Applied Mathematics and Computation 205 (2008) 832–840

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

A BYY scale-incremental EM algorithm for Gaussian mixture learning

Lei Li, Jinwen Ma *

Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China

ARTICLE INFO

Keywords:
Bayesian Ying–Yang (BYY) harmony learning
Gaussian mixture
EM algorithm
Model selection
Unsupervised image segmentation

ABSTRACT

Gaussian mixture model has been used extensively in the fields of information processing and data analysis. However, its model selection, i.e., the selection of number of components or Gaussians in the mixture, is still a difficult problem. Fortunately, the new established Bayesian Ying–Yang (BYY) harmony function provides an efficient criterion for the model selection of Gaussian mixture with a set of sample data. In this paper, we propose a BYY scale-incremental EM algorithm for Gaussian mixture learning via a component split rule to increase the BYY harmony function incrementally. Particularly, starting from two components and adding one component sequentially via the split rule after each EM procedure until a maximum number of components, the algorithm increases the scale of the mixture incrementally and leads to the maximization of the BYY harmony function, together with the correct model selection and a good parameter estimation of the Gaussian mixture. It is demonstrated well by the simulation experiments that this BYY scale-incremental EM algorithm can make both model selection and parameter estimation efficiently for Gaussian mixture modeling. Moreover, the BYY scale-incremental EM algorithm is successfully applied to two real-life data sets, including Iris data classification and unsupervised color image segmentation.

1.

As a typical statistical model, Gaussian mixture has been widely used in the fields of information processing and data analysis. In fact, there have been several statistical methods for its learning or modeling (e.g., the expectation–maximization (EM) algorithm [1] for maximum likelihood and the self-organizing network with hyper-ellipsoidal clustering [2]). Generally, the parameters of Gaussian mixture can be estimated via the EM algorithm under the maximum likelihood framework. Although the EM algorithm owns certain good convergence behaviors in certain situations (e.g., [3–7]), it generally has some weaknesses or limitations. Clearly, the EM algorithm is a local searching approach, thus "bad" initialization can make it get trapped in a local maxima. Moreover, it is based on the assumption that the number of Gaussians in the mixture is pre-known and fixed, otherwise it cannot work. However, in many instances, this crucial information is not available and the selection of an appropriate number of Gaussians must be made with the estimation of the parameters, which becomes a rather complicated problem [8,9]. As the number of Gaussians is just a scale of the Gaussian mixture, the selection of number of Gaussians in the mixture is generally referred to as the model selection for the mixture. Thus, as the number of Gaussians is not known in advance, the Gaussian mixture learning is a compound problem of model selection and parameter estimation.

The traditional approach to solving this compound problem is to choose a best number k^* of Gaussians via some selection criterion. As a matter of fact, there have been many existing selection criteria, and among them, Akaike's information

* Corresponding author.
E-mail address: jwma@math.pku.edu.cn (J. Ma).

0096-3003/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2008.05.076

criterion (AIC) [10] as well as its extensions (e.g., Bozdogan's information criteria [11]), Bayesian inference criterion (BIC) [12], minimum description length (MDL) criterion [13] and minimum message length (MML) criterion [14], are well-known. However, all the existing theoretic selection criteria have their limitations and often result in a wrong result. Moreover, the process of evaluating an information criterion or validity index incurs a large computational cost since we need to repeat the entire parameter estimation process at a large number of different values of k.

During 1990s, there appeared some new approaches to solving this problem. One approach was to utilize a kind of stochastic simulation to infer the optimal mixture model. The two typical implementations are the methods of Dirichlet processes [15] and reversible jump Markov chain Monte Carlo (RJMCMC) [16]. However, these stochastic simulation methods generally require a large number of samples via different sampling rules. Another approach was to implement the Bayesian inference by maximizing the variational function, which was known as the variational Bayesian learning [17,18]. But this VB method only maximizes a lower bound of the Bayesian inference probability and is still in lack of theoretical justifications.

Recently, a novel approach has been developed from the Bayesian Ying–Yang (BYY) harmony learning system and theory [10–22] with a feature that model selection can be made automatically during the parameter learning. Actually, it was already shown in [23] that the Gaussian mixture modeling problem in which the number of Gaussians is unknown can be equivalent to the maximization of a harmony function on a specific BI-directional architecture (BI-architecture) of the BYY system for the Gaussian mixture model and a gradient learning rule for maximization of this harmony function was also established. Later on, the conjugate, natural, adaptive gradient and fixed-point learning algorithms [24–26] were further proposed to improve the efficiency of the harmony function maximization. These BYY learning algorithms have the same behavior that an appropriate number of Gaussians can be automatically allocated for the sample data set, with the mixing proportions of the extra Gaussians attenuating to zero. That is, they can learn the parameters of the Gaussian mixture with automated model selection. In fact, this BYY harmony function and its model selection property were theoretically analyzed and proved under some wild conditions in [27]. Moreover, an annealing BYY learning algorithm [28] was also established on a backward architecture of the BYY system for the Gaussian mixture to search the global maximum of the harmony function, being expressed as a kind of deterministic annealing EM procedure. On the other hand, from point view of penalizing the Shannon entropy of the mixing proportions on maximum likelihood estimation (MLE), an entropy penalized MLE iterative algorithm was also proposed to make model selection automatically with parameter estimation on Gaussian mixture [29].

Although those automated model selection learning algorithms are quite efficient for Gaussian mixture learning in many situations, they must satisfy an assumption that k is larger than the number of actual Gaussians in the sample data. Clearly, we can easily overestimate the number of Gaussians in the sample data and set it to be k. But when k is much larger than the true value, these algorithms usually converge to a wrong result. Unfortunately, it is rather difficult to get an overestimate of the number of actual Gaussians in the sample data which is just slightly larger than the true number. In order to get rid of this difficulty, we can construct a scale-incremental learning algorithm by increasing the number of components one by one until it reaches the correct one with the best harmony, i.e., the maximization of the harmony function. In fact, Vlassis and Likas have already proposed such a kind of algorithm called the greedy EM algorithm [30], which was further discussed and strengthened in [31]. However, the stop criterion of this greedy EM algorithm is still based on the maximum likelihood and thus cannot guarantee the correctness of the final model selection, i.e., the maximum k.

In the current paper, we propose a BYY scale-incremental EM algorithm in which each component split operation tries to increase the harmony function and the stop criterion is based on the maximum harmony function. Since the maximization of the harmony function just corresponds to the correct model selection [27], the BYY scale-incremental EM algorithm leads to the correct model selection. Actually, it is demonstrated well by the simulation and practical experiments that this BYY scale-incremental EM algorithm is efficient for Gaussian mixture learning and its applications.

The rest of the paper is organized as follows. In Section 2, we revisit the EM algorithm for Gaussian mixtures. We further introduce the BYY learning system and the harmony function in Section 3. In Section 4, we present the BYY scale-incremental EM algorithm. Several simulation experiments and applications to classification of the Iris data and unsupervised color image segmentation are conducted in Section 5 to demonstrate the efficiency of the proposed BYY scale-incremental EM algorithm. Finally, we conclude briefly in Section 6.

2.

2.1. Gaussian mixture model

We begin with a brief description of Gaussian mixture model. Let X be a d-dimensional random variable, with $x = [x_1, x_2, \ldots, x_d]^T$ representing a particular value of X. Mathematically, X is called to be subject to a finite mixture model of k components in $\Re^d$ if its probability density is given as follows:

$$\Phi(x) = \sum_{i=1}^{k} \pi_i \phi(x|\theta_i) \quad \forall x \in \Re^d, \tag{1}$$

where each θ_i is the set of parameters defining the ith component (i.e., the probability density), and $\pi_i \in (0,1)$ $(i = 1, 2, \ldots, k)$ are the mixing proportions subject to $\sum_{i=1}^{k} \pi_i = 1$. Then, for the Gaussian mixture model, each component density $\phi(x|\theta_i)$ is a Gaussian probability density given by

$$\phi(x|\theta_i) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_j|} e^{-\frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)}, \tag{2}$$

m_j is the mean vector and Σ_j is the covariance matrix which is assumed positive definite. For clarity, we encapsulate all the parameters into one vector $\Theta = (\pi_1, \pi_2, \ldots \pi_k, \theta_1, \theta_2, \ldots, \theta_k)$, where $\theta_i = (\mu_i, \Sigma_i)$ represents the parameters of the ith Gaussian. In this way, according to Eq. (1), the density of the Gaussian mixture can be rewritten as

$$\Phi(x|\Theta) = \sum_{i=1}^{k} \pi_i \phi(x|\theta_i) = \sum_{i=1}^{k} \pi_i \phi(x|\mu_i, \Sigma_i). \tag{3}$$

For the Gaussian mixture learning or modeling, we usually only have a sample data set $S = \{x_1, x_2, \ldots x_N\}$ from the original Gaussian mixture and our aim is to estimate all the parameters in the original Gaussian mixture from these sample data. The difficulty relies on the blindness of the index of the Gaussian from which each sample was generated. In fact, if we know the Gaussian from which each sample x_t comes, the estimation of the parameters θ_k for each Gaussian becomes very simple. Oppositely, as these samples are not labeled, we need to estimate these incomplete data, which makes the estimation of the parameters in Θ_k more difficult. Fortunately, the EM algorithm for Gaussian mixtures can solve this difficulty with the help of the concept of the missing data and the expectation.

2.2. The EM algorithm for Gaussian mixtures

The well-known expectation–maximization (EM) algorithm [1] is designed to solve the maximum likelihood estimation problem for a probability model in which some random variable can be observed, while the other random variable cannot be observed. That is, it concerns the problem of missing or unobservable data. By alternatively implementing the expectation step to estimate the probability distribution of the unobservable random variable and the maximization step to increase the log-likelihood function of the model, the EM algorithm can finally lead to a local maximum of the log-likelihood function of the model on the sample data set. For the Gaussian mixture model we consider, the available sample data $S = \{x_1, x_2, \ldots x_N\}$ can be considered as the observable data, while the hidden indexes of these samples can be considered as the unobservable data. In this situation, the log-likelihood function can be expressed as follows:

$$\log p(S|\Theta_k) = \log \prod_{t=1}^{N} \phi(x_t|\Theta_k) = \sum_{t=1}^{N} \log \sum_{i=1}^{k} \pi_i \phi(x_t|\theta_i). \tag{4}$$

By applying the EM algorithm to the maximum likelihood estimation problem of the Gaussian mixture with the sample data set $S = \{x_1, x_2, \ldots x_N\}$, we can easily establish the EM algorithm for Gaussian mixtures, which enables us to update the parameters of the Gaussian mixture with the sample data such that the above log-likelihood function increases incrementally to a local maximum. Actually, the update of the parameters of the EM algorithm for Gaussian mixtures can be given by the following iterative equations for all k components [3]:

$$P(j|x_t) = \frac{\pi_j \phi(x_t|\theta_j)}{\sum_{i=1}^{k} \pi_i \phi(x_t|\theta_i)}, \tag{5}$$

$$\pi_j^+ = \frac{1}{N} \sum_{t=1}^{N} P(j|x_t), \tag{6}$$

$$\mu_j^+ = \frac{1}{\sum_{i=1}^{N} P(j|x_t)} \sum_{t=1}^{N} P(j|x_t) x_t, \tag{7}$$

$$\Sigma_j^+ = \frac{\sum_{t=1}^{N} P(j|x_t)(x_t - \mu_j^+)(x_t - \mu_j^+)^T}{\sum_{t=1}^{N} P(j|x_t)}. \tag{8}$$

As already pointed out in the previous section, the EM algorithm cannot guarantee to converge to the best solution, i.e., the consistent maximum likelihood estimate with the sample data set S. Generally, it is considered as a linearly convergent algorithm [3]. However, recent theoretical analysis has proved that the EM algorithm for Gaussian mixtures or the mixtures of densities from a class of exponential families tends to be asymptotically super-linear when the overlap of densities in the mixture tends to zero [4–6]. Moreover, it was also proved that the EM algorithm for Gaussian mixtures tends to converge to the correct solution as the overlap of densities in the mixture tends to zero [7]. Thus, the EM algorithm is better than the gradient-type algorithms. Nevertheless, as the EM algorithm tries to maximize the likelihood function, it certainly has no ability to make model selection for the Gaussian mixture with the sample data. In order to overcome these weaknesses, we will utilize the BYY harmony function instead of the log-likelihood function in our scale-incremental approach to the Gaussian mixture learning.

3.

We further introduce the Bayesian Ying–Yang (BYY) learning system and the harmony function on Gaussian mixture, which will be used to construct our scale-incremental EM algorithm. A BYY system describes each observation $x \in \subset \Re^d$ and its corresponding inner representation $y \in \subset \Re^m$ via the two types of Bayesian decomposition of the joint density $p(x,y) = p(x)p(y|x)$ and $q(x,y) = q(y)q(x|y)$, being called Yang and Ying machines, respectively. For analysis of the finite mixture, y is limited to be an integer variable, i.e., $y \in = \{1,2,\ldots,k\} \subset \Re$ with $m = 1$. Given a data set $D_x = \{x_t\}_{t=1}^N$, the task of learning on a BYY system consists of specifying all the aspects of $p(x), p(y|x), q(y), q(x|y)$ with a harmony learning principle implemented by maximizing the harmony functional

$$H(p||q) = \int p(y|x)p(x)\ln[q(x|y)q(y)]\,\mathrm{d}x\,\mathrm{d}y - \ln z_q, \tag{9}$$

where z_q is a regularization term [21].

If both $p(y|x)$ and $q(x|y)$ are parametric, i.e, from a family of probability densities with a parameter $\theta \in R^d$, the BYY system is called to have a Bi-directional architecture (BI-architecture). For the Gaussian mixture modeling, we use the following specific BI-architecture of the BYY system. $q(j) = \alpha_j$, $\alpha_j \geqslant 0$ and $\sum_{j=1}^k \alpha_j = 1$. Also, we ignore the regularization term z_q (i.e, set $z_q = 1$) and let $p(x)$ be the empirical density $p_0(x) = \frac{1}{N}\sum_{t=1}^N g(x - x_t)$, where $x \in = \Re^d$ and $g(\cdot)$ is a kind of kernel function (e.g., Gaussian function). Moreover, the BI-architecture is constructed with the following parametric from:

$$p(y = j|x) = \frac{\alpha_j q(x|\theta_j)}{q(x|\Theta_k)}, \quad q(x|\Theta_k) = \sum_{j=1}^k \alpha_j q(x|\theta_j), \tag{10}$$

where $q(x|\theta_j) = q(x|y = j)$ with θ_j consisting of all its parameters and $\Theta_k = \{\alpha_j, \theta_j\}_{j=1}^k$. Substituting these component densities into Eq. (9) and letting the kernel functions approach the delta functions $\delta(\cdot)$, we have

$$H(p||q) = J(\Theta_k) = \frac{1}{N}\sum_{t=1}^N \sum_{j=1}^k \frac{\alpha_j q(x_t|\theta_j)}{\sum_{i=1}^k \alpha_i q(x_t|\theta_i)} \ln[\alpha_j q(x_t|\theta_j)]. \tag{11}$$

This is, $H(p||q)$ becomes a harmony function $J(\Theta_k)$ on the parameters Θ_k of a finite mixture model. When $q(x|\theta_j)$ is a Gaussian mixture density given by Eq. (2), $J(\Theta_k)$ becomes a harmony function on Gaussian mixtures with the sample data set D_x. It has been demonstrated by the experiments in [23–26,28] and proved by the theoretical analysis in [27] that this harmony function reaches its global maximization when the number of Gaussians is just equal to that of the actual Gaussians or clusters in the sample data. Thus, we will use it as a new criterion for the model selection of the Gaussian mixture in our scale-incremental approach for the Gaussian mixture learning.

4. -

With the above preparations, we now begin to present our BYY scale-incremental EM algorithm. Given a sample data set $S = \{x_1, x_2, \ldots, x_N\}$ from an original mixture with $k^*(> 1)$ Gaussians and setting an initial number $k = 2$, we can use the (conventional) EM algorithm to get k estimated Gaussians with the associated parameters. When $k < k^*$, there are some estimated Gaussians, which cannot match the actual Gaussian and should be split into two or more Gaussians. Thus, the main task of the scale-incremental algorithm is to construct a split criterion so that the split operation can be combined with the EM algorithm dynamically and independently. Due to the BYY harmony function, we can construct the BYY harmony split criterion as well as the scale-incremental EM algorithm in the following two subsections.

4.1. BYY harmony split criterion

After each EM procedure with a fixed k, we get the estimated parameters Θ_k in the Gaussian mixture. According to Eq. (9), the harmony function $J(\Theta_k)$ can be further expressed in the sum form as follows:

$$J(\Theta_k) = \sum_{j=1}^k H_j(p_j||q_j), \tag{12}$$

where

$$H(p_j||q_j) = \frac{1}{N}\sum_{t=1}^N \frac{\alpha_j q(x_t|\theta_j)}{\sum_{i=1}^k \alpha_i q(x_t|\theta_i)} \ln[\alpha_j q(x_t|\theta_j)]. \tag{13}$$

Clearly, $H_j(p_j||q_j)$ denotes the harmony level of the jth Gaussian with respect to the corresponding actual Gaussian implied in the sample data. In order to improve the total harmony function, we can split the component or Gaussian with the least component harmony value $H_j(p_j||q_j)$. That is, if $H_r(p_r||q_r)$

operation on the rth component. Specifically, we divide it into two components i', j' with their parameters being designed as follows (refer to [32]).

According to the covariance matrix Σ_r, we compute its singular value decomposition $\Sigma_r = USV^{\mathrm{T}}$, where $S = \mathrm{diag}[s_1, s_2, \ldots, s_d]$ is a diagonal matrix with nonnegative diagonal elements in a descent order, U and V are two (standard) orthogonal matrices. Then, we further set $A = U\sqrt{S} = U\mathrm{diag}[\sqrt{s_1}, \sqrt{s_2}, \ldots, \sqrt{s_d}]$ and get the first column A_1 of A. Finally, we have the parameters for the two split Gaussians as follows:

$$\alpha_{i'} = \gamma\alpha_r, \alpha_{j'} = (1-\gamma)\alpha_r, \tag{14}$$

$$m_{i'} = m_r - (\alpha_{j'}/\alpha_{i'})^{1/2}\mu A_1, \tag{15}$$

$$m_{j'} = m_r + (\alpha_{i'}/\alpha_{j'})^{1/2}\mu A_1, \tag{16}$$

$$\Sigma_{i'} = (\alpha_{j'}/\alpha_{i'})\Sigma_r + ((\beta - \beta\mu^2 - 1)(\alpha_r/\alpha_{i'}) + 1)A_1A_1^{\mathrm{T}}, \tag{17}$$

$$\Sigma_{j'} = (\alpha_{i'}/\alpha_{j'})\Sigma_r + ((\beta\mu^2 - \beta - \mu^2)(\alpha_r/\alpha_{j'}) + 1)A_1A_1^{\mathrm{T}}, \tag{18}$$

where γ, μ, β are all equal to 0.5.

4.2. Procedure of BYY scale-incremental EM algorithm

According to the above BYY harmony split criterion, the procedure of the BYY scale-incremental EM algorithm can be summarized as follows:

1. Set $k = 2$ and the initial values of the parameters Θ_2.
2. At each k and with the parameters Θ_k, split the least harmony Gaussian $\phi(x|\theta_r)$ into two new Gaussians $\phi(x|\theta'_j)$ and $\phi(x|\theta''_j)$ according to Eqs. (14)–(18).
3. Perform the EM algorithm from the parameters of the remainder and split Gaussians and their mixing proportions to update the parameters Θ_{k+1} for the mixture of $k+1$ Gaussians.
4. If $J(\Theta_{k+1}) \leqslant J(\Theta_k)$, stop and get the result Θ_k at the Gaussian number k, otherwise, let $k = k+1$ with the parameters Θ_{k+1} and return to Step 2.

It can easily found from the above procedure that the split operation tries to increase the total harmony function and the stopping criterion tries to prevent from splitting too many Gaussians. Therefore, the BYY scale-incremental EM algorithm can find correct number of Gaussians in the sample data.

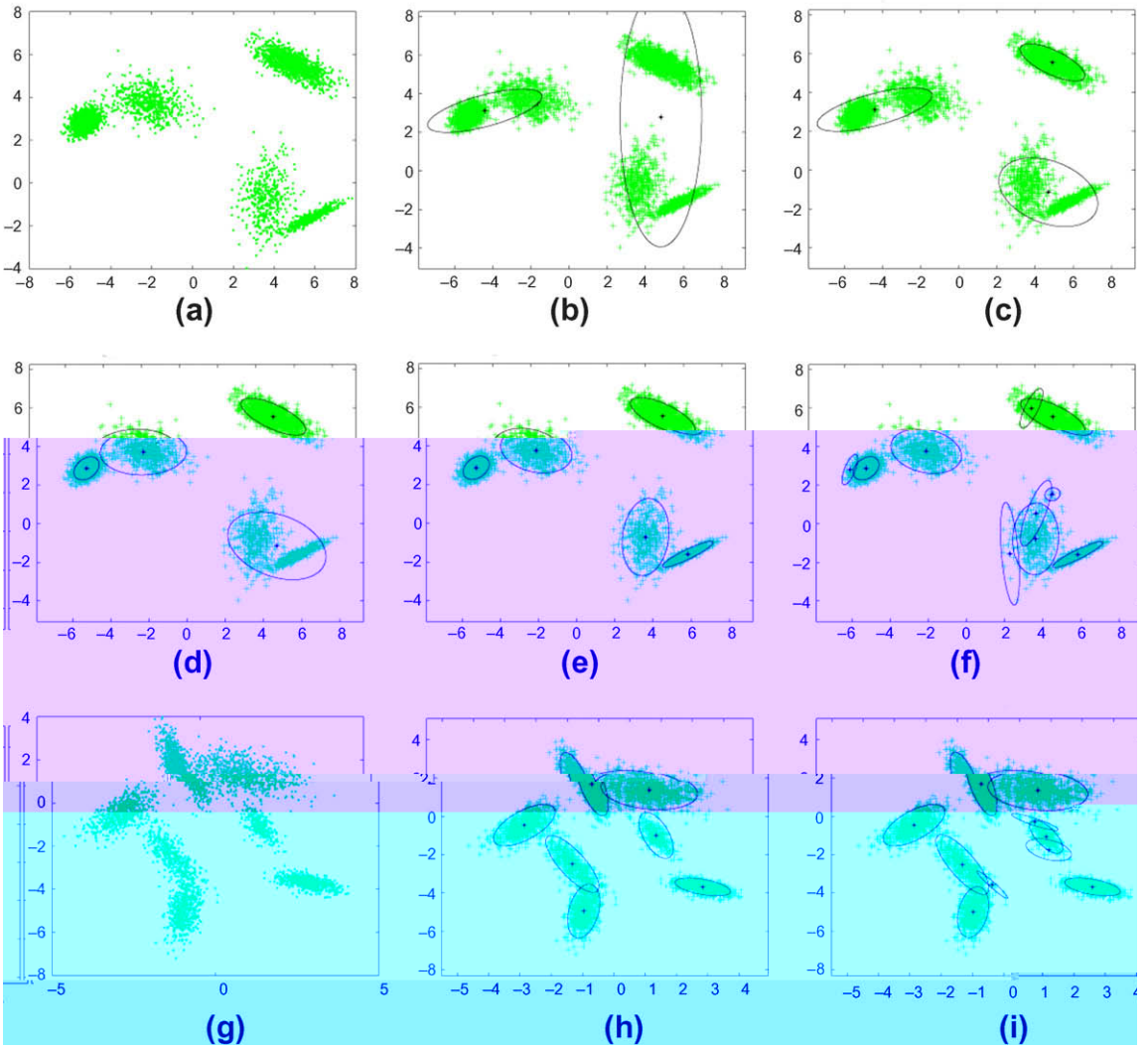
5.

In this section, several simulation experiments are carried out to demonstrate the BYY scale-incremental EM algorithm for Gaussian mixture learning on two data sets. Moreover, the BYY scale-incremental EM algorithm is applied to classification of the Iris data and unsupervised color image segmentation.

5.1. Simulation experiments

We conducted simulation experiments on two sets of samples drawn from a mixture of five or seven bivariate Gaussians densities (i.e., $d = 2$). As shown in Fig. 1a, the first data set consists of five Gaussians with certain degree of overlap. At $k = 2$, the initial values, shown in Fig. 1b, were set by conducting a procedure of k-means algorithm. We implemented the BYY scale-incremental EM algorithm on the data set from $k = 2$ and the algorithm always stopped as long as $J(\Theta_{k+1}) \leqslant J(\Theta_k)$.

The experimental results of the BYY scale-incremental EM algorithm on this data set are given in Fig. 1c–e at the three different steps. It can be clearly observed that after the scale-incremental learning, five Gaussians were finally located accu-



. 1. (a) The first set of sample data with five Gaussians; (b)–(e) the experimental results at the three steps of the BYY scale-incremental EM algorithm on the first sample data set; (f) the experimental result of the greedy EM algorithm on the first sample data set; (g) the second set of sample data with seven Gaussians; (h) the experimental result of the BYY scale-incremental EM algorithm on the second sample data set and (i) the experimental result of the greedy EM algorithm data set.

1
The cost times of the two algorithms on the both data sets

The data sets	The number of actual Gaussians or classes	The BYY scale-incremental EM algorithm	The Greedy EM algorithm
Set 1	5	5.891251 s	7.017382 s
Set 2	7	7.559194 s	8.012473 s

5.2. Classification of the Iris data

We further applied the BYY scale-incremental EM learning algorithm to the classification of the Iris data, which is a typical real data set for testing a classification algorithm. Actually, it consists of 150 samples of three classes where each class contains 50 samples and each sample or datum is four-dimensional and consists of measures of the plant morphology. Since our BYY scale-incremental EM algorithm is a kind of unsupervised learning algorithm, we would not use the class indexes of these samples. However, these pre-known class indexes would be used to check the classification accuracy of the BYY incremental EM learning algorithm on the Iris data.

2

The component splitting numbers of the two algorithms on the both data sets

The data sets	The number of actual Gaussians or classes	The BYY scale-incremental EM algorithm	The Greedy EM algorithm
Set 1	5	5	10
Set 2	7	7	9

. 2. (a) The experimental results on the color image segmentation; (b) The original color images; (c) the segmentation results of the BYY scale-incremental EM algorithm and the segmentation results of the greedy EM algorithm.

We started the BYY scale-incremental EM algorithm by setting $k = 2$ and the other initial parameters were obtained from the convergent result of the k-means algorithm on the Iris data. The algorithm was also stopped when $J(\Theta_{k+1}) \leqslant J(\Theta_k)$. For quick convergence of the algorithm, we set a low threshold value $T = 0.033$. When the mixing proportion of some Gaussian was less than T, we cancel this Gaussian in the mixture for the following learning iterations. It was shown by the experiments that the BYY scale-incremental EM learning algorithm could detect the three classes in the Iris data with an optimal classification accuracy of 97.4% (there are only four errors in the second class), which is considerably better than the classification accuracy 93.3% (there are ten errors) of the Greedy EM method [31], but slightly less than the optimal classification accuracy 98% (there are only three errors) of the maximum certainty partitioning method with a large number of linear mixing Gaussian kernels [33].

5.3. Applications to color image segmentation

We finally applied the BYY scale-incremental EM algorithm to the unsupervised color image segmentation which has been recognized as a promising and challenging topic in image processing [34]. In our experiments, we first transformed the original color images, as shown in the column (a) of Fig. 2, from the RGB coordinate into the YUV coordinate in the same way as did in [34]. Each pixel in the color image was then represented by a three dimensional real vector. The experimental results of the BYY scale-incremental EM algorithm on these color images are given in the column (b) of Fig. 2. For comparison, the experimental results of the greedy EM algorithm [31] are also given in the column (c) of Fig. 2. From these segmented images of the BYY scale-incremental EM algorithm, we can find that two or three objects (including the background) can be located accurately at the actual objects, respectively. Moreover, in comparison with the segmented images of the greedy EM algorithm, we can find that our proposed scale-incremental EM algorithm can get a more accurate segmentation on the contours of the objects in each image.

6.

We have investigated the Gaussian mixture learning for both the parameter estimation and model selection from the pointview of component splitting in the EM algorithm and established a scale-incremental learning algorithm with the help of the Bayesian Ying–Yang (BYY) harmony function and the EM algorithm. This BYY scale-incremental EM algorithm begins with two components and adds one component at each step via the BYY harmony split rule after each EM procedure until a maximum number k is reached at the maximization of the BYY harmony function. It is demonstrated well by the simulation experiments and the practical applications that the BYY scale-incremental EM algorithm achieves a good parameter estimation of the Gaussian mixture with correct model selection on a sample data set, and is always better than the greedy EM algorithm.

This work was supported by the Natural Science Foundation of China for grants 60471054 and 60771061.

[1] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39 (1977) 1–38.
[2] J. Mao, A.K. Jain, A self-organizing network for hyperellipsoidal clustering, IEEE Transactions on Neural Networks 7 (1) (1996) 16–29.
[3] R.A. Render, H.F. Walker, Mixture densities, maximum likelihood and the EM algorithm, SIAM Review 26 (2) (1984) 195–239.
[4] L. Xu, M.I. Jordan, On convergence properties of the EM algorithm for Gaussian mixtures, Neural Computation 8 (1996) 129–151.
[5] J. Ma, L. Xu, M.I. Jordan, Asymptotic convergence rate of the EM algorithm for Gaussian mixtures, Neural Computation 12 (12) (2000) 2881–2907.
[6] J. Ma, L. Xu, Asymptotic convergence properties of the EM algorithm with respect to the overlap in the mixture, Neurocomputing 68 (2005) 105–129.
[7] J. Ma, S. Fu, On the correct convergence of the EM algorithm for Gaussian mixtures, Pattern Recognition l38 (12) (2005) 2602–2611.
[8] J.A. Hartigan, Distribution problems in clustering, in: J. Van Ryzin (Ed.), Classification and Clustering, Academic Press, 1977, pp. 45–72.
[9] G.W. Millgan, M.C. Copper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 46 (1985) 187–199.
[10] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control AC-19 (1974) 716–723.
[11] H. Bozdogan, Model selection and Akaike's information criterion: the general theory and its analytical extensions, Psychometrika 52 (1987) 345–370.
[12] G. Scharz, Estimating the dimension of a model, The Annals of Statistics 6 (2) (1978) 461–464.
[13] J. Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific, Singapore, 1989.
[14] J. Oliver, R. Baxterand, C. Wallace, Unsupervised learning using MML, in: Proceedings of the 13th International Conference on Machine Learning, 1996, pp. 364–372.
[15] M.D. Escobar, M. West, Bayesian density estimation and inference using mixtures, Journal of the American Statistical Association 90 (430) (1995) 577–588.
[16] J.G. Peter, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (4) (1995) 711–732.
[17] S.J. Roberts et al., Bayesian approaches to Gaussian mixture modelling, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1133–1142.
[18] N. Ueda, Z. Ghahramani, Bayesian model search for mixture models based on optimizing variational bounds, Neural Networks 15 (2002) 1223–1241.
[19] L. Xu, Ying–Yang machine: a Bayesian–Kullback scheme for unified learnings and new results on vector quantization, in: Proceedings of the 1995 International Conference on Neural Information Processing (ICONIP'95), vol. 2, 1995, pp. 977–988.
[20] L. Xu, Bayesian Ying–Yang machine, clustering and number of clusters, Pattern Recognition Letters 18 (1997) 1167–1178.
[21] L. Xu, 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) (2001) 43–69.

[22] L. Xu, BYY harmony learning, structural RPCL and topological self-organizing on mixture modes, Neural Networks 15 (8–9) (2002) 1231–1237.
[23] J. Ma, T. Wang, L. Xu, A gradient BYY harmony learning rule on Gaussian mixture with automated model selection, Neurocomputing 56 (2004) 481–487.
[24] J. Ma et al., Conjugate and natural gradient rules for BYY harmony learning on Gaussian mixture with automated model selection, International Journal of Pattern Recognition and Artificial Intelligence 19 (5) (2005) 701–713.
[25] J. Ma, L. Wang, BYY harmony learning on finite mixture: adaptive gradient implementation and a floating RPCL mechanism, Neural Processing Letters 24 (1) (2006) 19–40.
[26] J. Ma, X. He, A fast fixed-point BYY harmony learning algorithm on Gaussian mixture with automated model selection, Pattern Recognition letters 29 (2008) 701–711.
[27] J. Ma, Automated model selection (AMS) on finite mixtures: a theoretical analysis, in: Proceedings of the 2006 International Joint Conference on Neural Networks (IJCNN06), 2006, pp. 8255–8261.
[28] J. Ma, J. Liu, The BYY annealing learning algorithm for Gaussian mixture with automated model selection, Pattern Recognition 40 (2007) 2029–2037.
[29] J. Ma, T. Wang, Entropy penalized automated model selection on Gaussian mixture, International Journal of Pattern Recognition and Artificial Intelligence 18 (8) (2004) 1501–1512.
[30] N. Vlassis, A. Likas, A greedy EM algorithm for Gaussian mixture learning, Neural Processing Letters 15 (2002) 77–87.
[31] J.J. Verbeek, N. Vlassis, B. Krose, Efficient greedy learning of Gaussian mixture models, Neural Computation 15 (2) (2003) 469–485.
[32] Z. Zhang et al., EM algorithms for Gaussian mixtures with split-and-merge operation, Pattern Recognition 36 (2003) 1973–1983.
[33] S.J. Robert, R. Everson, I. Rezek, Maximum certainty data partitioning, Pattern Recognition 33 (2000) 833–839.
[34] N. Boujeman, Generalized competitive clustering for image segmentation, in: Proceedings of the 19th International Conference of the North American Fuzzy Information Processing Society, 2000, pp. 133–137.