The EM algorithm for multi-dimensional Gaussian mixture model

Qian Wang*, Jian Wang**

* Department of Statistics, School of Mathematics and Statistics, Shandong University of Technology, Zibo, China
** Department of Statistics, School of Mathematics and Statistics, Shandong University of Technology, Zibo, China

Abstract- One of the research hotspots of the expectation-maximization (EM) algorithm is the parameter estimation of mixture models. Taking the Gaussian mixture model (GMM) in a multi-dimensional case as an example, we use the EM algorithm to estimate its parameters, and get the corresponding iterative formulas. Then, use R software to carry out numerical simulation experiments on the EM algorithm, and explore its convergence and sensitivity.

Index Terms- Gaussian mixture model, EM, Multi-dimensional, Sensitivity.

I. INTRODUCTION

The problem of parameter estimation with mixture models has been widely concerned. Before the development of computer, scholars generally used the moment method to estimate parameters of mixture models. Person [1] used moment method to solve the parameters of two component GMM. However, the moment method needs to solve nonlinear equations, and the increase of model components will increase the number of equations, which greatly increases the computational complexity. Bhning [2] summarized some algorithms for maximum likelihood estimation in mixture models, and pointed out that the EM algorithm proposed by Dempster et al. [3] is one of the most conventional algorithms for maximum likelihood estimation. The proposal of the EM algorithm advances the related research on the parameter estimation of the mixture model. The program of EM algorithm is simple and easy to implement. For its relevant theoretical research refer to Xu et al. [4]. Specifically, given the initial value of the model parameter $\boldsymbol{\theta}^{(0)}$, note that $\boldsymbol{\theta}^{(t)}$ is the estimated value of the parameter $\boldsymbol{\theta}$ of the t-th iteration. At iteration t + 1, E- step needs to calculate:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E(\log f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}).$$

The M step is to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ so that

$$Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) = \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

Repeat E-step and M-step until $\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|$ or $\|Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})\|$ to a sufficiently small value.

In view of the wide application of the GMM, Ali et al. [5] applied the GMM to moving target detection, and McNicholas et al. [6] applied the GMM to clustering analysis of microarray expression data. This paper aims at the problem of parameter estimation of the multi-dimensional GMM, and uses the EM algorithm and maximum likelihood estimation principle to derive the update formula of each unknown parameter. A large number of numerical simulation experiments are carried out by using the EM algorithm to verify the convergence of EM algorithm, and the sensitivity of the algorithm is discussed from some aspects.

II. PRELIMINARIES

Definition 1 Suppose $\boldsymbol{x}$ is a p-dimensional random vector that the probability density function is given by

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

where $\boldsymbol{x} \in E_p, E_p = \{\boldsymbol{x}: -\infty < x_i < \infty, for\ i = 1,2,\cdots,p\}$. Then it is said that $\boldsymbol{x}$ has a multivariate Gaussian distribution, denoted as $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Definition 2 The mixed distribution model takes the following form

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k f(\boldsymbol{y}|\boldsymbol{\psi}_k),$$

where $\theta = (\alpha_1, \cdots, \alpha_K, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_K)$ denotes the vector of unknown parameters, and $\boldsymbol{\xi}_i$ consists of the elements of the $\boldsymbol{\psi}_k$ known a priori to be distinct. K is the mixture component number, $f(y|\boldsymbol{\psi}_k)$ is the k-th component density function, α_k is the mixing weight or probability of the $f(y|\boldsymbol{\psi}_k)$ for $\sum_{k=1}^{K} \alpha_k = 1$, see also Geoffrey et al. [7].

III. MAIN RESULTS

Theorem 1 Let $\mathbf{y} = (\boldsymbol{y_1}, \cdots, \boldsymbol{y_n})$ be a randon sample from the following GMM

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k f(\boldsymbol{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$
$$= \sum_{k=1}^{K} \alpha_k \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_k)\right),$$

where $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \cdots, \alpha_K, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_K)$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ represent the mean vector and covariance matrix of the k-th Gaussian distribution, respectively. Then at iteration t + 1, the maximum likelihood estimation of $\boldsymbol{\theta}$ obtained by the EM algorithm is

$$\alpha_k^{(t+1)} = \frac{\sum_{i=1}^n w_{ik}^{(t)}}{n}, \boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^n w_{ik}^{(t)} \boldsymbol{y}_i}{\sum_{i=1}^n w_{ik}^{(t)}},$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^n w_{ik}^{(t)} (\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t+1)}) (\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{i=1}^n w_{ik}^{(t)}},$$

where $w_{ik}^{(t)} = \alpha_k^{(t)} f\left(\boldsymbol{y}_i \big| \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right) / \sum_{k'=1}^K \alpha_{k'}^{(t)} f\left(\boldsymbol{y}_i \big| \boldsymbol{\mu}_{k'}^{(t)}, \boldsymbol{\Sigma}_{k'}^{(t)}\right)$, $\boldsymbol{y}_i \in R^{1 \times p}$.

Proof Since the density of $\boldsymbol{y}_i$ conditioning on $\boldsymbol{\theta}$ is given by

$$f(\boldsymbol{y}_i | \boldsymbol{\theta}) = \sum_{k=1}^K \alpha_k f(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$= \sum_{k=1}^K \alpha_k \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_k)\right),$$

the likelihood and log-likelihood functions are given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n f(\boldsymbol{y}_i | \boldsymbol{\theta}) = \prod_{i=1}^n \sum_{k=1}^K \alpha_k f(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

and

$$l(\boldsymbol{\theta}) = \sum_{i=1}^n log\left\{\sum_{k=1}^K \alpha_k f(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}.$$

Maximizing the $l(\boldsymbol{\theta})$ is difficult, because the sum sign in log is not conducive to derivation. Next, we use the EM algorithm to solve the problem. Indicator variable $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n)$ is introduced, where $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{iK})$, which is defined as follows

$$z_{ik} = \begin{cases} 1, & \boldsymbol{y}_i \text{is from the component } k, \\ 0, & \text{otherwise.} \end{cases}$$

There is and only one takes 1 in $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{iK})$. In this case, the likelihood and log-likelihood functions are rewritten as

$$L'(\boldsymbol{\theta}) = \prod_{i=1}^n f(\boldsymbol{y}_i, \boldsymbol{z}_i | \boldsymbol{\theta})$$

$$= \prod_{i=1}^n f(\boldsymbol{y}_i | \boldsymbol{\theta}, \boldsymbol{z}_i) f(\boldsymbol{z}_i | \boldsymbol{\alpha})$$

$$= \prod_{i=1}^n \left(\left(\prod_{k=1}^K f(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{ik}}\right)\left(\prod_{k=1}^K \alpha_k^{z_{ik}}\right)\right)$$

$$= \prod_{i=1}^n \prod_{k=1}^K \left(\alpha_k f(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}},$$

and

$$l'(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K z_{ik}(log\,\alpha_k + log\,f(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)).$$

Note that $\boldsymbol{\theta}^{(t)}$ is the estimated vector of the parameter vector $\boldsymbol{\theta}$ at iteration t, then at the E-step, calculate the Q function as

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) = E\left(l'(\boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{\theta}^{(t)}\right)$$

$$= \sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(t)} \left(log\,\alpha_k^{(t)} + log\,f\left(\boldsymbol{y}_i \big| \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)\right)$$

where

$$E\left(z_{ik} | \boldsymbol{y}, \boldsymbol{\theta}^{(t)}\right) = w_{ik}^{(t)}$$

$$= \alpha_k^{(t)} f\left(\boldsymbol{y}_i \big| \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right) \Big/ \sum_{k'=1}^K \alpha_{k'}^{(t)} f\left(\boldsymbol{y}_i \big| \boldsymbol{\mu}_{k'}^{(t)}, \boldsymbol{\Sigma}_{k'}^{(t)}\right).$$

At the M-step, find the partial derivative of $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)})$ with respect to the unknown parameters, and set the derivative to 0 to obtain the parameter estimation of the iteration t+1 under the EM:

$$\alpha_k^{(t+1)} = \frac{\sum_{i=1}^n w_{ik}^{(t)}}{n}, \boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^n w_{ik}^{(t)} \boldsymbol{y}_i}{\sum_{i=1}^n w_{ik}^{(t)}},$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^n w_{ik}^{(t)} (\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t+1)}) (\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{i=1}^n w_{ik}^{(t)}}.$$

IV. EXPERIMENT ANALYSIS

In this section, we simulate four scenarios to explore the convergence and sensitivity of the EM algorithm from the perspectives of n, p, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Here we examine the running speeds of the EM algorithm by using the iterative time, and examine the performance by using the mean absolute error (MAE), which is expressed as

$$MAE(\hat{\theta}) = \frac{1}{K + Kp + Kp^2} \sum_{k=1}^K \left(|\alpha_k - \hat{\alpha}_k| + \|\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k\|_1 + \|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k\|_1\right)$$

Scenario1: Effect on the sample size n. We fix $K = 2$, $p = 2$, $\boldsymbol{\alpha} = (0.3, 0.7)$, $\boldsymbol{\mu} = \begin{pmatrix} 0.1 & 0.5 \\ 1 & 1.5 \end{pmatrix}$, $\boldsymbol{\Sigma}_k = diag(0.1, 0.1)$ and vary the sample size n with $n = \{10^4, 2 \times 10^4, 3 \times 10^4, 4 \times 10^4\}$. Initial values of parameters are set to $\boldsymbol{\alpha}^{(0)} = (0.4, 0.6)$, $\boldsymbol{\mu}^{(0)} = \begin{pmatrix} 0.2 & 0.6 \\ 1.2 & 1.6 \end{pmatrix}$, and $\boldsymbol{\Sigma}_k^{(0)} = diag(0.15, 0.15)$, respectively.

Scenario2: Effect on the p . We fix $n = 10^4$, $K = 2$, $\boldsymbol{\alpha} = (0.3, 0.7)$, $\boldsymbol{\mu}_k \sim U(k, k + 1)$, $\boldsymbol{\Sigma}_k = diag(0.1_1, \cdots, 0.1_p)$ and vary the p with $p = \{2, 3, 4, 5\}$. Initial values of parameters are set to $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}$, $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}$, and $\boldsymbol{\Sigma}_k^{(0)} = \boldsymbol{\Sigma}_k$, respectively.

Scenario3: Effect on the $\boldsymbol{\mu}$. We fix $n = 10^4$, $K = 2$, $p = 2$, $\boldsymbol{\alpha} = (0.3, 0.7)$, $\boldsymbol{\Sigma}_k = diag(0.1, 0.1)$ and vary the $\boldsymbol{\mu}$ with $\boldsymbol{\mu}_a = \begin{pmatrix} 0.1 & 0.5 \\ 1 & 1.5 \end{pmatrix}$, $\boldsymbol{\mu}_b = \begin{pmatrix} 0.1 & 0.5 \\ 2 & 2.5 \end{pmatrix}$, $\boldsymbol{\mu}_c = \begin{pmatrix} 0.1 & 0.5 \\ 3 & 3.5 \end{pmatrix}$, and $\boldsymbol{\mu}_d = \begin{pmatrix} 0.1 & 0.5 \\ 4 & 4.5 \end{pmatrix}$, respectively. Initial values of parameters are set to $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}$, $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}$, and $\boldsymbol{\Sigma}_k^{(0)} = \boldsymbol{\Sigma}_k$, respectively.

Scenario4: Effect on the $\boldsymbol{\Sigma}$. We fix $n = 10^4$, $K = 2$, $p = 2$, $\boldsymbol{\alpha} = (0.3, 0.7)$, $\boldsymbol{\mu} = \begin{pmatrix} 0.1 & 0.5 \\ 1 & 1.5 \end{pmatrix}$ and vary the $\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma}_a = (\boldsymbol{\Sigma}_1, diag(0.1, 0.1))$, $\boldsymbol{\Sigma}_b = (\boldsymbol{\Sigma}_1, diag(0.1, 0.2))$, $\boldsymbol{\Sigma}_c = (\boldsymbol{\Sigma}_1, diag(0.1, 0.3))$, $\boldsymbol{\Sigma}_d = (\boldsymbol{\Sigma}_1, diag(0.1, 0.4))$, $\boldsymbol{\Sigma}_e = (\boldsymbol{\Sigma}_1, diag(0.1, 0.5))$, where $\boldsymbol{\Sigma}_1 = diag(0.1, 0.1)$. Initial values of parameters are set to $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}$, $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}$, and $\boldsymbol{\Sigma}_k^{(0)} = \boldsymbol{\Sigma}_k$, respectively.

4.1 Convergence analysis

Taking scenario 1 as an example, we calculated the log-likelihood value at each iteration and showed the result in Figure 1. It can be seen from Figure 1 that as the number of iterations increase, the log-likelihood values gradually increase and tend to stable. It is shown that the EM algorithm has obvious convergence when solving the GMM.
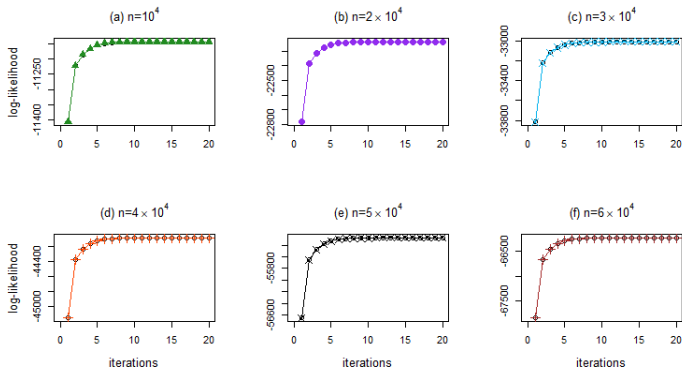

Figure 1: The values of the log-likelihood with different n in simulation

4.2 Sensitivity analysis

We summarized the results of the iteration time and MAE value of scenarios 1-4 in Figure 2. It can be seen from Panels (a) and (e) of Figure 2 that the increase of the sample size n will increase the iteration time of the algorithm, but when the sample size starts from $n = 4 \times 10^4$, the value of MAE is gradually decreasing, that is to say, the estimated value is closer to the true value. It can be seen from Panels (b) and (f) of Figure 2 that with the change of p, the increase trend of the iteration time and the value of MAE is almost the same. From Panels (c) and (g) of Figure 2, it can be seen that with the change of μ, the increase trend of the iteration time and the value of MAE shows a downward trend, that is, when the mean vector of each component is farther away, it will reduce the calculation time of the EM algorithm and improve the accuracy of the estimation. It can be seen from Panels (d) and (h) of Figure 2 that, on the contrary to the change trend of μ, when the covariance matrix of each component differs greatly, the calculation time of the EM algorithm will increase, and the value of MAE will increase.
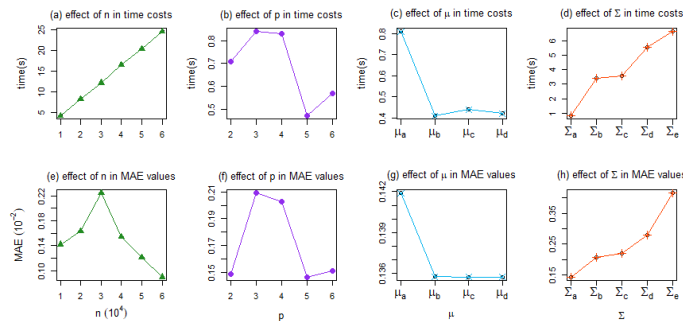

Figure 2: Performance results of the EM in simulation

V. CONCLUSION

In this paper, many simulation experiments have proved that the EM algorithm has obvious convergence when estimating the parameters of the multi-dimensional GMM. And in the case of a large sample size, if the distance between the mean vector of each component is far, and the covariance matrix difference of each component is small, the parameter value estimated by the EM algorithm will be closer to the true value.

APPENDIX

The R code of the EM algorithm is presented. (EM-code.docx)

REFERENCES

[1] Pearson, K. (1894). "Contributions to the Mathematical Theory of Evolution," Philosophical Transactions of the Royal Society A, 185, 71–110.

[2] Bhning, D. . (1995). "A review of reliable maximum likelihood algorithms for semiparametric mixture models," Journal of Statistical Planning and Inference, 47(1-2), 5-28.

[3] Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39(1), 1-22.

[4] Xu, L. , and Jordan, M. . (1996). "On convergence properties of the EM algorithm for Gaussian mixtures," Neural Computation, 8(1), 129-151.

[5] Ali, I. , Mille, J. , & Tougne, L. . (2014). "Adding a rigid motion model to foreground detection: application to moving object detection in rivers," Pattern Analysis & Applications,17(3), 567-585.

[6] Mcnicholas, P. D. , & Brendan, M. T. . (2010). "Model-based clustering of microarray expression data via latent Gaussian mixture models," Bioinformatics(21), 2705-12.

[7] Geoffrey J. McLachlan, Sharon X. Lee, and Suren I. Rathnayake. (2019). "Finite mixture models," Annual Review of Statistics and Its Application, 6(1), 355-378.

AUTHORS

First Author – Qian Wang, Post graduate student, Shandong University of Technology, waqian0715@163.com.

Second Author – Jian Wang, Associate Professor, Shandong University of Technology, wjzhenhua@126.com.

Correspondence Author – Jian Wang, wjzhenhua@126.com, 13706430826.