Gaussian mixture modelling by exploiting competitive stop EM algorithm

Kexin Jia\textsuperscript{1,*}, Yuxia Xin\textsuperscript{2} and Ting Cheng\textsuperscript{3}

\textsuperscript{1}East China Research Institute of Electronic Engineering
\textsuperscript{2}Hefei Information Technology University
\textsuperscript{3}University of Electronic Science and Technology

jiakexin@sina.com

Abstract. To improve the robustness of order selection and parameter learning for Gaussian mixture model (GMM), this paper proposes a competitive stop expectation-maximization (EM) algorithm, which is based on two stop conditions. The first condition is a Lilliefors test based multivariate (MV) normality criterion, which is used to determine whether to split a component into two different components. The EM algorithm stops splitting when all components have MV normality. The minimum description length (MDL) criterion is used in the second condition, which competes with the first condition to prevent the EM algorithm from over-splitting. Simulation experiments verify the effectiveness of the proposed algorithm.

1. Introduction
In univariate and multivariate statistical processing, the GMM is a very flexible and powerful probabilistic statistical model. This model is widely used in various fields of statistical modeling and analysis of observed data [1][2], such as signal processing, image processing, machine learning, pattern recognition, computer vision, etc.

The classic GMM fitting approach is the standard EM algorithm. The algorithm can obtain a maximum likelihood (ML) estimate of the parameters of the mixture model. However, when applying the standard EM algorithm to fit a GMM, we will face two problems [3][4]: the first problem is that the algorithm must know the order of GMM in advance, and the other problem is that the algorithm is very sensitive to the initial model parameters.

To solve the above problem, several EM variant algorithms are proposed. In [5], the proposed algorithm adds new components using a Bayesian splitting procedure. In [6], an EM variant algorithm is proposed which starts with only a component and uses a Mahalanobis distance-based normality test as split criterion. This algorithm stops when no component derives from MV normality distribution. The algorithm proposed in [7] also starts with only a component and employs entropy based Gaussian deficiency as split criterion.

To improve the robustness of order selection and parameter learning for GMM, a novel competitive stop EM (CSEM) algorithm is proposed. This algorithm is based on two stop conditions, starts with only one component, and finds the ML solution. The first condition is a Lilliefors test based multivariate (MV) normality criterion, which is used to determine whether to split a component into two different components and stop EM algorithm when no component diverges from MV Gaussian. The MDL criterion is used in the second condition, which prevents EM algorithm from over-splitting. This paper
is organized as follows. In section 2, The proposed CSEM algorithm are described. In section 3, the effectiveness of the algorithm is verified by simulation experiments. In section 4, the full paper is summarized.

2. CSEM algorithm
Let the observation sample set \( X \) be used to fit a GMM, the model includes \( K \) components \( \{L_k\}_{k=1}^K \), and each component \( L_k \) obeys a Gaussian distribution. The goal of the CSEM algorithm is to find \( K \) components \( \{L_k\}_{k=1}^K \). Fig.1 shows the processing flow of the CSEM algorithm.

![Processing flow of the CSEM algorithm](image)

**Figure 1.** Processing flow of the CSEM algorithm.

2.1. MV normality test
MV normality test is employed to determine whether to split a component into two different components and stop EM algorithm when no component diverges from MV Gaussian. This paper replaces a D-dimension normality test with D one-dimension normality tests which are based on Lilliefors test and perform normality test on each marginal sample set of \( X \).

Given \( L_k = \{x_{i1}, x_{i2}, \ldots, x_{ik}\}, k = 1, \ldots, K \), where, \( x_{ij} = [x_{i1}^{(j)}, \ldots, x_{iD}^{(j)}]^T \) is the \( i \)th sample vector in \( L_k \), \( N_k = |L_k| \) is the cardinality of \( L_k \). Let the Lilliefors test statistics of the \( d \)th marginal sample set \( \{x_{ij}^{(d)}\}_{i=1}^{N_k} \), denote \( KS_d^{(d)} \), that is,

\[
KS_d^{(d)} = \max_{i=1,2,\ldots,N_k} \left| \hat{F}_d \left( x_{ij}^{(d)} \right) - F_d \left( x_{ij}^{(d)} \right) \right|
\]

where \( F_d \left( x_{ij}^{(d)} \right) \) is the theoretical marginal Gaussian cumulative distribution function (MGCDF), and \( \hat{F}_d \left( x_{ij}^{(d)} \right) \) is the empirical MGCDF.

For \( \forall k = 1, \ldots, K \), If \( \max_d \{ KS_d^{(d)} \} \leq KS \), the \( k \)th component obeys a MV normal distribution, and the EM algorithm stops. Otherwise, the algorithm finds the component to be split. \( KS \) is the critical value of Lilliefors test statistics at significant level \( \lambda \).
In addition, the significant level $\eta$ of a MV normality test can be approximately computed as $\eta = 1 - (1 - \lambda)^\nu$. The above equation assumes that $KS_i^{(k)}$ are statistically independent and use the same significant level $\lambda$.

2.2. Search the component to be split

Let $\Omega$ be an index set, which includes the indexes of all the components that do not obey the MV normal distribution.

$$\Omega = \left\{ k \mid \max_j \left\{ KS_j^{(k)} \right\} > KS_k, k = 1, \ldots, K \right\}$$

(2)

For any component $L_k$ ($k \in \Omega$), its entropy can be expressed as

$$H(k) = -\int p(x|\theta) \log p(x|\theta) dx.$$  

(3)

Therefore, the index $k^*$ of the component to be split can be obtained by the following formula

$$k^* = \arg \max_{k \in \Omega} \{ H(k) \}$$

(4)

2.3. Splitting process

For cluster $L_k$, the mean $m_{\rho'}$ of the $p$th marginal sample set is computed by

$$m_{\rho'} = \frac{1}{N_{p'}} \sum_{j=1}^{N_{p'}} x_{k, j}^{(p')}$$

(5)

where $p'$ is determined by the following expression.

$$p' = \arg \max_p \{ KS_{p'}^{(k)} \}$$

(6)

In this section, the hyperplane $x = m_{\rho'}$ is used to split the component $L_k$ into two components $L_{k'}$ and $L_{k''}$ ($k' = K + 1$). The mean, covariance matrix and other information of these two components are used to initialize the standard EM algorithm.

2.4. Standard EM algorithm

The standard EM algorithm iteratively fits the GMM, with initial $\hat{\Theta}(0) = \{ \hat{\alpha}_k(0), \hat{\mu}_k(0), \hat{\Sigma}_k(0) \}_{k=1}^{K'}$ set as

$$\hat{\alpha}_k(0) = \left| L_k \right| \left| X \right|, k = 1, \ldots, K'$$

(7)

$$\hat{\mu}_k(0) = \frac{1}{\left| L_k \right|} \sum_{x \in L_k} x$$

(8)

$$\hat{\Sigma}_k(0) = \frac{1}{\left| L_k \right|} \sum_{x \in L_k} (x - \hat{\mu}_k(0))(x - \hat{\mu}_k(0))^T$$

(9)

The standard EM algorithm stops when

$$\frac{L(X|\hat{\Theta}(t+1)) - L(X|\hat{\Theta}(t))}{L(X|\hat{\Theta}(t))} < \varepsilon$$

(10)

where $\varepsilon > 0$. In practical applications, the variable $\varepsilon$ is usually equal to $10^{-5}$. 
2.5. Competitive stop condition
This paper proposes to use the MDL criterion as a competitive stop condition. When this condition is satisfied, the standard EM algorithm does not iteratively fit the GMM. This condition prevents the EM algorithm from over-splitting. The MDL criterion can be expressed as

$$L_2(\Theta) = \frac{1}{2} \left[ K - 1 + K \left( D + \frac{D(D+1)}{2} \right) \right] \log(N) - \log p(X|\Theta)$$

(11)

where, $p(X|\Theta)$ is the PDF of the GMM.

If $L_2^{(k)} < L_2^{(k-1)}$, the CSEM algorithm stops splitting and obtains a GMM with $K$ components. $L_2^{(k)}$ is the MDL value when the GMM order is $K$.

3. Experimental results
This section demonstrates the effectiveness of the proposed CSEM algorithm through three examples.

Example 1: The GMM in this example includes 5 Gaussian components, and each Gaussian component produces 500 samples. For the normality test, we set the significant level to 0.05. The parameters for the mixture generation are

$$\alpha_i = 0.2, \mu_i = [-1, -2]^T, \Sigma_i = \begin{bmatrix} 0.6 & 0.15 \\ 0.15 & 0.6 \end{bmatrix}$$

(12)

Fig. 2 shows the iterative process of the proposed CSEM algorithm. The algorithm starts by treating all samples as one component and computes the corresponding mean and covariance matrices. After many iterations, the correct model order is estimated. After 100 repeated experiments, the clustering correctness of proposed algorithm is 100% and the average EM iteration is 34.47. Fig.2(a) is the set of
original sample vectors with \( L_{2}^{(i)} = 1212.8 \). Fig. 2 (b), (c), (d), and (f) are the results of the first, second, third and fifth splitting operations with \( L_{2}^{(2)} = 974.3 \), \( L_{2}^{(3)} = 910.0 \), \( L_{2}^{(4)} = 870.9 \) and \( L_{2}^{(5)} = 852.3 \). Fig. 2 (e) is the result of the fourth splitting operations and this is the final splitting result because \( L_{2}^{(6)} = 850.7 \), competitive stop condition satisfies.

Example 2: To compare with other EM variants, we select both algorithms presented in literature [8] and [7]. The first is named as Backward-MDL2-EM (BMEM) algorithm. Its initialization method in this experiment is the same as that discussed in literature [8]. The second is named as Mahalanobis distance-based EM (MDEM) algorithm. This example uses three sets of test samples given in literature [7], each of which includes 500 test samples. Each algorithm is repeated 1000 times, and the correctness of the model order, the average and standard deviation of the iterations are computed.

The correctness of model order selection and the average and standard deviation (STD) of EM iterations, are performed for comparison. The results for the above three data sets are presented in Table I. It can be inferred that the number of EM iterations of proposed EM algorithm is very close to that of MDEM and has relatively low EM iterations when compared with BMEM algorithm. For data sets A and C, the mixture components do not overlap with each other, the correctness of model order selection is 100%. For data set B, the clustering correctness of our proposed CSEM algorithm is 84.8%.

Table 1. The results of three different EM algorithm for test sample sets.

| Algorithm     | Data set | Correctness (% | Average EM iterations | STD EM iterations |
|---------------|----------|----------------|------------------------|-------------------|
| CSEM \( \eta = 0.05 \) | A        | 100            | 35.861                 | 14.1943           |
|               | B        | 84.8           | 103.377                | 34.47             |
|               | C        | 100            | 179.783                | 34.9575           |
| MDEM          | A        | 87.6           | 24.451                 | 15.4107           |
|               | B        | 78.8           | 94.919                 | 38.6933           |
|               | C        | 72.9           | 126.833                | 38.9059           |
| BMEM \( k_{\text{min}} = 1, k_{\text{max}} = 10 \) | A        | 59.30          | 205.969                | 59.1639           |
|               | B        | 41.50          | 206.513                | 53.2611           |
|               | C        | 65.10          | 246.007                | 56.3084           |

Example 3: Mixtures of factor analyzers (MFA), proposed in [9], are basically mixtures of Gaussians. To verify the effectiveness of fitting an MFA, this example will use the proposed algorithm on a noisy shrinking spiral data. Fig. 3 shows that the proposed algorithm successfully extracts a one-dimensional manifold from the noisy shrinking spiral data. The selected model order is \( K = 12 \). We repeat the fitting algorithm 30 times, the model order selects \( K = 10 \) 2 times, \( K = 11 \) 11 times, \( K = 12 \) 12 times, and \( K = 13 \) 5 times.
4. Conclusion
To make the algorithm insensitive to the initial parameters and avoid the algorithm from converging to the local optimal solution, this paper combines Lilliefors test, MDL criterion and standard EM algorithm, and proposes a competitive stop EM algorithm for fitting a GMM. The algorithm is initialized as a single Gaussian component, and the Lilliefors test is used to check the normality of each component in the GMM. The algorithm selects the component with the largest entropy for splitting among the components that do not have normality. The MDL criterion acts as a competitive stop condition to prevent the algorithm from over-splitting. It is worth noting that this paper only uses a noisy shrinking spiral data to verify the effectiveness of the proposed algorithm in fitting an MFA. How to apply the proposed algorithm to other practical areas is the focus of further research.

References
[1] L. Zhao, J. Ma. A dynamic model selection algorithm for mixtures of Gaussian processes[J]. 2016 IEEE 13th International Conference on Signal Processing, 2016, 13:1095-1099.
[2] R. Fan, R. Huang, R. Diao. Gaussian Mixture Model-Based Ensemble Kalman Filter for Machine Parameter Calibration[J]. IEEE Transactions on Energy Conversion, 2018,33(3):1597-1599.
[3] G. McLachlan, D. Peel. Finite Mixture Models [M]. New York: John Wiley & Sons, 2000.
[4] L. Xu, M. Jordan. On Convergence Properties of the EM algorithm for Gaussian Mixtures [J]. Neural Computation, 1996, 8: 129-151.
[5] C. Constantinopoulos, A. Likas. Unsupervised learning of Gaussian mixtures based on variational component splitting [J]. IEEE transactions on Neural Networks, 2007, 18(3):745-755.
[6] D. Ververidis, C. Kotropoulos. Gaussian mixture modeling by exploiting the Mahalanobis distance [J]. IEEE transaction on signal processing, 2008, 56(7): 2797-2811.
[7] A. P. Benavent, F. E. Ruiz, J. M. Saez. Learning Gaussian mixture models with entropy-based criteria [J]. IEEE transactions on neural networks, 2009, 20(11): 1756-1771.
[8] M. A. Figueiredo, A. K. Jain. Unsupervised learning of finite mixture models [J]. IEEE transactions on pattern analysis and Machine intelligence, 2002, 24(3): 381-396.
[9] Z. Ghahramani, G. Hinton. The EM algorithm for mixtures of factor analyzer [M], Technical report CRG-TR-96-1, Univ. of Toronto, Canada, 1997.