Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2020 doi:10.20944/preprints202010.0550.v2

Conditional mixture model and its application for regression model

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 statistical parameter when data sample contains hidden part and observed part. EM is applied to learn finite mixture model in which the whole distribution of observed variable is average sum of partial distributions. Coverage ratio of every partial distribution is specified by the probability of hidden variable. An application of mixture model is soft clustering in which cluster is modeled by hidden variable whereas each data point can be assigned to more than one cluster and degree of such assignment is represented by the probability of hidden variable. However, such probability in traditional mixture model is simplified as a parameter, which can cause loss of valuable information. Therefore, in this research I propose a so-called conditional mixture model (CMM) in which the probability of hidden variable is modeled as a full probabilistic density function (PDF) that owns individual parameter. CMM aims to extend mixture model. I also propose an application of CMM which is called adaptive regression model (ARM). Traditional regression model is effective when data sample is scattered equally. If data points are grouped into clusters, regression model tries to learn a unified regression function which goes through all data points. Obviously, such unified function is not effective to evaluate response variable based on grouped data points. The concept "adaptive" of ARM means that ARM solves the ineffectiveness problem by selecting the best cluster of data points firstly and then evaluating response variable within such best cluster. In order words, ARM reduces estimation space of regression model so as to gain high accuracy in calculation.

Keywords: expectation maximization (EM) algorithm, finite mixture model, conditional mixture model, regression model, adaptive regression model (ARM).

1. Introduction

Suppose data has two parts such as hidden part X and observed part Y and we only know Y. A relationship between random variable X and random variable Y is specified by the joint probabilistic density function (PDF) denoted $f(X, Y \mid \Theta)$ where Θ is parameter. Given sample $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$ whose all Y_i (s) are mutually independent and identically distributed (iid), it is required to estimate Θ based on $\mathcal{Y}$ whereas X is unknown. Expectation maximization (EM) algorithm is applied to solve this problem when only $\mathcal{Y}$ is observed. EM has many iterations and each iteration has two steps such as expectation step (E-step) and maximization step (M-step). At some t^{th} iteration, given current parameter $\Theta^{(t)}$, the two steps are described as follows:

E-step:

The expectation $Q(\Theta \mid \Theta^{(t)})$ is determined based on current parameter $\Theta^{(t)}$, according to equation 1.1 (Nguyen, Tutorial on EM tutorial, 2020, p. 50).

$$Q\left(\Theta \middle| \Theta^{(t)}\right) = \sum_{i=1}^{N} \int_{X} f(X|Y_i, \Theta) \log\left(f(X, Y_i|\Theta')\right) \mathrm{d}X \tag{1.1}$$

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 $(t+1)^{\text{th}}$ iteration).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2020 doi:10.20944/preprints202010.0550.v2

EM algorithm will converge after some iterations, at that time we have the estimate $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^{*}$. Note, the estimate Θ^{*} is result of EM.

Especially, the random variable X represents latent class or latent component of random variable Y. Suppose X is discrete and ranges in $\{1, 2, \ldots, K\}$. As a convention, let k=X. Note, because all Y_i (s) are iid, let random variable Y represent every Y_i. The so-called probabilistic finite mixture model is represented by the PDF of Y, as follows:

$$f(Y|\Theta) = \sum_{k=1}^{K} \alpha_k f_k(Y|\theta_k) \tag{1.2}$$

Where,

$$\Theta = (\alpha_1, \alpha_2, \ldots, \alpha_K, \theta_1, \theta_2, \ldots, \theta_K)^T$$

$$\sum_{k=1}^{K} \alpha_k = 1$$

Note, the superscript "T" denotes transpose operator for vector and matrix. The $Q(\Theta \mid \Theta^{(t)})$ is re-defined for finite mixture model as follows (Nguyen, Tutorial on EM tutorial, 2020, p. 79):

$$Q\left(\Theta\middle|\Theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} P\left(k\middle|Y_i, \Theta^{(t)}\right)\log\left(\alpha_k f_k(Y_i|\theta_k)\right) \tag{1.3}$$

Where,

$$P\left(k\middle|Y_i, \Theta^{(t)}\right) = \frac{\alpha_k^{(t)} f_k\left(Y_i\middle|\theta_k^{(t)}\right)}{\sum_{l=1}^{K} \alpha_l^{(t)} f_l\left(Y_i\middle|\theta_l^{(t)}\right)} \tag{1.4}$$

If every $f_k(Y|\theta_k)$ distributes normally with mean μ_k and covariance matrix Σ_k such that $\theta_k = (\mu_k, \Sigma_k)^T$, the next parameter $\Theta^{(t+1)}$ is calculated at M-step of such t^{th} iteration given current parameter $\Theta^{(t)}$ as follows (Nguyen, Tutorial on EM tutorial, 2020, p. 85):

$$\begin{aligned}
\alpha_k^{(t+1)} &= \frac{1}{N}\sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right) \\
\mu_k^{(t+1)} &= \frac{\sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right) Y_i}{\sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right)} \\
\Sigma_k^{(t+1)} &= \frac{\sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right)\left(\left(Y_i - \mu_k^{(t+1)}\right)\left(Y_i - \mu_k^{(t+1)}\right)^T\right)}{\sum_{i=1}^{N} P\left(k\middle|Y_i, \Theta^{(t)}\right)}
\end{aligned} \tag{1.5}$$

Note, the conditional probability $P(k \mid Y_i, \Theta^{(t)})$ is calculated at E-step.

In the traditional finite mixture model, the parameter α_k is essentially the parameter of hidden random variable X when X is discrete, $\alpha_k = P(X=k)$. In other words, $P(X)$ is "reduced" at most. There is a problem of how to define and learn finite mixture model when $P(X)$ is still a full PDF which owns individual parameter. Such problem is solved by the definition of conditional mixture model (CMM) in the next section.

2. Conditional mixture model

Now let W and Y be two random variables and both of them are observed. I define the conditional PDF of Y given W as follows:

$$f(Y|W, \Theta) = \sum_{k=1}^{K} \frac{g_k(W|\varphi_k)}{\sum_{l=1}^{K} g_l(W|\varphi_l)} f_k(Y|W, \theta_k) \tag{2.1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2020 doi:10.20944/preprints202010.0550.v2

Where $g_k(W|\varphi_k)$ is the k^{th} PDF of W which can be considered PDF of X for the k^{th} component. Equation 2.1 specifies the so-call conditional mixture model (CMM) when random variable Y is dependent on another random variable W. It is possible to consider that the parameter α_k in the traditional mixture model specified by equation 1.2 is:

$$\alpha_k = \frac{g_k(W|\varphi_k)}{\sum_{l=1}^{K} g_l(W|\varphi_l)}$$

It is deduced that hidden variable $X=k$ in CMM is represented by $g_k(W|\varphi_k)$ with a full of necessary parameters φ_k. When the sum $\sum_{l=1}^{K} g_l(W|\varphi_l)$ is considered as constant, we have:

$$f(Y|W,\Theta) = \frac{1}{\sum_{l=1}^{K} g_l(W|\varphi_l)} \sum_{k=1}^{K} g_k(W|\varphi_k) f_k(Y|W,\theta_k) \propto \sum_{k=1}^{K} g_k(W|\varphi_k) f_k(Y|W,\theta_k)$$

Where the sign "$\propto$" indicates proportion. The quasi-conditional PDF of Y given W is defined to be proportional to the conditional PDF of Y given W as follows:

$$\tilde{f}(Y|W,\Theta) = \sum_{k=1}^{K} g_k(W|\varphi_k) f_k(Y|W,\theta_k) \tag{2.2}$$

Where the parameter of CMM is $\Theta = (\varphi_1, \varphi_2, \ldots, \varphi_K, \theta_1, \theta_2, \ldots, \theta_K)^T$. Of course, we have:

$$f(Y|W,\Theta) \propto \tilde{f}(Y|W,\Theta)$$

Given sample $\mathcal{Z} = \{Z_1 = \{W_1, Y_1\}, Z_2 = \{W_2, Y_2\}, \ldots, \}, Z_N = \{W_N, Y_N\})\}$ of size N in which all X_i (s) are iid and all y_i (s) are iid, we need to learn CMM. Let W and Y represent every W_i and every Y_i, respectively. When applying EM along with the quasi-conditional PDF $\tilde{f}(Y|W,\Theta)$ to estimate Θ, the $Q(\Theta \mid \Theta^{(t)})$ is re-defined as follows:

$$Q\left(\Theta\middle|\Theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} P\left(k\middle|W_i,Y_i,\Theta^{(t)}\right)\log\left(g_k(W_i|\varphi_k) f_k(Y_i|W_i,\theta_k)\right) \tag{2.3}$$

Where $P(k \mid W_i, Y_i)$ is determined according to Bayes' rule,

$$P\left(k\middle|W_i,Y_i,\Theta^{(t)}\right) = \frac{g_k\left(W_i\middle|\varphi_k^{(t)}\right) f_k\left(Y_i\middle|W_i,\theta_k^{(t)}\right)}{\sum_{l=1}^{K} g_l\left(W_i\middle|\varphi_l^{(t)}\right) f_l\left(Y_i\middle|W_i,\theta_l^{(t)}\right)} \tag{2.4}$$

We need to maximize $Q(\Theta \mid \Theta^{(t)})$ at M-step of some t^{th} iteration given current parameter $\Theta^{(t)}$. Expectedly, the next parameter $\Theta^{(t+1)}$ is solution of the equation created by setting the first-order derivative of $Q(\Theta \mid \Theta^{(t)})$ with regard to Θ to be zero. The first-order partial derivatives of $Q(\Theta \mid \Theta^{(t)})$ with regard to φ_k and θ_k are:

$$\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial \varphi_k} = \sum_{i=1}^{N} P\left(k\middle|W_i,Y_i,\Theta^{(t)}\right)\frac{\partial \log\left(g_k(W_i|\varphi_k)\right)}{\partial \varphi_k}$$

$$\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial \theta_k} = \sum_{i=1}^{N} P\left(k\middle|W_i,Y_i,\Theta^{(t)}\right)\frac{\partial \log\left(f_k(Y_i|W_i,\theta_k)\right)}{\partial \theta_k}$$

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

$$\begin{cases} \sum_{i=1}^{N} P\left(k\middle|W_i,Y_i,\Theta^{(t)}\right)\frac{\partial \log\left(g_k(W_i|\varphi_k)\right)}{\partial \varphi_k} = \mathbf{0}^T \\ \sum_{i=1}^{N} P\left(k\middle|W_i,Y_i,\Theta^{(t)}\right)\frac{\partial \log\left(f_k(Y_i|W_i,\theta_k)\right)}{\partial \theta_k} = \mathbf{0}^T \end{cases} \tag{2.5}$$

How to solve the equation 2.5 depends on individual applications. The next section describes an application of CMM.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2020 doi:10.20944/preprints202010.0550.v2

3. Adaptive regression model

Traditional regression model is effective when data sample is scattered equally. If data points are grouped into clusters with their nature, regression model tries to learn a unified regression function which goes through all data points. Obviously, such unified function is not effective to evaluate response variable based on grouped data points. Alternately, if it is possible to select a right cluster for evaluating response variable, the value of response variable will be more precise. Therefore, selective evaluation is the main idea of adaptive regression model (ARM). The main ideology of ARM to group sample into clusters and build respective regression functions for clusters in parallel. CMM is applied to solve this problem, in other words, ARM is an application of CMM. There may be other applications of CMM but here I focus on ARM.

Given a n-dimension random variable $W = (w_1, w_2,\ldots, w_n)^T$ which is called regressors, a linear regression function is defined as

$$y = \beta_0 + \sum_{j=1}^{n} \beta_j w_j \tag{3.1}$$

Where y is the random variable called response variable and each β_j is called regressive coefficient. According to linear regression model, y conforms multinormal distribution, as follows:

$$f(y|W,\theta) = f(y|W,\alpha,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\beta^T W)^2}{2\sigma^2}\right) \tag{3.2}$$

Where $\beta = (\beta_0, \beta_1,\ldots, \beta_n)^T$ is called regressive parameter of $f(y \mid W, \beta, \sigma^2)$. Therefore, mean and variance of $f(y \mid W, \beta, \sigma^2)$ are $\beta^T X$ and σ^2, respectively. Note, $f(y \mid W, \beta, \sigma^2)$ is called regressive PDF of y. As a convention, we denote:

$$\beta^T W = \beta_0 + \sum_{j=1}^{n} \beta_j w_j$$

Given sample $\mathcal{Z} = \{Z_1 = \{W_1, y_1\}, Z_2 = \{W_2, y_2\},\ldots, \}, Z_N = \{W_N, y_N\})\}$ of size N in which all X_i (s) are iid and all y_i (s) are iid. Let $W = (w_1, w_2,\ldots, w_n)^T$ and y represent every $W_i = (w_{i1}, w_{i2},\ldots, w_{in})$ and every y_i, respectively. Let $\boldsymbol{W}$ and $\boldsymbol{y}$ be a matrix and a vector extracted from $\mathcal{Z}$ as follows:

$$\boldsymbol{W} = \begin{pmatrix} 1 & w_{11} & w_{12} & \cdots & w_{1n} \\ 1 & w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & w_{N1} & w_{N2} & \cdots & w_{Nn} \end{pmatrix}$$
$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \tag{3.3}$$

When applying EM to estimate Θ, by following equation 2.3, the $Q(\Theta \mid \Theta^{(t)})$ for ARM is re-defined as follows:

$$Q\left(\Theta\middle|\Theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} P\left(k\middle|W_i, Y_i, \Theta^{(t)}\right)\log\left(g_k(W_i|\varphi_k) f_k(y_i|W_i,\theta_k)\right) \tag{3.4}$$

Where,

$$P\left(k\middle|W_i, y_i, \Theta^{(t)}\right) = \frac{g_k\left(W_i\middle|\varphi_k^{(t)}\right) f_k\left(y_i\middle|W_i,\theta_k^{(t)}\right)}{\sum_{l=1}^{K} g_l\left(W_i\middle|\varphi_l^{(t)}\right) f_l\left(y_i\middle|W_i,\theta_l^{(t)}\right)} \tag{3.5}$$

Where the parameter of ARM is $\Theta = (\varphi_1, \varphi_2,\ldots, \varphi_K, \theta_1, \theta_2,\ldots, \theta_K)^T$ but each φ_k and each θ_k are resolved more complexly. The definition of $Q(\Theta \mid \Theta^{(t)})$ implies that sample $\mathcal{Z}$ can be grouped into K clusters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2020 doi:10.20944/preprints202010.0550.v2

The function $f_k(y \mid W, \theta_k)$ is the k^{th} regressive PDF of y.

$$f_k(y|W,\theta_k) = f_k(y|W,\beta_k,\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left(-\frac{(y-\beta_k^T W)^2}{2\sigma_k^2}\right) \quad (3.6)$$

Obviously, we have:

$$\begin{aligned}\theta_k &= (\beta_k, \sigma_k^2)^T \\ \beta_k &= (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kn})^T \\ \beta^T W &= \beta_0 + \sum_{j=1}^{n} \beta_j w_j\end{aligned} \quad (3.7)$$

For convenience, suppose the k^{th} PDF of W denoted $g_k(W|\varphi_k)$ is multinormal PDF as follows:

$$g_k(W|\varphi_k) = g_k(W|\mu_k,\Sigma_k) = (2\pi)^{-\frac{n}{2}}|\Sigma_k|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(W-\mu_k)^T\Sigma_k^{-1}(W-\mu_k)\right) \quad (3.8)$$

Where,

$$\varphi_k = (\mu_k, \Sigma_k)^T$$

We need to maximize $Q(\Theta \mid \Theta^{(t)})$ at M-step of some t^{th} iteration given current parameter $\Theta^{(t)}$. Expectedly, the next parameter $\Theta^{(t+1)}$ is solution of the equation created by setting the first-order derivative of $Q(\Theta \mid \Theta^{(t)})$ with regard to Θ to be zero.

The first-order partial derivative of $Q(\Theta \mid \Theta^{(t)})$ with regard to β_k is:

$$\begin{aligned}\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial\beta_k} &= \sum_{i=1}^{N} P\left(k\middle|X_i, y_i, \Theta^{(t)}\right)\frac{\partial\log\left(f_k(y_i|W_i,\theta_k)\right)}{\partial\beta_k} \\ &= \sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)\left(y_i - \left(\beta_k^{(t)}\right)^T W_i\right)W_i^T\end{aligned}$$

By referring (Nguyen & Shafiq, Mixture Regression Model for Incomplete Data, 2018, pp. 11-13), the next parameter $\beta_k^{(t+1)}$ is solution of the equation $\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial\beta_k} = \mathbf{0}^T$ where $\mathbf{0}$ is zero vector, as follows:

$$\beta_k^{(t+1)} = \left(\boldsymbol{W}^T\boldsymbol{U}_k^{(t)}\right)^{-1}\boldsymbol{W}^T V_k^{(t)} \quad (3.9)$$

Where,

$$\boldsymbol{U}_k^{(t)} = \begin{pmatrix} u_{10}^{(t)}(k) & u_{11}^{(t)}(k) & \cdots & u_{1n}^{(t)}(k) \\ u_{20}^{(t)}(k) & u_{21}^{(t)}(k) & \cdots & u_{2n}^{(t)}(k) \\ \vdots & \vdots & \ddots & \vdots \\ u_{N0}^{(t)}(k) & u_{N1}^{(t)}(k) & \cdots & u_{Nn}^{(t)}(k) \end{pmatrix} \quad (3.10)$$

$$u_{ij}^{(t)}(k) = w_{ij}P\left(k\middle|W_i, y_i, \beta_k^{(t)}, (\sigma_k^2)^{(t)}\right)$$

And,

$$V_k^{(t)} = \begin{pmatrix} v_0^{(t)}(k) \\ v_1^{(t)}(k) \\ \vdots \\ v_n^{(t)}(k) \end{pmatrix} \quad (3.11)$$

$$v_i^{(t)}(k) = y_i P\left(k\middle|W_i, y_i, \beta_k^{(t)}, (\sigma_k^2)^{(t)}\right)$$

The first-order partial derivative of $Q(\Theta \mid \Theta^{(t)})$ with regard to σ_k^2 is:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2020 doi:10.20944/preprints202010.0550.v2

$$\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial \sigma_k^2} = \sum_{i=1}^{N} P\left(k\middle|X_i, y_i, \Theta^{(t)}\right) \frac{\partial \log\left(f_k(y_i|W_i, \theta_k)\right)}{\partial \sigma_k^2}$$
$$= \sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)(y_i - \beta_k^T W_i)^2 - \left(\sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)\right)\sigma_k^2$$

The next parameter $(\sigma_k^2)^{(t+1)}$ which is solution of the equation $\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial \sigma_k^2} = 0$ is:

$$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)\left(y_i - \left(\beta_k^{(t+1)}\right)^T W_i\right)^2}{\sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)} \quad (3.12)$$

Where $\beta_k^{(t+1)}$ is specified in equation 3.9.

The first-order partial derivative of $Q(\Theta \mid \Theta^{(t)})$ with regard to μ_k is:

$$\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial \mu_k} = \sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right) \frac{\partial \log\left(g_k(W_i|\theta_k)\right)}{\partial \mu_k}$$
$$= \left(\sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)(W_i - \mu_k)^T\right)\Sigma_k^{-1}$$

The next parameter $\mu_k^{(t+1)}$ which is solution of the equation $\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial \mu_k} = \mathbf{0}^T$ is:

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} P\left(k\middle|W_i, Y_i, \Theta^{(t)}\right)W_i}{\sum_{i=1}^{N} P\left(k\middle|W_i, Y_i, \Theta^{(t)}\right)} \quad (3.13)$$

The first-order partial derivative of $Q(\Theta \mid \Theta^{(t)})$ with regard to Σ_k is (Nguyen, Tutorial on EM tutorial, 2020, pp. 83-84):

$$\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial \Sigma_k} = \sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right) \frac{\partial \log\left(g_k(W_i|\theta_k)\right)}{\partial \Sigma_k}$$
$$= \sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)\left((W_i - \mu_k)(W_i - \mu_k)^T\right) - \left(\sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)\right)\Sigma_k$$

The next parameter $\Sigma_k^{(t+1)}$ which is solution of the equation $\frac{\partial Q\left(\Theta\middle|\Theta^{(t)}\right)}{\partial \Sigma_k} = (\mathbf{0})$ where $(\mathbf{0})$ is zero matrix is:

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)\left(\left(W_i - \mu_k^{(t+1)}\right)\left(W_i - \mu_k^{(t+1)}\right)^T\right)}{\sum_{i=1}^{N} P\left(k\middle|W_i, y_i, \Theta^{(t)}\right)} \quad (3.14)$$

Where $\mu_k^{(t+1)}$ is specified in equation 3.13.

In general, at some t^{th} iteration, given current parameter $\Theta^{(t)}$, the two steps of EM for ARM are described as follows:

E-step:

The conditional probability $P(k \mid Y_i, \Theta^{(t)})$ is calculated based on current parameter $\Theta^{(t)} = (\varphi_1^{(t)}, \varphi_2^{(t)}, \ldots, \varphi_K^{(t)}, \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_K^{(t)})^T$, according to equation 3.5.

$$P\left(k\middle|W_i, y_i, \Theta^{(t)}\right) = \frac{g_k\left(W_i\middle|\varphi_k^{(t)}\right)f_k\left(y_i\middle|W_i, \theta_k^{(t)}\right)}{\sum_{l=1}^{K} g_l\left(W_i\middle|\varphi_l^{(t)}\right)f_l\left(y_i\middle|W_i, \theta_l^{(t)}\right)}$$

Where,

$$\theta_k = (\beta_k, \sigma_k^2)^T$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2020 doi:10.20944/preprints202010.0550.v2

$$\varphi_k = (\mu_k, \Sigma_k)^T$$

M-step:

The next parameter $\Theta^{(t+1)} = (\varphi_1^{(t+1)}, \varphi_2^{(t+1)}, \ldots, \varphi_K^{(t+1)}, \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_K^{(t+1)})^T$, which is a maximizer of $Q(\Theta \mid \Theta^{(t)})$ with subject to Θ, is calculated by equation 3.19, equation 3.12, , equation 3.13, and equation 3.14 with current parameter $\Theta^{(t)}$.

$$\beta_k^{(t+1)} = \left(\boldsymbol{W}^T \boldsymbol{U}_k^{(t)}\right)^{-1} \boldsymbol{W}^T V_k^{(t)}$$

$$(\sigma_k^2)^{(t+1)} = \frac{\sum_{i=1}^{N} P\left(k \middle| W_i, y_i, \Theta^{(t)}\right) \left(y_i - \left(\beta_k^{(t+1)}\right)^T W_i\right)^2}{\sum_{i=1}^{N} P\left(k \middle| W_i, y_i, \Theta^{(t)}\right)}$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} P\left(k \middle| W_i, Y_i, \Theta^{(t)}\right) W_i}{\sum_{i=1}^{N} P\left(k \middle| W_i, Y_i, \Theta^{(t)}\right)}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{N} P\left(k \middle| W_i, y_i, \Theta^{(t)}\right) \left(\left(W_i - \mu_k^{(t+1)}\right)\left(W_i - \mu_k^{(t+1)}\right)^T\right)}{\sum_{i=1}^{N} P\left(k \middle| W_i, y_i, \Theta^{(t)}\right)}$$

Where $\boldsymbol{U}_k^{(t)}$ and $V_k^{(t)}$ are specified by equation 3.10 and equation 3.11, respectively.

EM algorithm will converge after some iterations, at that time we have the estimate $\Theta^{(t)} = \Theta^{(t+1)} = \Theta^* = (\varphi_1^*, \varphi_2^*, \ldots, \varphi_K^*, \theta_1^*, \theta_2^*, \ldots, \theta_K^*)^T$. As a result, ARM is specified by the estimate Θ^*. It can be said that Θ^* is ARM. Given any data point W, ARM select the best cluster v whose PDF $g_v(W \mid \varphi_v^*)$ is maximal, as follows:

$$v = \underset{k}{\operatorname{argmax}}\, g_k(W|\varphi_k^*) \tag{3.15}$$

Then ARM evaluates the response variable Y given regressor W with regard to such best cluster v as follows:

$$Y = (\beta_v^*)^T W = \beta_{v0}^* + \sum_{j=1}^{n} \beta_{vj}^* w_j \tag{3.16}$$

Instead of selecting the best cluster for evaluation, ARM can make an average over K clusters for evaluating Y as follows:

$$\begin{aligned} Y &= \frac{1}{\sum_{l=1}^{K} g_l(W|\varphi_l)} \sum_{k=1}^{K} g_k(W|\varphi_k)((\beta_k^*)^T W) \\ &= \frac{1}{\sum_{l=1}^{K} g_l(W|\varphi_l)} \sum_{k=1}^{K} g_k(W|\varphi_k) \left(\beta_{k0}^* + \sum_{j=1}^{n} \beta_{kj}^* w_j\right) \end{aligned} \tag{3.17}$$

In general, equation 3.16 is the main one used to evaluate the regression function because the concept "adaptive" implies that ARM selects the best cluster (adaptive cluster) for evaluation.

4. Conclusions

The main ideology of CMM is to improve competence of mixture model, in which the probability of hidden variable is turned back its original form of PDF with full of parameters. As a result, its application ARM takes advantages of such hidden parameters in order to select best group or best cluster for making prediction of response value. In order words, ARM reduces estimation space of regression model so as to gain high accuracy in calculation. However, a new problem raised for CMM as well as ARM is how to pre-define the number K of clusters or components when CMM currently set fixed K. In the future, I will research some methods (Hoshikawa, 2013, p. 5) to pre-define K. Alternately, CMM can be improved or modified so that the number of clusters is updated in runtime (Nguyen & Shafiq, Mixture

 doi:10.20944/preprints202010.0550.v2

Regression Model for Incomplete Data, 2018, p. 16); in other words, there is no pre-definition of *K* and so *K* is determined dynamically.

References

Hoshikawa, T. (2013, June 30). Mixture regression for observational data, with application to functional regression models. *arXiv preprint*. Retrieved September 4, 2018, from https://arxiv.org/pdf/1307.0170

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

Nguyen, L., & Shafiq, A. (2018, December 31). Mixture Regression Model for Incomplete Data. (L. d. Istael, Ed.) *Revista Sociedade Científica, 1*(3), 1-25. doi:10.5281/zenodo.2528978