The Competitive EM Algorithm for Gaussian Mixtures with BYY Harmony Criterion

Hengyu Wang, Lei Li, and Jinwen Ma*

Department of Information Science, School of Mathematical Sciences and LAMA, Peking University, Beijing, 100871, China
jwma@math.pku.edu.cn

Abstract. Gaussian mixture has been widely used for data modeling and analysis and the EM algorithm is generally employed for its parameter learning. However, the EM algorithm may be trapped into a local maximum of the likelihood and even leads to a wrong result if the number of components is not appropriately set. Recently, the competitive EM (CEM) algorithm for Gaussian mixtures, a new kind of split-and-merge learning algorithm with certain competitive mechanism on estimated components of the EM algorithm, has been constructed to overcome these drawbacks. In this paper, we construct a new CEM algorithm through the Bayesian Ying-Yang (BYY) harmony stop criterion, instead of the previously used MML criterion. It is demonstrated by the simulation experiments that our proposed CEM algorithm outperforms the original one on both model selection and parameter estimation.

Keywords: Competitive EM (CEM) algorithm, Gaussian mixture, Bayesian Ying-Yang (BYY) harmony learning, Model selection

1 Introduction

As a powerful statistical tool, Gaussian mixture has been widely used in the fields of signal processing and pattern recognition. In fact, there already exist several statistical methods for the Gaussian mixture modeling, such as the k-means algorithm [2]) and the Expectation-Maximization (EM) algorithm [3]. However, the EM algorithm cannot determine the correct number of Gaussians in the mixture for a sample data set because the likelihood to be maximized is actually a increasing function of the number of Gaussians. Moreover, a "bad" initialization usually makes it trapped at a local maximum, and sometimes the EM algorithm converges to the boundary of the parameter space.

Conventionally, the methods of model selection for Gaussian mixture, i.e., determining a best number k^* of Gaussians for a sample data set, are based on certain selection criteria such as Akaike's Information Criterion (AIC) [4], Bayesian Information Criteria (BIC) [5], and Minimum Message Length (MML) criterion [6]. However, these criteria have certain limitations and often lead to a wrong result.

* Corresponding author.

D.-S. Huang et al. (Eds.): ICIC 2008, LNCS 5226, pp. 552–560, 2008.

Recently, with the development of the Bayesian Ying-Yang (BYY) harmony learning system and theory [7,8], there have emerged a new kind of learning algorithms [10,10,11,12] on the Gaussian mixture modeling based on the maximization of the harmony function, which is equivalent to the BYY harmony learning on certain architectures of the BYY learning system related to the Gaussian mixture model. These learning algorithms can automatically determine the number of Gaussians for the sample data set during parameter learning. Moreover, the successes of these algorithms also show that the harmony function can be served as an efficient criterion of model selection on Gaussian mixture.

Although these BYY harmony learning algorithms are quite efficient for the Gaussian mixture modeling, especially on automated model selection, they need an assumption that the number k of Gaussians in the mixture should be slightly larger than the true number k^* of Gaussians in the sample data. In fact, if k is smaller or too much larger than k^*, they may converge to a wrong result. On the other hand, it is still a difficult problem to estimate a reasonable upper bound of k^* with a sample data set. One possible way to overcome this difficulty is to introduce the split-and-merge operation or competitive mechanism into the EM algorithm to make the model selection dynamically [13]. However, the MML model selection criterion used in such a competitive EM (CEM) algorithm is not very efficient for the model selection on Gaussian mixture.

In this paper, we propose a new CEM algorithm which still works in a split-and-merge mode by using the BYY harmony criterion, i.e., the maximization of the harmony function, as a stop criterion. The simulation experiments demonstrate that our proposed CEM algorithm outperforms the original CEM algorithm on both model selection and parameter estimation.

2 The EM Algorithm for Gaussian Mixtures

We consider the following Gaussian mixture model:

$$p(x|\Theta_k) = \sum_{i=1}^{k} \alpha_i p(x|\theta_i) = \sum_{i=1}^{k} \alpha_i p(x|\mu_i, \Sigma_i), \tag{1}$$

where k is number of components in the mixture, $\alpha_i (\geq 0)$ are the mixing proportions of components satisfying $\sum_{i=1}^{k} \alpha_i = 1$ and each component density $p(x|\theta_i)$ is a Gaussian probability density function given by:

$$p(x|\theta_i) = p(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}, \tag{2}$$

where μ_i is the mean vector and Σ_i is the covariance matrix which is assumed positive definite. For clarity, we let Θ_k be the collection of all the parameters in the mixture, i.e., $\Theta_k = (\theta_i, \cdots, \theta_k, \alpha_1, \cdots, \alpha_k)$.

Given a set of N i.i.d. samples, $\mathcal{X} = \{x_t\}_{t=1}^N$, the log-likelihood function for the Gaussian mixture model is expressed as follows:

$$\log p(\mathcal{X}|\Theta_k) = \log \prod_{t=1}^{N} p(x_t|\Theta_k) = \sum_{t=1}^{N} \log \sum_{i=1}^{k} \alpha_i p(x_t|\theta_i), \tag{3}$$

which can be maximized to get a Maximum Likelihood (ML) estimate of Θ_k via the following EM algorithm:

$$\alpha_i^+ = \frac{1}{n} \sum_{t=1}^{N} P(i|x_t); \mu_i^+ = \sum_{t=1}^{N} x_t P(i|x_t) / \sum_{t=1}^{N} P(i|x_t) \tag{4}$$

$$\Sigma_i^+ = [\sum_{t=1}^{N} P(i|x_t)(x_t - \mu_i^+)(x_t - \mu_i^+)^T] / \sum_{t=1}^{N} P(i|x_t), \tag{5}$$

where $P(i|x_t) = \alpha_i p(x_t|\theta_i) / \sum_{i=1}^{k} \alpha_i p(x_t|\theta_i)$ are the posterior probabilities.

Although the EM algorithm can have some good convergence properties, it clearly has no ability to determine the appropriate number of the components for a sample data set because it is based on the maximization of the likelihood function. In order to overcome this difficulty, we will use the BYY harmony function instead of the likelihood function for our Gaussian mixture learning.

3 BYY Learning System and Harmony Function

In a BYY learning system, each observation $x \in X \subset \mathcal{R}^d$ and its corresponding inner representation $y \in Y \subset \mathcal{R}^m$ are described with two types of Bayesian decomposition: $p(x, y) = p(x)p(y|x)$ and $q(x, y) = q(y)q(x|y)$, which are called them Yang and Ying machine, respectively. Given a data set $D_x = \{x_1, \cdots, x_n\}$ from the Yang or observation space, the task of learning on a BYY system consist of specifying all the aspect of $p(y|x)$,$p(x)$,$q(x|y)$,$q(y)$ with a harmony learning principle implemented by maximizing the function:

$$H(p \parallel q) = \int p(y|x)p(x) \ln[q(x|y)q(y)]dxdy. \tag{6}$$

For the Gaussian mixture model, we let y be limited to be an integer variable $y = \{1, \cdots, k\} \subset R$ and utilize the following specific BI-Architecture of the BYY system:

$$p(x) = p_0(x) = \frac{1}{N} \sum_{t=1}^{N} G(x - x_t); \qquad p(y = i|x) = \alpha_i q(x|\theta_i)/q(x|\Theta_k); \tag{7}$$

$$q(x|\Theta_k) = \sum_{i=1}^{k} \alpha_i q(x|\theta_i); \quad q(y) = q(y = i) = \alpha_i > 0; \quad \sum_{i=1}^{k} \alpha_i = 1, \tag{8}$$

where $G(\cdot)$ is a kernel function and $q(x|y=i) = q(x|\theta_i)$ is a Gaussian density function. In this architecture, $q(y)$ is a free probability function, $q(x|y)$ is a Gaussian density, and $p(y|x)$ is constructed from $q(y)$ and $q(x|y)$ under the Bayesian law. Θ_k=$\{\alpha_i, \theta_i\}_{i=1}^k$ is the collection of the parameters in this BYY learning system. Obviously $p(x|\Theta_k) = \sum_{i=1}^k \alpha_i q(x|\theta_i)$ is a Gaussian mixture model under the architecture of the BYY system.

Putting all these component densities into Eq.(6) and letting the kernel functions approach the delta functions, $H(p||q)$ reduces to the following harmony function:

$$H(p \parallel q) = J(\Theta_k) = \frac{1}{N} \sum_{t=1}^{N} \sum_{i=1}^{k} \frac{\alpha_i q(x_t|\theta_i)}{\sum_{j=1}^k \alpha_j q(x_t|\theta_j)} \ln[\alpha_i q(x_t|\theta_i)]. \tag{9}$$

Actually, it has been shown by the experiments and theoretical analysis that as this harmony function arrives at the global maximum, a number of Gaussians will match the actual Gaussians in the sample data, respectively, with the mixing proportions of the extra Gaussians attenuating to zero. Therefore, the maximization of the harmony function can be used as a reasonable criterion for model selection on Gaussian mixture.

4 The Competitive EM Algorithm with BYY Harmony Criterion

4.1 The Split and Merge Criteria

In order to overcome the weaknesses of the EM algorithm, we can introduce certain split-and-merge operation into the EM algorithm such that a competitive learning mechanism can be implemented on the estimated Gaussians obtained from the EM algorithm. We begin to introduce the split and merge criteria used in [13].

In order to do so, we define the local density $f_i(x;\Theta_k)$ as a modified empirical distribution given by:

$$f_i(x;\Theta_k) = \sum_{t=1}^{N} \delta(x - x_t) P(i|x_t;\Theta_k) / \sum_{t=1}^{N} P(i|x_t;\Theta_k), \tag{10}$$

where $\delta(\cdot)$ is the Kronecker function and $P(i|x_t;\Theta_k)$ is the posterior probability. Then, the local Kullback divergence can be used to measure the distance between the local data density $f_i(x;\Theta_k)$ and the estimated density $p(x|\theta_i)$ of the i-th component in the mixture:

$$D_i(\Theta_k) = \int f_i(x;\Theta_k) \log \frac{f_i(x;\Theta_k)}{p(x|\theta_i)} dx. \tag{11}$$

Thus, the split probability of the i-th component is assumed to be proportional to $D_i(\Theta_k)$:

$$P_{split}(i;\Theta_k) = (1/N_{\Theta_k}) \cdot D_i(\Theta_k), \tag{12}$$

where N_{Θ_k} is a regularization factor.

On the other hand, if the i-th and j-th components should be merged into a new component denoted as the i'-th component and the parameters of the new Gaussian mixture become Θ'_{k-1}, the merge probability of the i-th and j-th components is assumed to be inversely proportional to $D_{i'}(\Theta'_{k-1})$:

$$P_{merge}(i,k;\Theta_k) = (1/N_{\Theta_k}) \cdot (\beta / D_{i'}(\Theta'_{k-1})), \tag{13}$$

where β is also a regularization factor determined by experience. That is, as the split probability of the new component merged from the two old components becomes larger, the merge probability of these two components becomes smaller.

With the above split and merge probabilities, we can construct the following split and merge criteria.

Merge Criterion: If the i-th and j-th components gave the highest merge probability, we may merge them with the parameters θ_l of the new component or Gaussian as follows ([14]):

$$\alpha_l = \alpha_i + \alpha_j; \mu_l = (\alpha_i \mu_i + \alpha_j \mu_j)/\alpha_l; \tag{14}$$

$$\Sigma_l = \{\alpha_i[\Sigma_i + (\mu_i - \mu_l)(\mu_i - \mu_l)^T] \tag{15}$$

$$+\alpha_j[\Sigma_j + (\mu_j - \mu_l)(\mu_j - \mu_l)^T)]\}/\alpha_. \tag{16}$$

Split Criterion: If the r-th component has the highest split probability, we can split it into two components, called the i-th and j-th components. In order to do so, we get the singular value decomposition of the covariance matrix $\Sigma_r = USV^T$, where $S = diag[s_1, s_2, \cdots, s_d]$ is a diagonal matrix with nonnegative diagonal elements in a descent order, U and V are two (standard) orthogonal matrices. Then, we set $A = U\sqrt{S} = Udiag[\sqrt{s_1}, \sqrt{s_2}, \cdots, \sqrt{s_d}]$ and get the first column A_1 of A. Finally, we have the parameters for the two split components as follows ([14]):

$$\alpha_i = \gamma\alpha_r, \alpha_j = (1-\gamma)\alpha_r; \tag{17}$$

$$\mu_i = m_r - (\alpha_j/\alpha_i)^{1/2}\mu A_1, \mu_j = m_r + (\alpha_i/\alpha_j)^{1/2}\mu A_1; \tag{18}$$

$$\Sigma_i = (\alpha_j/\alpha_i)\Sigma_r + ((\eta - \eta\lambda^2 - 1)(\alpha_r/\alpha_i) + 1)A_1 A_1^T; \tag{19}$$

$$\Sigma_j = (\alpha_i/\alpha_j)\Sigma_r + ((\eta\lambda^2 - \eta - \lambda^2)(\alpha_r/\alpha_j) + 1)A_1 A_1^T, \tag{20}$$

where $\gamma, \mu, \lambda, \eta$ are all equal to 0.5. For clarity, we denote the parameters of the new mixture by Θ'_{k+1}.

4.2 The Proposed CEM Algorithm

The performance of the CEM algorithm strongly relies on the stop criterion. We now use the BYY harmony learning criterion as the stop criterion instead of the MML criterion in [6]. That is, our split-and merge EM learning process tries to maximize the harmony function $J(\Theta)$ given in Eq.(9). Specifically, in each iteration, if the harmony learning function $J(\Theta_k)$ increases, the split or merge operation will be accepted, otherwise it will be reject.

For quick convergence, during each stage of the algorithm, the component with a mixing proportion α_i being less than a threshold $\epsilon > 0$ will be discarded from the Gaussian mixture directly. With such a component annihilation mechanism, the algorithm can escape a lot of computation and converge to the solution more rapidly.

With all the preparations, we can summarize the CEM algorithm for Gaussian mixtures with the BYY harmony criterion as follows.

Step 1: Initialization. Select k and set the initial parameters Θ_k as randomly as possible. And set $l = 0$.

Step 2: Implement the (conventional) EM algorithm and get the new parameters as $\Theta_k(l)$ of the Gaussian mixture.

Step 3: Implement the Split and merge operations on the estimated components of the Gaussian mixture with $\Theta_k(l)$ and obtain the the new parameters $\Theta'_{k-1}(l)$ and $\Theta'_{k+1}(l)$ from the merge and split operations, respectively.

Step 4: Compute $Acc(M) = J(\Theta'_{k-1}(l)) - J(\Theta_k(l))$. If $Acc(M) > 0$, accept the merge operation, update the estimated mixture by $\Theta'_{k-1}(l)$, and go to Step 6. Otherwise, go to Step 5.

Step 5: Compute $Acc(S) = J(\Theta'_{k+1}(l)) - J(\Theta_k(l))$. If $Acc(S) > 0$, accept the split operation, update the estimated mixture by $\Theta'_{k+1}(l)$, and go to Step 6. Otherwise, stop the algorithm and get the parameters of the Gaussian mixture.

Step 6: Remove the components whose mixing proportions are lower than the pre-defined threshold $\epsilon > 0$. With the remaining parameters and k, let $l = l + 1$ and go to Step 2.

Clearly, this proposed CEM algorithm increases the harmony function on each stage and reach its maximum at end. The maximization of the harmony function as well as the EM algorithm guarantee the new CEM algorithm can lead to a good result on both model selection and parameter estimation for the Gaussian mixture modeling with a sample data set, which will be demonstrated by the simulation experiments in the next section.

5 Experimental Results

In this section, some simulation experiments were conducted to demonstrate the performance of the proposed CEM algorithm with BYY harmony criterion. Actually, the proposed CEM algorithm was implemented on two different synthetic data sets, each of which contained 3000 samples. Moreover, the proposed CEM algorithm was compared with the original CEM algorithm.

The first sample data set consisted of seven Guassians. As shown in the left subfigure of Fig.1, the proposed CEM algorithm was initialized with $k = 5$ from the parameters obtained by the k-means algorithm. After some iterations, as shown in the middle subfigure of Fig.1, one initial component was split into two Gaussians by the algorithm. Such a split process continued until all the

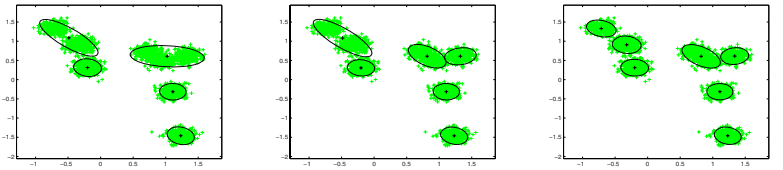

Fig. 1. The Experimental Results of the Proposed CEM Algorithm on the First Sample Data Set

components accurately matched the actual Gaussians in the sample data set, respectively, which is shown in the right subfigure of Fig.1. The algorithm stopped as $J(\Theta_k)$ reached its maximum and selected the correct number of Gaussians in the sample data set.

The second sample data set consisted of eight Gaussians. As shown in the left subfigure of Fig.2, the proposed CEM algorithm was implemented on the second sample data set initially with $k = 12$, which was much larger than the number of actual Gaussians. As shown in the middle subfigure of Fig.2, two pairs of the initially estimated Gaussians were selected to have been merged into two new Gaussians. The algorithm finally stopped with the eight actual Gaussians estimated accurately, being shown in the right subfigure of Fig.2.

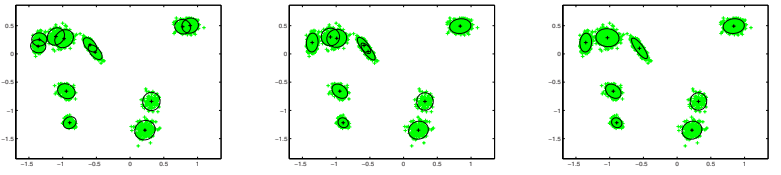

Fig. 2. The Experimental Results of the Proposed CEM Algorithm on the Second Sample Data Set

For comparison, we also implemented the original CEM algorithm with the MLL criterion [6] on the second sample data set. From Fig.3, it can be observed that the original CEM algorithm was initialized at the same situation, but led to

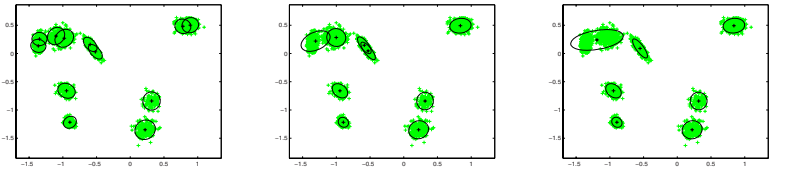

Fig. 3. The Experimental Results of the Original CEM Algorithm on the Second Sample Data Set

a wrong result on both model selection and parameter estimation at last. Therefore, our proposed CEM algorithm can outperform the original CEM algorithm on both model selection and parameter estimation in certain cases.

6 Conclusions

We have investigated the competitive EM algorithm from the view of the Bayesian Ying-Yang (BYY) harmony learning and proposed a new CEM algorithm with the BYY harmony criterion. The proposed competitive EM algorithm can automatically detect the correct number of Gaussians for a sample data set and obtain a good estimation of the parameters for the Gaussian mixture modeling through a series of the split and merge operations on the estimated Gaussians obtained from the EM algorithm. It is demonstrated well by the simulation experiments that the proposed CEM algorithm can achieve a better solution for the Gaussian mixture modeling on both model selection and parameter estimation on a sample data set.

Acknowledgements

This work was supported by the Natural Science Foundation of China for grant 60771061. The authors acknowledge Mr. Gang Chen for his helpful discussions.

References

1. Hartigan, J.A.: Distribution Problems in Clustering. Classification and Clustering. In: Van Ryzin, J. (ed.) Distribution Problems in Clustering, pp. 45–72. Academic press, New York (1977)
2. Jain, A.K., Dubes, R.C.: Algorithm for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximun Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society B 39, 1–38 (1977)
4. Akaike, H.: A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control AC-19, 716–723 (1974)
5. Scharz, G.: Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464 (1978)
6. Figueiredo, M.A.T., Jain, A.K.: Unsupervised Learning of Finite Mixture Models. IEEE Transations on Pattern Analysis and Machine Intelligence 24(3), 381–396 (2002)
7. Xu, L.: Best Harmony, Unified RPCL and Automated Model Selection for Unsupervised and Supervised Learning on Gaussian Mixtures, Three-Layer Nets and ME-RBF-SVM Models. International Journal of Neural Systems 11(1), 43–69 (2001)
8. Xu, L.: BYY Harmony Learning, Structural RPCL, and Topological Self-organizing on Mixture Modes. Neural Networks 15, 1231–1237 (2002)
9. Ma, J., Wang, T., Xu, L.: A Gradient BYY Harmony Learning Rule on Gaussian Mixture with Automated Model Selection. Neurocomputing 56, 481–487 (2004)

10. Ma, J., Wang, L.: BYY Harmony Learning on Finite Mixture: Adaptive Gradient Implementation and a Floating RPCL Mechanism. Neural Processing Letters 24(1), 19–40 (2006)
11. Ma, J., Liu, J.: The BYY Annealing Learning Algorithm for Gaussian Mixture with Automated Model Selection. Pattern Recognition 40, 2029–2037 (2007)
12. Ma, J., He, X.: A Fast Fixed-point BYY Harmony Learning Algorithm on Gaussian Mixture with Automated Model Selection. Pattern Recognition Letters 29(6), 701–711 (2008)
13. Zhang, B., Zhang, C., Yi, X.: Competitive EM Algorithm for Finite Mixture Models. Pattern Recognition 37, 131–144 (2004)
14. Zhang, Z., Chen, C., Sun, J., et al.: EM Algorithms for Gaussian Mixtures with Split-and-Merge Operation. Pattern Recogniton 36, 1973–1983 (2003)