Erasmus University Rotterdam

Tatev Karen Aslanyan

Finite Mixture Models

with Negative Binomial distribution using EM-algorithm

In order to estimate the parameters of the finite mixture model, the method of maximum likelihood(ML) can be used. However, the log-likelihood function in ML has, in general, multiple local maxima and gives poor results. ML estimation based on numerical optimization algorithm does not work smoothly, and this is mainly because the number of components is unknown. Another, better approach for estimating the parameters in finite mixture model is the Expectation-Maximization algorithm (EM) described in Cameron, (2005). The finite mixture regression model is given by:

$$y_{it} = \alpha_{s_i} + x'_{it}\beta_{s_i} + \varepsilon_{it} \tag{0.1}$$

where t = 1,...,T_i and i = 1,...,N. The segment of i^{th} observation is $s_i \in \{1,...,S\}$. Suppose, $P(s_i = s) = \lambda_s$ such that $\sum_{s=1}^{S} \lambda_s = 1$ with $p_s \geq 0$ where λ_s represents the weight of component s. So, α and β can take S different values $\alpha_1,...,\alpha_S$ and $\beta_1,...,\beta_S$with corresponding weights $p_1,...,p_S$ respectively. The likelihood function is described as follows:

$$L = \prod_{i=1}^{N} \sum_{i=1}^{S} \lambda_s \prod_{t=1}^{T_i} p(y_{it} \mid x_{it}, \theta_s) \tag{0.2}$$

where $p(y_{it} \mid x_{it},\theta_s)$ represents the conditional distribution of y_{it}. Suppose, w_{is} indicates whether i belongs to segment s such that it takes value 1 if i^{th} observation belongs to group s:

$$w_{is} = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{if else} \end{cases} \tag{0.3}$$

The complete data likelihood function and complete log-likelihood functions are defined as follows:

$$\begin{aligned} L = p(y, s \mid x, \theta) &= \prod_{i=1}^{N} \prod_{s=1}^{S} \Big(\lambda_s P(y_i \mid X_i, \theta_s)\Big)^{z_{is}} = \prod_{i=1}^{N} \prod_{s=1}^{S} \Big(\lambda_s \prod_{t=1}^{T_i} P(y_{it} \mid x_{it}, \theta_s)\Big)^{z_{is}} \\ l = \ln\, p(y, s \mid x, \theta) &= \sum_{i=1}^{N} \sum_{s=1}^{S} w_{is} \Big(\ln(\lambda_s) + \sum_{t=1}^{T_i} \ln(P(y_{it} \mid x_{it}, \theta_s))\Big) \\ &= \sum_{i=1}^{N} \sum_{s=1}^{S} w_{is} \ln(\lambda_s) + \sum_{i=1}^{N} \sum_{s=1}^{S} w_{is} \sum_{t=1}^{T_i} \ln(P(y_{it} \mid x_{it}, \theta_s)) \end{aligned} \tag{0.4}$$

EM algorithm

The EM algorithm is iterative process consisting of two-steps:

E-step: Determine the expectation of the complete log-likelihood with respect to s|y with current estimate of θ ($\hat{\theta}$), that is: $E_{s|y}[\ln(P(y,s|\ x,\hat{\theta}))]$.

M-step: Maximize the expected value for parameter θ to update new estimator $\hat{\theta}^*$, that is:

arg max$_\theta$ $E_{s|y}[\ln(P(y,s|\ x,\theta))]$ where different initialization estimates can be used.

In order to adjust the EM algorithm to the negative binomial distribution described earlier we adjust the weights defined in Equation 0.3. Suppose, $w_i = [w_{i1},...,w_{iS}]'$ is S dimensional vector containing all weights assigned to observation and W is the matrix containing w_i vectors for all N observations.

E-step

In E-step the expectation of complete log-likelihood (l) function conditional on data which is given and the parameters that have to be estimated. We denote by $\hat{\Lambda} = (\hat{\lambda}_1,...,\hat{\lambda}_S)'$,$\hat{B} = (\hat{\beta}_1,...,\hat{\beta}_S)'$ and $\hat{D} = (\hat{d}_1,...,\hat{d}_S)'$ the estimates of corresponding parameters of the model where β_s represents the slope coefficient vector of s^{th} segment with dimension J and d_s is the dispersion parameter of segment s. Consequently, the expectation of l is defined as follows:

$$E_s[l \mid \hat{\Lambda}, \hat{B}, \hat{D}] = \sum_{i=1}^{N}\sum_{s=1}^{S} E[w_{is} \mid y_i]\ln(\hat{\lambda}_s) + \sum_{i=1}^{N}\sum_{s=1}^{S} E[w_{is} \mid y_i]\sum_{t=1}^{T_i}\ln(P(y_{it} \mid x_{it}, \hat{\beta}_s, \hat{d}_s)) \tag{0.5}$$

Conditional distribution of weight w_{is} based Bayes theorem is defined as follows:

$$P[w_{is} \mid y_i] = \frac{\prod_{s=1}^{S}\hat{\lambda}_s(P(y_i \mid X_i, \hat{\beta}_s, \hat{\beta}_s))^{w_{is}}}{\sum_{i=1}^{N}\prod_{s=1}^{S}\hat{\lambda}_s P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)} \tag{0.6}$$

Therefore,

$$E[w_{is} \mid y_i] = \frac{\hat{\lambda}_s P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)}{\sum_{s=1}^{S}\hat{\lambda}_s P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)} \tag{0.7}$$

M-step

After performing E-step the $E_s[l \mid \hat{\Lambda}, \hat{B}, \hat{D}]$ should be maximized such that $\lambda_s \in [0,1]$ and $\lambda_s = 1 - \sum_{s=1}^{S-1}\lambda_s$. We perform this step by using Lagrangian optimization method and obtain the following complete likelihood function and the corresponding first order derivatives as follows:

$$
\begin{aligned}
L &= \sum_{i=1}^{N}\sum_{s=1}^{S} E[w_{is} \mid y_i]\ln\hat{\lambda}_s + \sum_{i=1}^{N}\sum_{s=1}^{S} E[w_{is} \mid y_i]\ln P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s) - k(\sum_{s=1}^{S}\hat{\lambda}_s - 1) \\
\frac{\partial L}{\partial \hat{\lambda}_s} &= \sum_{i=1}^{N}\frac{E[w_{is} \mid y_i]}{\hat{\lambda}_s} = 0 \qquad \Leftrightarrow \qquad k = 0 \\
\frac{\partial L}{\partial \hat{\beta}_{js}} &= \sum_{i=1}^{N} E[w_{is} \mid y_i]\frac{\partial \ln P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)}{\partial \hat{\beta}_{js}} = 0 \\
\frac{\partial L}{\partial \hat{d}_s} &= \sum_{i=1}^{N} E[w_{is} \mid y_i]\frac{\partial \ln P(y_i \mid X_i, \hat{\beta}_s, \hat{d}_s)}{\partial \hat{d}_s} = 0
\end{aligned}
\tag{0.8}
$$

where k is Lagrange multiplier. Only one for the three equation from Equation 0.8 has an analytical solution, namely the first equation, last two equation cannot be solved analytically. Therefore, the Broyden - Fletcher - Goldfarb - Shanno(BFGS) algorithm will be used to solve the last two equations.

FMNB Model Selection

We use Finite Mixture model in combination with negative binomial distribution as a classification method, as we mentioned earlier. In K-means we, in advance, have assumed that there should be 3 clusters of customers: Good, Better and Best. However, with this model, we aim to find the optimal number of clusters using the data itself. However, it is impossible to select the number of segments in RFM model by using the standard t-test. The LR test for $_s = 0$ has no standard deviation and consequently, parameters α_s and β_s are not identified in this case. This problem is known as Devis problem and therefore, to avoid it, we use Bayesian Information Criteria(BIC) to evaluate the performance of the model for the different number of segments. The consistency property of BIC means that it is guaranteed to select the true model as the sample size grows infinitely large. BIC overcomes AIC's problem by making the parameter inclusion threshold more stringent as the sample size grows. More specifically, according to **?**, the Type I error of BIC goes to zero whereas the AIC's does not. It is known that AIC will fail to select the true model with non-vanishing probability as N becomes large, even in the case that the true model is under consideration. Thus, the consistency of BIC makes it quite attractive. However, the situation changes, in the case that the number of parameters in the true model is infinite. Then the AIC is asymptotically efficient in MSE of estimation, whereas BIC is not. In our case, there are few parameters in the model, and the number of observations is quite large which indicates to use the BIC information criterion for model selection. The BIC is defined as follows:

$$
\begin{aligned}
BIC &= -2\ln L + C\ln\sum_{i=1}^{N} T_i \\
C &= S - 1 + (J+1)S
\end{aligned}
\tag{0.9}
$$