^{x_{i\bar{m}_j}}$$

Where,

$$\Theta_{obs}(i)=\left(\frac{p_{\bar{m}_{i1}}}{P_{obs}(i)},\frac{p_{\bar{m}_{i2}}}{P_{obs}(i)},\dots,\frac{p_{\bar{m}_{i|\bar{M}_i|}}}{P_{obs}(i)}\right)^T$$

$$P_{obs}(i)=\sum_{j=1}^{|\bar{M}_i|}p_{\bar{m}_{ij}} \tag{2.49}$$

$$K_{obs}(i)=\sum_{j=1}^{|\bar{M}_i|}x_{\bar{m}_{ij}}$$

Obviously, $\Theta_{obs}(i)$ is extracted from Θ given indicator $\bar{M}_i$ or M_i.

The conditional PDF of X_{mis} given X_{obs} is calculated based on the PDF of X and the marginal PDF of X_{obs} as follows:

$$f(X_{mis}|\Theta_M)=f(X_{mis}|X_{obs},\Theta)=\frac{f(X_{obs},X_{mis}|\Theta)}{f(X_{obs}|\Theta_{obs})}$$

$$=\frac{\frac{K!}{\prod_{j=1}^{n}(x_j!)}\prod_{j=1}^{n}p_j^{x_j}}{\frac{K_{obs}!}{\prod_{j=1}^{|\bar{M}|}x_{\bar{m}_j}!}\prod_{j=1}^{|\bar{M}|}\left(\frac{p_{\bar{m}_j}}{P_{obs}}\right)^{x_{\bar{m}_j}}}$$

$$=\frac{K!}{K_{obs}!}\frac{\prod_{j=1}^{|\bar{M}|}x_{\bar{m}_j}!}{\prod_{j=1}^{n}(x_j!)}*\frac{\prod_{j=1}^{n}p_j^{x_j}}{\prod_{j=1}^{|\bar{M}|}\left(\frac{p_{\bar{m}_j}}{P_{obs}}\right)^{x_{\bar{m}_j}}}$$

$$=\frac{K!}{K_{obs}!\prod_{j=1}^{|M|}\left(x_{m_j}!\right)}*\left(\prod_{j=1}^{|M|}p_{m_j}^{x_{m_j}}\right)*\left(\prod_{j=1}^{|\bar{M}|}p_{\bar{m}_j}^{x_{\bar{m}_j}}\left(\frac{P_{obs}}{p_{\bar{m}_j}}\right)^{x_{\bar{m}_j}}\right)$$

$$=\frac{K!}{K_{obs}!\prod_{j=1}^{|M|}\left(x_{m_j}!\right)}*\left(\prod_{j=1}^{|M|}p_{m_j}^{x_{m_j}}\right)*\left(\prod_{j=1}^{|\bar{M}|}(P_{obs})^{x_{\bar{m}_j}}\right)$$

$$=\frac{K!}{K_{obs}!\prod_{j=1}^{|M|}\left(x_{m_j}!\right)}*\left(\prod_{j=1}^{|M|}p_{m_j}^{x_{m_j}}\right)*((P_{obs})^{K_{obs}})$$

This implies that the conditional PDF of X_{mis} given X_{obs} is multinomial PDF of K trials.

$$f(X_{mis}|X_{obs},\Theta_M)=f(X_{mis}|X_{obs},\Theta)$$
$$=\frac{K!}{K_{obs}!\prod_{j=1}^{|M|}\left(x_{m_j}!\right)}*\left(\prod_{j=1}^{|M|}p_{m_j}^{x_{m_j}}\right)*((P_{obs})^{K_{obs}}) \tag{2.50}$$

Therefore,

$$f\left(X_{mis}(i)\middle|X_{obs}(i),\Theta_{M_i}\right)=f(X_{mis}(i)|X_{obs}(i),\Theta)$$
$$=\frac{K!}{K_{obs}(i)!\prod_{j=1}^{|M_i|}\left(x_{im_j}!\right)}*\left(\prod_{j=1}^{|M_i|}p_{m_{ij}}^{x_{im_j}}\right)*\left(\left(P_{obs}(i)\right)^{K_{obs}(i)}\right)$$

Where

$$P_{obs}(i) = \sum_{j=1}^{|\bar{M}_i|} p_{\bar{m}_{ij}}$$

$$K_{obs}(i) = \sum_{j=1}^{|\bar{M}_i|} x_{\bar{m}_{ij}}$$

Obviously, the parameter Θ_{M_i} of the conditional PDF $f\left(X_{mis}(i)\middle|X_{obs}(i), \Theta_{M_i}\right)$ is:

$$\Theta_{M_i} = u\left(\Theta, X_{obs}(i)\right) = \begin{pmatrix} p_{m_1} \\ p_{m_2} \\ \vdots \\ p_{m_k} \\ P_{obs}(i) = \sum_{j=1}^{|\bar{M}_i|} p_{\bar{m}_{ij}} \end{pmatrix} \tag{2.51}$$

Therefore, equation 2.51 to extract Θ_{M_i} from Θ given $X_{obs}(i)$ is an instance of equation 2.15. It is easy to check that

$$\sum_{j=1}^{|M_i|} x_{m_{ij}} + K_{obs}(i) = K_{mis}(i) + K_{obs}(i) = K$$

$$\sum_{j=1}^{|M_i|} p_{m_{ij}} + P_{obs}(i) = \sum_{j=1}^{|M_i|} p_{m_{ij}} + \sum_{j=1}^{|\bar{M}_i|} p_{\bar{m}_{ij}} = \sum_{j=1}^{n} p_j = 1$$

At E-step of some t^{th} iteration, given current parameter $\Theta^{(t)} = (p_1^{(t)}, p_2^{(t)}, \ldots, p_n^{(t)})^T$, the sufficient statistic of X is calculated according to equation 2.22. Let,

$$\tau^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}^{(t)}\right)\right\}$$

The sufficient statistic of $X_{obs}(i)$ is:

$$\tau\left(X_{obs}(i)\right) = \left(x_{i\bar{m}_1}, x_{i\bar{m}_2}, \ldots, x_{i\bar{m}_{|\bar{M}_i|}}\right)^T$$

The sufficient statistic of $X_{mis}(i)$ with regard to $f\left(X_{mis}(i)\middle|X_{obs}(i), \Theta_{M_i}\right)$ is:

$$\tau\left(X_{mis}(i)\right) = \begin{pmatrix} x_{im_1} \\ x_{im_2} \\ \vdots \\ x_{im_{|M_i|}} \\ \sum_{j=1}^{|\bar{M}_i|} x_{\bar{m}_{ij}} \end{pmatrix}$$

We also have:

$$E\left(\tau(X_{mis})\middle|\Theta_{M_i}^{(t)}\right) = \int_{X_{mis}} f\left(X_{mis}\middle|X_{obs}, \Theta_{M_i}^{(t)}\right)\tau(X_{mis})\mathrm{d}X_{mis} = \begin{pmatrix} Kp_{m_1} \\ Kp_{m_2} \\ \vdots \\ Kp_{m_{|M_i|}} \\ \sum_{j=1}^{|\bar{M}_i|} Kp_{\bar{m}_{ij}} \end{pmatrix}$$

Therefore, the sufficient statistic of X at E-step of some t^{th} iteration given current parameter $\Theta^{(t)} = (p_1^{(t)}, p_2^{(t)}, \ldots, p_n^{(t)})^T$ is defined as follows:

$$\tau^{(t)} = \left(\bar{x}_1^{(t)}, \bar{x}_2^{(t)}, \ldots, \bar{x}_n^{(t)}\right)^T$$
$$\bar{x}_j^{(t)} = \frac{1}{N}\sum_{i=1}^{N} \begin{cases} x_{ij} \text{ if } j \notin M_i \\ K p_j^{(t)} \text{ if } j \in M_i \end{cases} \forall j \quad (2.52)$$

Equation 2.52 is an instance of equation 2.11, which compose $\tau(X)$ from $\tau(X_{obs})$ and $\tau(X_{mis})$ when $f(X|\Theta)$ is multinomial PDF.

At M-step of some t^{th} iteration, we need to maximize $Q_1(\Theta'|\Theta)$ with following constraint

$$\sum_{j=1}^{n} p_j = 1$$

According to equation 2.19, we have:

$$Q_1(\Theta'|\Theta) = \sum_{i=1}^{N} E\left(\log\left(b(X_{obs}(i), X_{mis})\right)\middle|\Theta_{M_i}\right) + (\Theta')^T \sum_{i=1}^{N}\left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}\right)\right\} - N\log\left(a(\Theta')\right)$$

Where quantities $b(X_{obs}(i), X_{mis})$ and $a(\Theta')$ belongs to the PDF $f(X|\Theta)$ of X. Because there is the constraint $\sum_{j=1}^{n} p_j = 1$, we use Lagrange duality method to maximize $Q_1(\Theta'|\Theta)$. The Lagrange function $la(\Theta', \lambda \mid \Theta)$ is sum of $Q_1(\Theta'|\Theta)$ and the constraint $\sum_{j=1}^{n} p_j = 1$, as follows:

$$la(\Theta', \lambda|\Theta) = Q_1(\Theta'|\Theta) + \lambda\left(1 - \sum_{j=1}^{n} p_j'\right)$$
$$= \sum_{i=1}^{N} E\left(\log\left(b(X_{obs}(i), X_{mis})\right)\middle|\Theta_{M_i}\right)$$
$$+ (\Theta')^T \sum_{i=1}^{N}\left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}\right)\right\} - N\log\left(a(\Theta')\right) + \lambda\left(1 - \sum_{j=1}^{n} p_j'\right)$$

Where $\Theta' = (p_1', p_2', \ldots, p_n')^T$. Note, $\lambda \geq 0$ is called Lagrange multiplier. Of course, $la(\Theta', \lambda \mid \Theta)$ is function of Θ' and λ. The next parameter $\Theta^{(t+1)}$ that maximizes $Q_1(\Theta'|\Theta)$ is solution of the equation formed by setting the first-order derivative of Lagrange function regarding Θ' and λ to be zero.

The first-order partial derivative of $la(\Theta', \lambda \mid \Theta)$ with regard to Θ' is:

$$\frac{\partial la(\Theta', \lambda|\Theta)}{\partial \Theta'} = \sum_{i=1}^{N}\left(E\left(\tau(X_{obs}(i), X_{mis})\middle|\Theta_{M_i}\right)\right)^T - N\log'\left(a(\Theta')\right)$$
$$= \sum_{i=1}^{N}\left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}\right)\right\}^T - N\log'\left(a(\Theta')\right) - (\lambda, \lambda, \ldots, \lambda)^T$$

By referring table 1.2, we have:

$$\log'\left(a(\Theta')\right) = \left(E(\tau(X)|\Theta')\right)^T = \int_X f(X|\Theta)\left(\tau(X)\right)^T \mathrm{d}X$$

Thus,

$$\frac{\partial la(\Theta', \lambda|\Theta)}{\partial \Theta'} = \sum_{i=1}^{N}\left\{\tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis})\middle|\Theta_{M_i}\right)\right\}^T - N\left(E(\tau(X)|\Theta')\right)^T - (\lambda, \lambda, \ldots, \lambda)^T$$

The first-order partial derivative of $la(\Theta', \lambda \mid \Theta)$ with regard to λ is:

$$\frac{\partial la(\Theta', \lambda|\Theta)}{\partial \lambda} = 1 - \sum_{j=1}^{n} p_j'$$

Therefore, at M-step of some t^{th} iteration, given current parameter $\Theta^{(t)} = (p_1^{(t)}, p_2^{(t)}, \ldots, p_n^{(t)})^T$, the next parameter $\Theta^{(t+1)}$ is solution of the following equation:

$$\begin{cases} \sum_{i=1}^{N} \left\{ \tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis}) \middle| \Theta_{M_i}^{(t)}\right) \right\}^T \\ -N\left(E(\tau(X)|\Theta)\right)^T - (\lambda, \lambda, \ldots, \lambda) = \mathbf{0}^T \\ 1 - \sum_{j=1}^{n} p_j = 0 \end{cases}$$

This implies:

$$\begin{cases} E(\tau(X)|\Theta) = \tau^{(t)} - \begin{pmatrix} \lambda/N \\ \lambda/N \\ \lambda/N \\ \lambda/N \end{pmatrix} \\ \sum_{j=1}^{n} p_j = 1 \end{cases}$$

Where,

$$\tau^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \left\{ \tau\left(X_{obs}(i)\right), E\left(\tau(X_{mis}) \middle| \Theta_{M_i}^{(t)}\right) \right\}$$

Due to

$$E(\tau(X)|\Theta) = \int_X \tau(X) f(X|\Theta) \mathrm{d}X = (Kp_1, Kp_2, \ldots, Kp_n)^T$$

$$\tau^{(t)} = \left(\bar{x}_1^{(t)}, \bar{x}_2^{(t)}, \ldots, \bar{x}_n^{(t)}\right)^T$$

$$\bar{x}_j^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} x_{ij} \text{ if } j \notin M_i \\ Kp_j^{(t)} \text{ if } j \in M_i \end{cases} \forall j$$

We obtain n equations $Kp_j = -\lambda/N + \bar{x}_j^{(t)}$ and 1 constraint $\sum_{j=1}^{n} p_j = 1$. Therefore, we have:

$$p_j = -\frac{\lambda}{KN} + \frac{1}{KN} \sum_{i=1}^{N} \begin{cases} x_{ij} \text{ if } j \notin M_i \\ Kp_j^{(t)} \text{ if } j \in M_i \end{cases} \forall j$$

Summing n equations above, we have:

$$1 = \sum_{j=1}^{n} p_j = -\frac{\lambda}{KN} + \frac{1}{KN} \sum_{j=1}^{n} \left(\sum_{i=1}^{N} \begin{cases} x_{ij} \text{ if } j \notin M_i \\ Kp_j^{(t)} \text{ if } j \in M_i \end{cases} \right)$$

$$= -\frac{\lambda}{KN} + \frac{1}{KN} \sum_{i=1}^{N} \left(\sum_{j=1}^{|\bar{M}_i|} x_{i\bar{m}_j} + \sum_{j=1}^{|M_i|} Kp_{m_j}^{(t)} \right)$$

Suppose every missing value x_{im_j} is estimated by Kp_{m_j} such that:

$$\sum_{j=1}^{|\bar{M}_i|} x_{\bar{m}_{ij}} = \sum_{j=1}^{|M_i|} Kp_{m_j}^{(t)}$$

We obtain:

$$1 = -\frac{\lambda}{KN} + \frac{1}{KN}\sum_{i=1}^{N}\left(\sum_{j=1}^{|\bar{M}_i|} x_{i\bar{m}_j} + \sum_{j=1}^{|M_i|} x_{im_j}\right) = -\frac{\lambda}{KN} + \frac{1}{KN}\sum_{i=1}^{N} K = -\frac{\lambda}{KN} + 1$$

This implies

$$\lambda = 0$$

Such that

$$p_j = \frac{1}{KN}\sum_{i=1}^{N}\begin{cases} x_{ij} \text{ if } j \notin M_i \\ Kp_j^{(t)} \text{ if } j \in M_i \end{cases} \forall j$$

Therefore, at M-step of some t^{th} iteration, given current parameter $\Theta^{(t)} = (p_1^{(t)}, p_2^{(t)},\ldots, p_n^{(t)})^T$, the next parameter $\Theta^{(t+1)}$ is specified by following equation.

$$p_j^{(t+1)} = \frac{1}{KN}\sum_{i=1}^{N}\begin{cases} x_{ij} \text{ if } j \notin M_i \\ Kp_j^{(t)} \text{ if } j \in M_i \end{cases} \forall j \quad (2.53)$$

In general, given sample $\mathcal{X} = \{X_1, X_2,\ldots, X_N\}$ whose X_i (s) are iid is MCAR data and $f(X|\Theta)$ is multinomial PDF of K trials, GEM for handling missing data is summarized in table 2.3.

M-step:

Given $\tau^{(t)}$ and $\Theta^{(t)} = (p_1^{(t)}, p_2^{(t)},\ldots, p_n^{(t)})^T$, the next parameter $\Theta^{(t+1)}$ is specified by equation 2.53.

$$p_j^{(t+1)} = \frac{1}{KN}\sum_{i=1}^{N}\begin{cases} x_{ij} \text{ if } j \notin M_i \\ Kp_j^{(t)} \text{ if } j \in M_i \end{cases} \forall j$$

Table 2.3. E-step and M-step of GEM algorithm for handling missing data given multinomial PDF

In table 2.3, E-step is implied in how to perform M-step. As aforementioned, in practice we can stop GEM after its first iteration was done, which is reasonable enough to handle missing data. Next section includes two examples of handling missing data with multinormal distribution and multinomial distribution.

3. Numerical examples

It is necessary to have an example for illustrating how to handle missing data with multinormal PDF.

Example 3.1. Given sample of size two, $\mathcal{X} = \{X_1, X_2\}$ in which $X_1 = (x_{11}=1, x_{12}=?, x_{13}=3, x_{14}=?)^T$ and $X_2 = (x_{21}=?, x_{22}=2, x_{23}=?, x_{24}=4)^T$ are iid. Therefore, we also have $Z_1 = (z_{11}=0, z_{12}=1, z_{13}=0, z_{14}=1)^T$ and $Z_2 = (z_{21}=1, z_{22}=0, z_{23}=1, z_{24}=0)^T$. All Z_i (s) are iid too.

	x_1	x_2	x_3	x_4
X_1	1	?	3	?
X_2	?	2	?	4

	z_1	z_2	z_3	z_4
Z_1	0	1	0	1
Z_2	1	0	1	0

Of course, we have $X_{obs}(1) = (x_{11}=1, x_{13}=3)^T$, $X_{mis}(1) = (x_{12}=?, x_{14}=?)^T$, $X_{obs}(2) = (x_{22}=2, x_{24}=4)^T$ and $X_{mis}(2) = (x_{21}=?, x_{23}=?)^T$. We also have $M_1 = \{m_{11}=2, m_{12}=4\}$, $\bar{M}_1 = \{\bar{m}_{11}=1, \bar{m}_{12}=3\}$, $M_2 = \{m_{21}=1, m_{22}=3\}$, and $\bar{M}_2 = \{\bar{m}_{21}=2, \bar{m}_{22}=4\}$. Let X and Z be random variables representing every X_i and every Z_i, respectively. Suppose $f(X|\Theta)$ is multinormal PDF and missingness variable Z follows binomial distribution of 4 trials according to equation 2.26 and equation 2.27. Dimension of X is 4. We will estimate $\Theta = (\mu, \Sigma)^T$ and $\Phi = p$ based on $\mathcal{X}$.

$$\mu = (\mu_1, \mu_2, \mu_3, \mu_4)^T$$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix}$$

The parameters μ and Σ are initialized arbitrarily as zero vector and identity vector whereas p is initialized 0.5 as follows:

$$\mu^{(1)} = \left(\mu_1^{(1)} = 0, \mu_2^{(1)} = 0, \mu_3^{(1)} = 0, \mu_4^{(1)} = 0\right)^T$$

$$\Sigma^{(1)} = \begin{pmatrix} \sigma_{11}^{(1)} = 1 & \sigma_{12}^{(1)} = 0 & \sigma_{13}^{(1)} = 0 & \sigma_{14}^{(1)} = 0 \\ \sigma_{21}^{(1)} = 0 & \sigma_{22}^{(1)} = 1 & \sigma_{23}^{(1)} = 0 & \sigma_{24}^{(1)} = 0 \\ \sigma_{31}^{(1)} = 0 & \sigma_{32}^{(1)} = 0 & \sigma_{33}^{(1)} = 1 & \sigma_{34}^{(1)} = 0 \\ \sigma_{41}^{(1)} = 0 & \sigma_{42}^{(1)} = 0 & \sigma_{43}^{(1)} = 0 & \sigma_{44}^{(1)} = 1 \end{pmatrix}$$

$$p^{(1)} = 0.5$$

At 1st iteration, E-step, we have:

$$X_{obs}(1) = (x_1 = 1, x_3 = 3)^T$$

$$\mu_{mis}(1) = \left(\mu_2^{(1)} = 0, \mu_4^{(1)} = 0\right)^T$$

$$\Sigma_{mis}(1) = \begin{pmatrix} \sigma_{22}^{(1)} = 1 & \sigma_{24}^{(1)} = 0 \\ \sigma_{42}^{(1)} = 0 & \sigma_{44}^{(1)} = 1 \end{pmatrix}$$

$$\mu_{obs}(1) = \left(\mu_1^{(1)} = 0, \mu_3^{(1)} = 0\right)^T$$

$$\Sigma_{obs}(1) = \begin{pmatrix} \sigma_{11}^{(1)} = 1 & \sigma_{13}^{(1)} = 0 \\ \sigma_{31}^{(1)} = 0 & \sigma_{33}^{(1)} = 1 \end{pmatrix}$$

$$V_{obs}^{mis}(1) = \begin{pmatrix} \sigma_{21}^{(1)} = 0 & \sigma_{23}^{(1)} = 0 \\ \sigma_{41}^{(1)} = 0 & \sigma_{43}^{(1)} = 0 \end{pmatrix}$$

$$V_{mis}^{obs}(1) = \begin{pmatrix} \sigma_{12}^{(1)} = 0 & \sigma_{14}^{(1)} = 0 \\ \sigma_{32}^{(1)} = 0 & \sigma_{34}^{(1)} = 0 \end{pmatrix}$$

$$\mu_{M_1}^{(1)} = \mu_{mis}(1) + \left(V_{obs}^{mis}(1)\right)\left(\Sigma_{obs}(1)\right)^{-1}\left(X_{obs}(1) - \mu_{obs}(1)\right)$$

$$= \left(\mu_{M_1}^{(1)}(2) = 0, \mu_{M_1}^{(1)}(4) = 0\right)^T$$

$$\Sigma_{M_1}^{(1)} = \Sigma_{mis}(1) - \left(V_{obs}^{mis}(1)\right)\left(\Sigma_{obs}(1)\right)^{-1}\left(V_{mis}^{obs}\right) = \begin{pmatrix} \Sigma_{M_1}^{(1)}(2,2) = 1 & \Sigma_{M_1}^{(1)}(2,4) = 0 \\ \Sigma_{M_1}^{(1)}(4,2) = 0 & \Sigma_{M_1}^{(1)}(4,4) = 1 \end{pmatrix}$$

$$X_{obs}(2) = (x_2 = 2, x_4 = 4)^T$$

$$\mu_{mis}(2) = \left(\mu_1^{(1)} = 0, \mu_3^{(1)} = 0\right)^T$$

$$\Sigma_{mis}(2) = \begin{pmatrix} \sigma_{11}^{(1)} = 1 & \sigma_{13}^{(1)} = 0 \\ \sigma_{31}^{(1)} = 0 & \sigma_{33}^{(1)} = 1 \end{pmatrix}$$

$$\mu_{obs}(2) = \left(\mu_2^{(1)} = 0, \mu_4^{(1)} = 0\right)^T$$

$$\Sigma_{obs}(2) = \begin{pmatrix} \sigma_{22}^{(1)} = 1 & \sigma_{24}^{(1)} = 0 \\ \sigma_{42}^{(1)} = 0 & \sigma_{44}^{(1)} = 1 \end{pmatrix}$$

$$V_{obs}^{mis}(2) = \begin{pmatrix} \sigma_{12}^{(1)} = 0 & \sigma_{14}^{(1)} = 0 \\ \sigma_{32}^{(1)} = 0 & \sigma_{34}^{(1)} = 0 \end{pmatrix}$$

$$V_{mis}^{obs}(2) = \begin{pmatrix} \sigma_{21}^{(1)} = 0 & \sigma_{23}^{(1)} = 0 \\ \sigma_{41}^{(1)} = 0 & \sigma_{43}^{(1)} = 0 \end{pmatrix}$$

$$\mu_{M_2}^{(1)} = \mu_{mis}(2) + \left(V_{obs}^{mis}(2)\right)\left(\Sigma_{obs}(2)\right)^{-1}\left(X_{obs}(2) - \mu_{obs}(2)\right)$$
$$= \left(\mu_{M_2}^{(1)}(1) = 0, \mu_{M_2}^{(1)}(3) = 0\right)^T$$

$$\Sigma_{M_2}^{(1)} = \Sigma_{mis}(2) - \left(V_{obs}^{mis}(2)\right)\left(\Sigma_{obs}(2)\right)^{-1}\left(V_{mis}^{obs}\right) = \begin{pmatrix} \Sigma_{M_2}^{(1)}(1,1) = 1 & \Sigma_{M_2}^{(1)}(1,3) = 0 \\ \Sigma_{M_2}^{(1)}(3,1) = 0 & \Sigma_{M_2}^{(1)}(3,3) = 1 \end{pmatrix}$$

$$\bar{x}_1^{(1)} = \frac{1}{2}\left(x_{11} + \mu_{M_2}^{(1)}(1)\right) = 0.5$$
$$\bar{x}_2^{(1)} = \frac{1}{2}\left(\mu_{M_1}^{(1)}(2) + x_{22}\right) = 1$$
$$\bar{x}_3^{(1)} = \frac{1}{2}\left(x_{13} + \mu_{M_2}^{(1)}(3)\right) = 1.5$$
$$\bar{x}_4^{(1)} = \frac{1}{2}\left(\mu_{M_1}^{(1)}(4) + x_{24}\right) = 2$$

$$s_{11}^{(1)} = \frac{1}{2}\left((x_{11})^2 + \left(\Sigma_{M_2}^{(1)}(1,1) + \left(\mu_{M_2}^{(1)}(1)\right)^2\right)\right) = 1$$
$$s_{12}^{(1)} = s_{21}^{(1)} = \frac{1}{2}\left(x_{11}\mu_{M_1}^{(1)}(2) + \mu_{M_2}^{(1)}(1)x_{22}\right) = 0$$
$$s_{13}^{(1)} = s_{31}^{(1)} = \frac{1}{2}\left(x_{11}x_{13} + \left(\Sigma_{M_2}^{(1)}(1,3) + \mu_{M_2}^{(1)}(1)\mu_{M_2}^{(1)}(3)\right)\right) = 1.5$$
$$s_{14}^{(1)} = s_{41}^{(1)} = \frac{1}{2}\left(x_{11}\mu_{M_1}^{(1)}(4) + \mu_{M_2}^{(1)}(1)x_{24}\right) = 0$$
$$s_{22}^{(1)} = \frac{1}{2}\left(\left(\Sigma_{M_1}^{(1)}(2,2) + \left(\mu_{M_1}^{(1)}(2)\right)^2\right) + (x_{22})^2\right) = 2.5$$
$$s_{23}^{(1)} = s_{32}^{(1)} = \frac{1}{2}\left(\mu_{M_1}^{(1)}(2)x_{13} + x_{22}\mu_{M_2}^{(1)}(3)\right) = 0$$
$$s_{24}^{(1)} = s_{42}^{(1)} = \frac{1}{2}\left(\left(\Sigma_{M_1}^{(1)}(2,4) + \mu_{M_1}^{(1)}(2)\mu_{M_1}^{(1)}(4)\right) + x_{22}x_{24}\right) = 4$$
$$s_{33}^{(1)} = \frac{1}{2}\left((x_{13})^2 + \left(\Sigma_{M_2}^{(1)}(3,3) + \left(\mu_{M_2}^{(1)}(3)\right)^2\right)\right) = 5$$
$$s_{34}^{(1)} = s_{43}^{(1)} = \frac{1}{2}\left(x_{13}\mu_{M_1}^{(1)}(4) + \mu_{M_2}^{(1)}(3)x_{24}\right) = 0$$
$$s_{44}^{(1)} = \frac{1}{2}\left(\left(\Sigma_{M_1}^{(1)}(4,4) + \left(\mu_{M_1}^{(1)}(4)\right)^2\right) + (x_{24})^2\right) = 8.5$$

At 1st iteration, M-step, we have:

$$\mu_1^{(2)} = \bar{x}_1^{(1)} = 0.5$$
$$\mu_2^{(2)} = \bar{x}_2^{(1)} = 1$$
$$\mu_3^{(2)} = \bar{x}_3^{(1)} = 1.5$$
$$\mu_4^{(2)} = \bar{x}_4^{(1)} = 2$$

$$\sigma_{11}^{(2)} = s_{11}^{(1)} - \left(\bar{x}_1^{(1)}\right)^2 = 0.75$$
$$\sigma_{12}^{(2)} = \sigma_{21}^{(2)} = s_{12}^{(1)} - \bar{x}_1^{(1)}\bar{x}_2^{(1)} = -0.5$$

$$\sigma_{13}^{(2)} = \sigma_{31}^{(2)} = s_{13}^{(1)} - \bar{x}_1^{(1)}\bar{x}_3^{(1)} = 0.75$$
$$\sigma_{14}^{(2)} = \sigma_{41}^{(2)} = s_{14}^{(1)} - \bar{x}_1^{(1)}\bar{x}_4^{(1)} = -1$$
$$\sigma_{22}^{(2)} = s_{22}^{(1)} - \left(\bar{x}_2^{(1)}\right)^2 = 1.5$$
$$\sigma_{23}^{(2)} = \sigma_{32}^{(2)} = s_{23}^{(1)} - \bar{x}_2^{(1)}\bar{x}_3^{(1)} = -1.5$$
$$\sigma_{24}^{(2)} = \sigma_{42}^{(2)} = s_{24}^{(1)} - \bar{x}_2^{(1)}\bar{x}_4^{(1)} = 2$$
$$\sigma_{33}^{(2)} = s_{33}^{(1)} - \left(\bar{x}_3^{(1)}\right)^2 = 2.75$$
$$\sigma_{34}^{(2)} = \sigma_{43}^{(2)} = s_{34}^{(1)} - \bar{x}_3^{(1)}\bar{x}_4^{(1)} = -3$$
$$\sigma_{44}^{(2)} = s_{44}^{(1)} - \left(\bar{x}_4^{(1)}\right)^2 = 4.5$$

$$p^{(2)} = \frac{c(Z_1) + c(Z_2)}{4 * 2} = \frac{2+2}{4*2} = 0.5$$

At 2nd iteration, E-step, we have:

$$X_{obs}(1) = (x_1 = 1, x_3 = 3)^T$$
$$\mu_{mis}(1) = \left(\mu_2^{(2)} = 1, \mu_4^{(2)} = 2\right)^T$$
$$\Sigma_{mis}(1) = \begin{pmatrix} \sigma_{22}^{(2)} = 1.5 & \sigma_{24}^{(2)} = 2 \\ \sigma_{42}^{(2)} = 2 & \sigma_{44}^{(2)} = 4.5 \end{pmatrix}$$
$$\mu_{obs}(1) = \left(\mu_1^{(2)} = 0.5, \mu_3^{(2)} = 1.5\right)^T$$
$$\Sigma_{obs}(1) = \begin{pmatrix} \sigma_{11}^{(2)} = 0.75 & \sigma_{13}^{(2)} = 0.75 \\ \sigma_{31}^{(2)} = 0.75 & \sigma_{33}^{(2)} = 2.75 \end{pmatrix}$$
$$V_{obs}^{mis}(1) = \begin{pmatrix} \sigma_{21}^{(2)} = -0.5 & \sigma_{23}^{(2)} = -1.5 \\ \sigma_{41}^{(2)} = -1 & \sigma_{43}^{(2)} = -3 \end{pmatrix}$$
$$V_{mis}^{obs}(1) = \begin{pmatrix} \sigma_{12}^{(2)} = -0.5 & \sigma_{14}^{(2)} = -1 \\ \sigma_{32}^{(2)} = -1.5 & \sigma_{34}^{(2)} = -3 \end{pmatrix}$$

$$\mu_{M_1}^{(2)} = \mu_{mis}(1) + \left(V_{obs}^{mis}(1)\right)\left(\Sigma_{obs}(1)\right)^{-1}\left(X_{obs}(1) - \mu_{obs}(1)\right)$$
$$= \left(\mu_{M_1}^{(2)}(2) \cong 0.17, \mu_{M_1}^{(2)}(4) \cong 0.33\right)^T$$

$$\Sigma_{M_1}^{(2)} = \Sigma_{mis}(1) - \left(V_{obs}^{mis}(1)\right)\left(\Sigma_{obs}(1)\right)^{-1}\left(V_{mis}^{obs}\right) = \begin{pmatrix} \Sigma_{M_1}^{(2)}(2,2) \cong 0{,}67 & \Sigma_{M_1}^{(2)}(2,4) \cong 0.33 \\ \Sigma_{M_1}^{(2)}(4,2) \cong 0.33 & \Sigma_{M_1}^{(2)}(4,4) \cong 1.17 \end{pmatrix}$$

$$X_{obs}(2) = (x_2 = 2, x_4 = 4)^T$$
$$\mu_{mis}(2) = \left(\mu_1^{(2)} = 0.5, \mu_3^{(2)} = 1.5\right)^T$$
$$\Sigma_{mis}(2) = \begin{pmatrix} \sigma_{11}^{(2)} = 0.75 & \sigma_{13}^{(2)} = 0.75 \\ \sigma_{31}^{(2)} = 0.75 & \sigma_{33}^{(2)} = 2.75 \end{pmatrix}$$
$$\mu_{obs}(2) = \left(\mu_2^{(2)} = 1, \mu_4^{(2)} = 2\right)^T$$
$$\Sigma_{obs}(2) = \begin{pmatrix} \sigma_{22}^{(2)} = 1.5 & \sigma_{24}^{(2)} = 2 \\ \sigma_{42}^{(2)} = 2 & \sigma_{44}^{(2)} = 4.5 \end{pmatrix}$$

$$V_{obs}^{mis}(2) = \begin{pmatrix} \sigma_{12}^{(2)} = -0.5 & \sigma_{14}^{(2)} = -1 \\ \sigma_{32}^{(2)} = -1.5 & \sigma_{34}^{(2)} = -3 \end{pmatrix}$$

$$V_{mis}^{obs}(2) = \begin{pmatrix} \sigma_{21}^{(2)} = -0.5 & \sigma_{23}^{(2)} = -1.5 \\ \sigma_{41}^{(2)} = -1 & \sigma_{43}^{(2)} = -3 \end{pmatrix}$$

$$\mu_{M_2}^{(2)} = \mu_{mis}(2) + \left(V_{obs}^{mis}(2)\right)\left(\Sigma_{obs}(2)\right)^{-1}\left(X_{obs}(2) - \mu_{obs}(2)\right)$$

$$= \left(\mu_{M_2}^{(2)}(1) \cong 0.05, \mu_{M_2}^{(2)}(3) = 0.14\right)^T$$

$$\Sigma_{M_2}^{(2)} = \Sigma_{mis}(2) - \left(V_{obs}^{mis}(2)\right)\left(\Sigma_{obs}(2)\right)^{-1}\left(V_{mis}^{obs}\right) = \begin{pmatrix} \Sigma_{M_2}^{(2)}(1,1) \cong 0.52 & \Sigma_{M_2}^{(2)}(1,3) \cong 0.07 \\ \Sigma_{M_2}^{(2)}(3,1) \cong 0.07 & \Sigma_{M_2}^{(2)}(3,3) \cong 0.7 \end{pmatrix}$$

$$\bar{x}_1^{(2)} = \frac{1}{2}\left(x_{11} + \mu_{M_2}^{(2)}(1)\right) \cong 0.52$$

$$\bar{x}_2^{(2)} = \frac{1}{2}\left(\mu_{M_1}^{(2)}(2) + x_{22}\right) \cong 1.1$$

$$\bar{x}_3^{(2)} = \frac{1}{2}\left(x_{13} + \mu_{M_2}^{(2)}(3)\right) \cong 1.57$$

$$\bar{x}_4^{(2)} = \frac{1}{2}\left(\mu_{M_1}^{(2)}(4) + x_{24}\right) \cong 2.17$$

$$s_{11}^{(2)} = \frac{1}{2}\left((x_{11})^2 + \left(\Sigma_{M_2}^{(2)}(1,1) + \left(\mu_{M_2}^{(2)}(1)\right)^2\right)\right) \cong 0.76$$

$$s_{12}^{(2)} = s_{21}^{(2)} = \frac{1}{2}\left(x_{11}\mu_{M_1}^{(2)}(2) + \mu_{M_2}^{(2)}(1)x_{22}\right) \cong 0.13$$

$$s_{13}^{(2)} = s_{31}^{(2)} = \frac{1}{2}\left(x_{11}x_{13} + \left(\Sigma_{M_2}^{(2)}(1,3) + \mu_{M_2}^{(2)}(1)\mu_{M_2}^{(2)}(3)\right)\right) \cong 1.54$$

$$s_{14}^{(2)} = s_{41}^{(2)} = \frac{1}{2}\left(x_{11}\mu_{M_1}^{(2)}(4) + \mu_{M_2}^{(2)}(1)x_{24}\right) \cong 0.17$$

$$s_{22}^{(2)} = \frac{1}{2}\left(\left(\Sigma_{M_1}^{(2)}(2,2) + \left(\mu_{M_1}^{(2)}(2)\right)^2\right) + (x_{22})^2\right) \cong 2.35$$

$$s_{23}^{(2)} = s_{32}^{(2)} = \frac{1}{2}\left(\mu_{M_1}^{(2)}(2)x_{13} + x_{22}\mu_{M_2}^{(2)}(3)\right) \cong 0.39$$

$$s_{24}^{(2)} = s_{42}^{(2)} = \frac{1}{2}\left(\left(\Sigma_{M_1}^{(2)}(2,4) + \mu_{M_1}^{(2)}(2)\mu_{M_1}^{(2)}(4)\right) + x_{22}x_{24}\right) \cong 4.19$$

$$s_{33}^{(2)} = \frac{1}{2}\left((x_{13})^2 + \left(\Sigma_{M_2}^{(2)}(3,3) + \left(\mu_{M_2}^{(2)}(3)\right)^2\right)\right) \cong 4.86$$

$$s_{34}^{(2)} = s_{43}^{(2)} = \frac{1}{2}\left(x_{13}\mu_{M_1}^{(1)}(4) + \mu_{M_2}^{(2)}(3)x_{24}\right) \cong 0.77$$

$$s_{44}^{(2)} = \frac{1}{2}\left(\left(\Sigma_{M_1}^{(2)}(4,4) + \left(\mu_{M_1}^{(2)}(4)\right)^2\right) + (x_{24})^2\right) \cong 8.64$$

At 2nd iteration, M-step, we have:

$$\mu_1^{(3)} = \bar{x}_1^{(2)} \cong 0.52$$

$$\mu_2^{(3)} = \bar{x}_2^{(2)} \cong 1.1$$

$$\mu_3^{(3)} = \bar{x}_3^{(2)} \cong 1.57$$

$$\mu_4^{(3)} = \bar{x}_4^{(2)} \cong 2.17$$

$$\sigma_{11}^{(3)} = s_{11}^{(2)} - \left(\bar{x}_1^{(2)}\right)^2 \cong 0.49$$
$$\sigma_{12}^{(3)} = \sigma_{21}^{(3)} = s_{12}^{(2)} - \bar{x}_1^{(2)}\bar{x}_2^{(2)} \cong -0.44$$
$$\sigma_{13}^{(3)} = \sigma_{31}^{(3)} = s_{13}^{(2)} - \bar{x}_1^{(2)}\bar{x}_3^{(2)} \cong 0.72$$
$$\sigma_{14}^{(3)} = \sigma_{41}^{(3)} = s_{14}^{(2)} - \bar{x}_1^{(2)}\bar{x}_4^{(2)} \cong -0.96$$
$$\sigma_{22}^{(3)} = s_{22}^{(2)} - \left(\bar{x}_2^{(2)}\right)^2 \cong 1.17$$
$$\sigma_{23}^{(3)} = \sigma_{32}^{(3)} = s_{23}^{(2)} - \bar{x}_2^{(2)}\bar{x}_3^{(2)} \cong -1.31$$
$$\sigma_{24}^{(3)} = \sigma_{42}^{(3)} = s_{24}^{(2)} - \bar{x}_2^{(2)}\bar{x}_4^{(2)} \cong 1.85$$
$$\sigma_{33}^{(3)} = s_{33}^{(2)} - \left(\bar{x}_3^{(2)}\right)^2 \cong 2.4$$
$$\sigma_{34}^{(3)} = \sigma_{43}^{(3)} = s_{34}^{(2)} - \bar{x}_3^{(2)}\bar{x}_4^{(2)} \cong -2.63$$
$$\sigma_{44}^{(3)} = s_{44}^{(2)} - \left(\bar{x}_4^{(2)}\right)^2 \cong 3.94$$

$$p^{(3)} = \frac{c(Z_1) + c(Z_2)}{4 * 2} = \frac{2+2}{4*2} = 0.5$$

Because the sample is too small for GEM to converge to an exact maximizer with small enough number of iterations, we can stop GEM at the second iteration with $\Theta^{(3)} = \Theta^* = (\mu^*, \Sigma^*)^T$ and $\Phi^{(3)} = \Phi^* = p^*$ when difference between $\Theta^{(2)}$ and $\Theta^{(3)}$ is insignificant.

$$\mu^* = (\mu_1^* = 0.52, \mu_2^* = 1.1, \mu_3^* = 1.57, \mu_4^* = 2.17)^T$$

$$\Sigma^* = \begin{pmatrix} \sigma_{11}^* = 0.49 & \sigma_{12}^* = -0.44 & \sigma_{13}^* = 0.72 & \sigma_{14}^* = -0.96 \\ \sigma_{21}^* = -0.44 & \sigma_{22}^* = 1.17 & \sigma_{23}^* = -1.31 & \sigma_{24}^* = 1.85 \\ \sigma_{31}^* = 0.72 & \sigma_{32}^* = -1.31 & \sigma_{33}^* = 2.4 & \sigma_{34}^* = -2.63 \\ \sigma_{41}^* = -0.96 & \sigma_{42}^* = 1.85 & \sigma_{43}^* = -2.63 & \sigma_{44}^* = 3.94 \end{pmatrix}$$

$$p^* = 0.5$$

As aforementioned, because X_{mis} is a part of X and $f(X_{mis} \mid \Theta_M)$ is derived directly from $f(X|\Theta)$, in practice we can stop GEM after its first iteration was done, which is reasonable enough to handle missing data.

As aforementioned, an interesting application of handling missing data is to fill in or predict missing values. For instance, the missing part $X_{mis}(1)$ of $X_1 = (x_{11}=1, x_{12}=?, x_{13}=3, x_{14}=?)^T$ is fulfilled by $\mu_{M_1}^*$ according to equation 2.44 as follows:

$$x_{12} = \mu_2^* = 1.1$$
$$x_{14} = \mu_4^* = 2.17$$

It is necessary to have an example for illustrating how to handle missing data with multinomial PDF.

Example 3.2. Given sample of size two, $\mathcal{X} = \{X_1, X_2\}$ in which $X_1 = (x_{11}=1, x_{12}=?, x_{13}=3, x_{14}=?)^T$ and $X_2 = (x_{21}=?, x_{22}=2, x_{23}=?, x_{24}=4)^T$ are iid.

	x_1	x_2	x_3	x_4
X_1	1	?	3	?
X_2	?	2	?	4

Of course, we have $X_{obs}(1) = (x_{11}=1, x_{13}=3)^T$, $X_{mis}(1) = (x_{12}=?, x_{14}=?)^T$, $X_{obs}(2) = (x_{22}=2, x_{24}=4)^T$ and $X_{mis}(2) = (x_{21}=?, x_{23}=?)^T$. We also have $M_1 = \{m_{11}=2, m_{12}=4\}$, $\bar{M}_1 = \{\bar{m}_{11}=1, \bar{m}_{12}=3\}$, $M_2 = \{m_{21}=1, m_{22}=3\}$, and $\bar{M}_2 = \{\bar{m}_{21}=2, \bar{m}_{22}=4\}$. Let X be random variable representing every X_i. Suppose $f(X|\Theta)$ is multinomial PDF of 10 trials. We will estimate $\Theta = (p_1, p_2, p_3, p_4)^T$. The parameters p_1, p_2, p_3, and p_2 are initialized arbitrarily as 0.25 as follows:

$$\Theta^{(1)} = \left(p_1^{(1)} = 0.25, p_2^{(1)} = 0.25, p_3^{(1)} = 0.25, p_4^{(1)} = 0.25\right)^T$$

At 1st iteration, M-step, we have:

$$p_1^{(2)} = \frac{1}{10 * 2}(1 + 10 * 0.25) = 0.175$$
$$p_2^{(2)} = \frac{1}{10 * 2}(10 * 0.25 + 2) = 0.225$$
$$p_3^{(2)} = \frac{1}{10 * 2}(3 + 10 * 0.25) = 0.275$$
$$p_4^{(2)} = \frac{1}{10 * 2}(10 * 0.25 + 4) = 0.325$$

We stop GEM after the first iteration was done, which results the estimate $\Theta^{(2)} = \Theta^* = (p_1^*, p_2^*, p_3^*, p_4^*)^T$ as follows:

$$p_1^* = 0.175$$
$$p_2^* = 0.225$$
$$p_3^* = 0.275$$
$$p_4^* = 0.325$$

4. Conclusions

In general, GEM is a powerful tool to handle missing data, which is not so difficult except that how to extract the parameter Θ_M of the conditional PDF $f(X_{mis} \mid X_{obs}, \Theta_M)$ from the whole parameter Θ of the PDF $f(X|\Theta_M)$ is most important with note that only $f(X|\Theta)$ is defined firstly and then $f(X_{mis} \mid X_{obs}, \Theta_M)$ is derived from $f(X|\Theta)$. Therefore, equation 2.15 is cornerstone of this method. Note, equation 2.35 and 2.51 are instances of equation 2.15 when $f(X|\Theta)$ is multinormal PDF or multinomial PDF.

References

Burden, R. L., & Faires, D. J. (2011). *Numerical Analysis* (9th Edition ed.). (M. Julet, Ed.) Brooks/Cole Cengage Learning.

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

Hardle, W., & Simar, L. (2013). *Applied Multivariate Statistical Analysis.* Berlin, Germany: Research Data Center, School of Business and Economics, Humboldt University.

Josse, J., Jiang, W., Sportisse, A., & Robin, G. (2018). *Handling missing values.* Inria. Julie Josse. Retrieved October 12, 2020, from http://juliejosse.com/wp-content/uploads/2018/07/LectureNotesMissing.html

Nguyen, L. (2020). *Tutorial on EM algorithm.* MDPI. Preprints. doi:10.20944/preprints201802.0131.v8

Ta, P. D. (2014). *Numerical Analysis Lecture Notes.* Vietnam Institute of Mathematics, Numerical Analysis and Scientific Computing. Hanoi: Vietnam Institute of Mathematics. Retrieved 2014

Wikipedia. (2014, August 4). *Karush–Kuhn–Tucker conditions*. (Wikimedia Foundation) Retrieved November 16, 2014, from Wikipedia website: http://en.wikipedia.org/wiki/Karush–Kuhn–Tucker_conditions

Wikipedia. (2016, March September). *Exponential family*. (Wikimedia Foundation) Retrieved 2015, from Wikipedia website: https://en.wikipedia.org/wiki/Exponential_family