Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 2029–2037
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

# The BYY annealing learning algorithm for Gaussian mixture with automated model selection

Jinwen Ma*, Jianfeng Liu

*D I S , S M S LMAM, P U , B 100871, C*

Received 8 December 2005; received in revised form 19 December 2006; accepted 20 December 2006

**Abstract**

Bayesian Ying–Yang (BYY) learning has provided a new mechanism that makes parameter learning with automated model selection via maximizing a harmony function on a backward architecture of the BYY system for the Gaussian mixture. However, since there are a large number of local maxima for the harmony function, any local searching algorithm, such as the hard-cut EM algorithm, does not work well. In order to overcome this difficulty, we propose a simulated annealing learning algorithm to search the global maximum of the harmony function, being expressed as a kind of deterministic annealing EM procedure. It is demonstrated by the simulation experiments that this BYY annealing learning algorithm can efficiently and automatically determine the number of clusters or Gaussians during the learning process. Moreover, the BYY annealing learning algorithm is successfully applied to two real-life data sets, including Iris data classification and unsupervised color image segmentation.


*K* : Bayesian Ying–Yang (BYY) learning; Gaussian mixture; Automated model selection; Simulated annealing; Unsupervised image segmentation

## 1. Introduction

As a powerful tool for data clustering or partitioning, Gaussian mixture model has been extensively studied in the literature for either parameter estimation or learning with a sample data set. Although there are several statistical methods to do such a task, e.g., the -means algorithm [1] and the EM algorithm [2], it is usually assumed that the number of clusters or Gaussians is pre-known. However, in many cases this key information is not available and then the appropriate number of Gaussians must be selected along with the estimation of parameters in the mixture, which becomes a rather complicated and difficult task [3].

As the number of Gaussians is just a scale of the Gaussian mixture model, the selection of number of Gaussians in the Gaussian mixture modeling is generally referred to as model selection. Thus, the general Gaussian mixture modeling is actually a compound problem of parameter learning (namely, estimation and model selection). In fact, this compound problem has been investigated by many researchers from different directions. The traditional approach was to choose an optimal number of Gaussians in the mixture via certain selection criterion. There have been several statistical selection criteria for this purpose. Among them, Akaike's information criterion (AIC) [4] as well as its extensions, such as the consistent Akaike's information criteria (CAIC) [5], are well known. However, all the existing statistical selection criteria have their limitations and often lead to a wrong result. Moreover, the process of evaluating a criterion incurs a large computational cost since we need to repeat the entire parameter learning process as different numbers of mixtures are estimated.

Since 1990s, there have appeared some new approaches to solve this problem. One approach was to use a kind of stochastic simulation to infer the optimal mixture model. The two typical implementations are the methods of Dirichlet processes [6] and reversible jump Markov chain Monte Carlo (RJMCMC) [7]. However, these stochastic simulation methods generally require a large number of samples through different sampling rules. Another approach was to introduce a kind of deterministic annealing in the partition learning process [8]. Certainly,

* Corresponding author. Tel.: +86 10 62760609; fax: +86 10 62751801.
*E-* : jwma@math.pku.edu.cn (J. Ma).

0031-3203/$30.00 
doi:10.1016/j.patcog.2006.12.028

there were several deterministic annealing methods that try to overcome the local convergence problem of the cost-based parameter estimation with the number of Gaussians or clusters fixed and given in advance (i.e., without the model selection procedure) (e.g., Refs. [9–11]). Recently, this compound problem was also investigated using maximum certainty partitioning [12] and variational Bayesian learning [13].

Alternatively, the Bayesian Ying–Yang (BYY) harmony learning system and theory, which was first proposed in 1995 [14] and then systematically developed and summarized in Refs. [15–17], has also established a theoretical foundation to solve this compound problem. The BYY harmony learning acts as a general statistical learning framework, which is not only useful for understanding several existing major learning approaches but also for tackling the learning problem on a set of finite samples with a new learning mechanism that makes model selection automatically during parameter learning. Particularly, for solving the current problem of interest, we can implement a mechanism of parameter learning with automated model selection on a certain BYY system for the Gaussian mixture via maximizing a harmony function, which is reduced from the harmony measure between the Ying and Yang machines. In fact, in Refs. [18–20], this mechanism was already implemented on a bidirectional architecture (BI-architecture) of the BYY system via some gradient-type learning algorithms in order to solve this compound problem of parameter learning and model selection. Moreover, a backward architecture (B-architecture) of the BYY system can also be applied to solving it. Actually, by simply ignoring the regularization term in the harmony function on this BYY system, a direct maximization of the harmony function leads to a discrete optimal problem with a hard-cut EM algorithm [15], which suffers from the difficulty of being stuck at a local maximum solution.

In the current paper, we follow the simulated annealing idea of gradually shifting the maximum likelihood (or equivalently the Kullback divergence learning) to the harmony learning suggested in [16,17] and present a simulated annealing procedure for searching the global maximum of the harmony function on the B-architecture of the BYY system, such that model selection and parameter learning on the Gaussian mixture can be accomplished simultaneously and efficiently. Namely, a BYY annealing learning algorithm is constructed for parameter learning on the Gaussian mixture with automated model selection. It is demonstrated by experiments that the BYY annealing learning algorithm can always perform model selection automatically during the parameter learning and leads to a good clustering or partitioning result.

In the sequel, the BYY harmony learning system and architecture are briefly introduced, and the BYY annealing learning algorithm is derived in Section 2. Several simulation and practical experiments are conducted in Section 3 to demonstrate the efficiency of the proposed annealing learning algorithm. Finally, we conclude the paper in Section 4.

## 2. The BYY annealing learning algorithm

A BYY system describes each observation $x \in \mathcal{X} \subset R^n$ and its corresponding inner representation $y \in \mathcal{Y} \subset R^m$ via two types of Bayesian decomposition of the joint probability density functions: $p(x, y) = p(x)p(y|x)$ and $q(x, y) = q(x|y)q(y)$, which are called $Y$ and $Y$, respectively. Given a data set $D_x = \{x_t\}_{t=1}^N$, the goal of on a BYY system is to extract the hidden probabilistic structure of with the help of from specifying all the aspects of $p(y|x)$, $p(x)$, $q(x|y)$, $q(y)$ via a harmony learning principle implemented by maximizing the following functional:

$$H(p||q) = \int p(y|x)p(x)\, \ln[q(x|y)q(y)]\, \mathrm{d}x\, \mathrm{d}y - \ln z_q, \quad (1)$$

where $z_q$ is a regularization term. That is, the harmony learning principle attempts to minimize the difference between $p(x, y)$ and $q(x, y)$, plus certain regularization. The theoretical details are referred in [16,17].

The BYY system is called to have a B-architecture (short for Backward architecture) if $q(x|y)$ is parametric, i.e., from a family of probability densities with a parameter $\theta$, while $p(y|x)$ is free to be determined by learning. For the Gaussian mixture modeling, we can use the following B-architecture of the BYY system. The inner representation is discrete, i.e., $y \in \mathcal{Y} = \{1, 2, \ldots, k\} \subset R$ with $m = 1$ and $q(y = j) = \alpha_j$ with $\alpha_j \geqslant 0$ 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 directly given by the empirical probability density $p_0(x) = (1/N)\sum_{t=1}^N g(x - x_t)$, where $x \in \mathcal{X} = R^n$ and $g(\cdot)$ is a kind of kernel function (e.g., Gaussian function). Moreover, each $q(x|y = j) = q(x|\theta_j)$ is a Gaussian probability density $q(x|m_j, \Sigma_j)$, as given by

$$q(x|\theta_j) = q(x|m_j, \Sigma_j) = \frac{1}{(2\pi)^{1/n}|\Sigma_j|^{1/2}} \mathrm{e}^{-(1/2)(x-m_j)^{\mathrm{T}}\Sigma_j^{-1}(x-m_j)}, \quad (2)$$

where $m_j$ denotes the mean vector and $\Sigma_j$ denotes the covariance matrix which is assumed to be positive definite. Moreover, $p(y|x)$ is a probability distribution that is free to be determined under the general constraints: $p(j|x) \geqslant 0$ and $\sum_{j=1}^k p(j|x) = 1$.

Putting all these component densities into Eq. (1) and letting the kernel functions approach the delta functions $\delta(x)$, $H(p||q)$ reduces to the following harmony function:

$$J(\Theta_k) = \frac{1}{N}\sum_{t=1}^N \sum_{j=1}^k p(j|x_t)\, \ln[\alpha_j q(x_t|m_j, \Sigma_j)], \quad (3)$$

on the parameters $\Theta_k = \{\alpha_j, m_j, \Sigma_j$

$p(j|x_t)$ leads to the following hard-cut form:

$$p(y|x_t) = \begin{cases} 1 & \text{if } y = \text{argmax}[\alpha_j q(x_t|m_j, \Sigma_j)], \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

which, together with the maximization with respect to other parameters, leads to the hard-cut EM algorithm suggested in Ref. [15]. As pointed out in Ref. [16], it is the nature by the harmony learning that the global maximization of $J(\Theta_k)$ leads to automatical detection of $k^*$ as long as is initially selected to be greater than $k^*$. On the other hand, the winner-take-all (WTA) competition mechanism by Eq. (4) makes the maximization of Eq. (3) a discrete optimization that is very easy to be trapped into a local maximum.

With the above background, we now derive a simulated annealing BYY harmony learning algorithm that can make each $p(j|x_t)$ gradually shift from a soft version to the WTA hard-cut version by Eq. (4). Specifically, in the light of Eq. (60) of Ref. [17], we consider

$$L_\lambda(\Theta_k) = J(\Theta_k) + \lambda O_N(p(y|x)), \quad (5)$$

where

$$O_N(p(y|x)) = -\frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\ \ln p(j|x_t), \quad (6)$$

and $\lambda \geqslant 0$. When $\lambda = 1$, the maximum of $L_\lambda(\Theta_k)$ is just the Kullback divergence learning that is equivalent to maximum likelihood learning directly on $q(x|\hat{\Theta}_k)$ [15]. However, when $\lambda = 0$, Eq. (5) reduces to Eq. (3). Thus, as $\lambda$ reduces from one to zero, the maximization of $L_\lambda(\Theta_k)$ will make $p(j|x_t)$ shift from a soft version in the conventional EM algorithm to the WTA or hard-cut version by Eq. (4). If we can let $\lambda \to 0$ from $\lambda_0 = 1$ appropriately in a simulated annealing procedure, the maximum of $L_\lambda(\Theta_k)$ will correspond to the global maximum of $J(\Theta_k)$ with a high probability .

Specifically, the parameters in $\Theta_k$ can be divided into two groups: $\Theta_1$ and $\Theta_2$, where $\Theta_1 = \{p(j|x_t),\ t = 1, \ldots, N,\ j = 1, \ldots, k\}$ and $\Theta_2 = \hat{\Theta}_k$. Then, we have

$$\max_{\Theta_k} L_\lambda(\Theta_k) = \max_{\Theta_1,\Theta_2} L_\lambda(\Theta_k) = \max_{\Theta_1,\Theta_2} L_\lambda(\Theta_1, \Theta_2),$$

which can be implemented by an alternative maximization iterative procedure:

S 1: Fix $\Theta_2 = \Theta_2^{old}$, get $\Theta_1^{new} = \text{argmax}_{\Theta_1} L_\lambda(\Theta_1, \Theta_2)$.

S 2: Fix $\Theta_1 = \Theta_1^{old}$, get $\Theta_2^{new} = \text{argmax}_{\Theta_2} L_\lambda(\Theta_1, \Theta_2)$.

This iterative procedure is guaranteed to reduce $L_\lambda(\Theta_k)$ until it converges to a local maximum when $\lambda$ is fixed. Furthermore, $\Theta_1^{new}$ and $\Theta_2^{new}$ can be solved in detail as follows.

On the one hand, we fix $\Theta_2$ and solve the maximum of $\Theta_1$. Since $\sum_{j=1}^{k} p(j|x_t) = 1$ for each $x_t$, we introduce $N$ Lagrange multipliers $\beta_1, \ldots, \beta_N$, and construct the following Lagrange function:

$$L_\lambda(\Theta_k, \beta_1, \ldots, \beta_N) = L_\lambda(\Theta_k) + \sum_{t=1}^{N}\beta_t\left(\sum_{j=1}^{k} p(j|x_t) - 1\right). \quad (7)$$

By letting the derivatives of $L_\lambda(\Theta_k, \beta_1, \ldots, \beta_N)$ with respect to all $\beta_t$ and $p(j|x_t)$ be zeros, we obtain a series of equations:

$$\ln[\alpha_j q(x_t|m_j, \Sigma_j)] - \lambda(1 + \ln\ p(j|x_t)) + \beta_t = 0, \quad (8)$$

$$\sum_{j=1}^{k} p(j|x_t) = 1, \quad (9)$$

for $t = 1, \ldots, N$; $j = 1, \ldots, k$. From these equations, we have a unique solution for $\Theta_1$:

$$p(j|x_t) = \frac{[\alpha_j q(x_t|m_j, \Sigma_j)]^{1/\lambda}}{\sum_{i=1}^{k}[\alpha_i q(x_t|m_i, \Sigma_i)]^{1/\lambda}},$$
$$t = 1, \ldots, N;\ \ j = 1, \ldots, k. \quad (10)$$

On the other hand, we fix $\Theta_1$ and solve the maximum of $\Theta_2$. Since $\sum_{j=1}^{k}\alpha_j = 1$, we introduce another Lagrange multiplier $\beta$ and construct the following Lagrange function:

$$L_\lambda(\Theta_k, \beta) = L_\lambda(\Theta_k) + \beta\left(\sum_{j=1}^{k}\alpha_j - 1\right). \quad (11)$$

By letting the derivatives of $L_\lambda(\Theta_k, \beta)$ with respect to $\beta$ and all the parameters in $\Theta_2$ be zeros, we have another series of equations as follows:

$$\frac{1}{N}\sum_{t=1}^{N} p(j|x_t)\frac{1}{\alpha_j} - \beta = 0, \quad (12)$$

$$\frac{1}{N}\sum_{t=1}^{N} p(j|x_t)(x_t - m_j) = 0, \quad (13)$$

$$\frac{1}{2N}\sum_{t=1}^{N} p(j|x_t)\Sigma_j^{-1}[(x_t - m_j)(x_t - m_j)^{\mathrm{T}} - \Sigma_j]\Sigma_j = 0, \quad (14)$$

$$\sum_{j=1}^{k}\alpha_j = 1, \quad (15)$$

for $j = 1, \ldots, k$. By solving this series of equations, we have the following unique solution for $\Theta_2$:

$$\hat{\alpha}_j = \frac{1}{N}\sum_{t=1}^{N} p(j|x_t), \quad (16)$$

$$\hat{m}_j = \frac{1}{\sum_{t=1}^{N} p(j|x_t)}\sum_{t=1}^{N} p(j|x_t)x_t, \quad (17)$$

$$\hat{\Sigma}_j = \frac{1}{\sum_{t=1}^{N} p(j|x_t)}\sum_{t=1}^{N} p(j|x_t)(x_t - \hat{m}_j)(x_t - \hat{m}_j)^{\mathrm{T}}. \quad (18)$$

From the above derivations, we have already established an alternative optimization algorithm for maximizing $L_\lambda(\Theta_k)$. It takes the same form as the standard EM algorithm for the Gaussian mixtures, but differs in the E-step, in which the posteriori probabilities $p(j|x_t)$ tend to be the hard-cut version as $\lambda \to 0$.

If $\lambda$ attenuates appropriately along time, this alternative maximization algorithm anneals to search for the global maximum of $J(\Theta_k)$, which further leads to automated model selection with parameter estimation. That is, when we set to be greater than $k^*$ (i.e., the number of actual Gaussians in the sample data set $D_x$), the annealing learning algorithm can make $k^*$ Gaussians in the estimated mixture match the actual Gaussians upon convergence and force the mixing proportions of the other $(k-k^*)$ extra Gaussians to vanish (i.e., eliminate them from the mixture automatically). For clarity and convenience, we refer to the derived algorithm as the BYY annealing learning algorithm.

Interestingly, the BYY annealing learning algorithm derived here takes a similar form as that of the deterministic annealing EM algorithm proposed in Ref. [11]. Actually, the annealing parameter $\beta$ $(0<\beta_{\min}\leqslant\beta\leqslant 1)$ in the deterministic annealing EM algorithm serves as $1/\lambda$ in the BYY annealing learning algorithm. However, the deterministic annealing EM algorithm makes $\beta$ gradually approach to 1 so that it can search for the global maximum of the likelihood function for overcoming the local maxima problem associated with the conventional EM algorithm. Therefore, the deterministic EM algorithm leads to a good maximum likelihood estimate, but it has no ability to make model selection for the Gaussian mixture. On the contrary, the BYY annealing learning algorithm makes $1/\lambda$ gradually tend to the positive infinity or $\lambda \to 0$ so that it tries to globally maximize the harmony function for the Gaussian mixtures. Hence, the BYY annealing learning algorithm has the ability to make model selection for the Gaussian mixture modeling. As a result, these two annealing-type algorithms anneal in different ways and achieve different goals; nevertheless, they take the similar forms and can be considered to belong to the same family of deterministic annealing EM algorithms.

## 3. Experimental results

In this section, several simulation experiments are carried out to demonstrate the performance of the BYY annealing learning algorithm for automated model selection as well as clustering on the sample data from a Gaussian mixture with certain degree of overlap. Moreover, the BYY annealing learning algorithm is applied to two real-life data sets, including Iris data classification and unsupervised color image segmentation.

### 3.1. O

As shown in Fig. 1, four typical sets of sample data from different Gaussian mixtures were used in our simulation experiments. The sample data in each set were randomly and independently generated from a mixture of four or three bivariate Gaussians with a certain degree of overlap on the plane coordinate system (i.e., $d=2$). These four sets of sample data are quite representative. The Gaussians (i.e., clusters) in $\mathscr{S}_1$ are sphere-shaped, with the equal number of samples. But those in $\mathscr{S}_2$ are ellipse-shaped, with different numbers of samples. Moreover, $\mathscr{S}_3$ consists of three very flat Gaussians and $\mathscr{S}_4$ has a small number of samples, with the same structure as $\mathscr{S}_2$.

We ran the BYY annealing learning algorithm on the given four sample data sets, respectively, by letting $k>k^*$ and setting the stopping criterion: $|J(\Theta_k^{new})-J(\Theta_k^{old})|<10^{-7}$. The initial parameters were randomly selected within certain intervals. However, it was found by the experiments that if the initial mean vectors of Gaussians are trained by the rival penalized competitive learning (RPCL) algorithm [21] on the sample data with a small number of iterations, the BYY annealing learning algorithm converges more quickly. Thus, we always selected the initial mean vectors of Gaussians with the aid of a short RPCL process. For the annealing parameter $\lambda$, we let

$$\lambda=\lambda(t)=\frac{1}{a(1-\mathrm{e}^{-b(t-1)})+c}, \tag{19}$$

where denotes the iteration time. In this case, $a=100$, $b=\ln 10/10\,000$ and $c=0.5$.

The experimental results of the BYY annealing learning algorithm on the four sample sets in the case of $k=8$ and $k^*=4$ (or 3), are given in Figs. 2–5, respectively. In order to vividly describe a Gaussian distribution in the estimated mixture obtained from the algorithm, we use the graded contour lines of its probability density starting from the center point (i.e., the mean vector), and gradually expanding unless the density is less than $\mathrm{e}^{-3}$. From each of these four figures, we observed that there are four (or three) Gaussians, which accurately match the actual ones in the sample data set. Also, we can find that the mixing proportions $\alpha_j$ of the extra Gaussians were forced to be zeros. That is, the BYY annealing learning algorithm can detect the correct number of the Gaussians or clusters in each of these sample data sets. We also observed that an extra Gaussian can be stable with any shape while its mixing proportion is attenuating to zero. Frequently, it degenerates to a point.

In addition to the Gaussian number detection, we further found that the clustering or partitioning result according to the converged posteriori probabilities $p(j|x_t)(=0,1)$ on the given sample data is generally as good as the conventional EM algorithm for the Gaussian mixtures with $k=k^*$. That is, by discarding those extra Gaussians, the final posteriori probabilities $p(j|x_t)$ can lead to a reasonable partition on the sample data.

Through further experiments on these sample sets, we also compared the BYY annealing learning algorithm with other state-of-the-art statistical approaches. In comparison with the simulation results of the gradient-type learning algorithms [18–20] on these four data sets, we have found that our annealing learning algorithm converges more accurately and stably than the gradient-type learning algorithms, for both tasks of correct number detection and data partitioning. Although it takes the form of simulated annealing, our annealing learning algorithm generally converges as quickly as the gradient-type learning algorithms.

In comparison with the hard-cut EM algorithm [15] and the methods of RJMCMC and Dirichlet processes, the BYY annealing learning algorithm has a better convergence behavior. Generally, the BYY annealing learning algorithm is not sensitive to the initial values of the parameters and always leads to a good result. On the contrary, the hard-cut EM algorithm

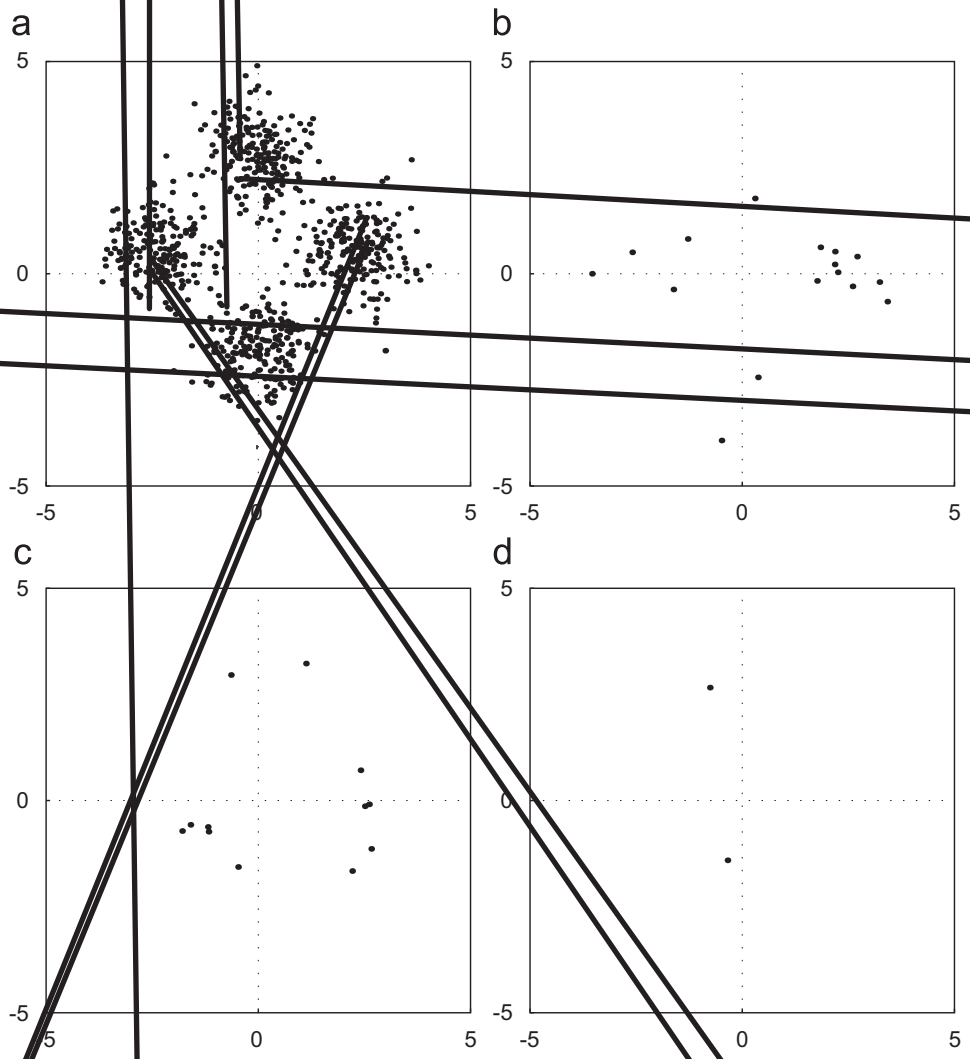


Fig. 1. Four sets of sample data used in the experiments. (a) Set 1 2; (c) set

t) of

estimated Gaussians. To evaluate the classification performance of the algorithm, we compute the classification accuracy given the real categories of the 150 samples.

In our experiments, for ease of classification we first regularized the original Iris data via dividing them by some integers so that they can be located within a reasonable region. Since the

different, we used the four different integers: 35, 20, 25, and 10, to regularize the first to fourth components of the Iris data,

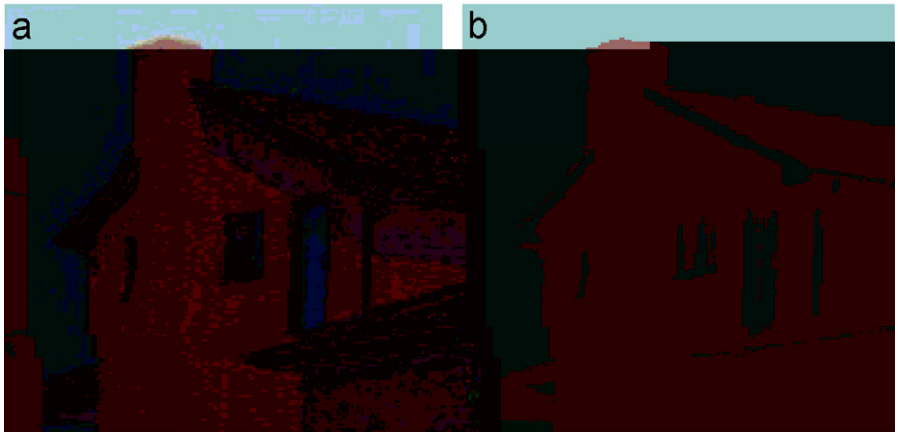


Fig. 6. The segmentation result on the color image of house. (a) The original color image of house; (b) the segmented image via the BYY annealing learning algorithm (after 21 iterations).

we cancel the $j$th Gaussian in the mixture in the later learning iterations. In this case, the learning process was stopped when $|J(\Theta_k^{new}) - J(\Theta_k^{old})| < 10^{-5}$. It was shown in the experiments that the BYY annealing learning algorithm can always correctly detect the three actual classes of the Iris data within 150 iterations. Moreover, the optimal classification accuracy of the BYY annealing learning algorithm could reach 98% (there are only two errors in the second class and one error in the third class), which is as good as the optimal classification accuracy of the maximum certainty partitioning method with a large number of linear mixing Gaussian kernels [12].

### 3.3. O

Finally, we applied the BYY annealing learning algorithm to the problem of unsupervised color image segmentation, which has been recognized as a promising and challenging topic in image processing [23]. Segmenting a digital color image into homogenous regions corresponding to the objects (including the background) is a fundamental problem in image processing. When the number of objects in an image is not known in advance, the image segmentation problem is in an unsupervised mode and becomes rather difficult in practice. If we consider each object as a Gaussian distribution, the whole color image can be regarded as a Gaussian mixture in the data or color space. Then, the BYY annealing learning algorithm provides a new tool for solving this unsupervised color image segmentation problem. In the following, we applied it to the unsupervised color image segmentation on three typical color images that are expressed in the three-dimensional color space by the RGB system. Specifically, we used each Gaussian in the algorithm to represent an object in a color image and set to be greater than the number $k^*$ of the actual objects in the image. When the mixing proportion of the estimated Gaussians are less than a small threshold $T$, we eliminate these Gaussians immediately. Finally, the pixels in the image are partitioned according to the posteriori probabilities $p(j|x_t)$ of the final estimated Gaussians.

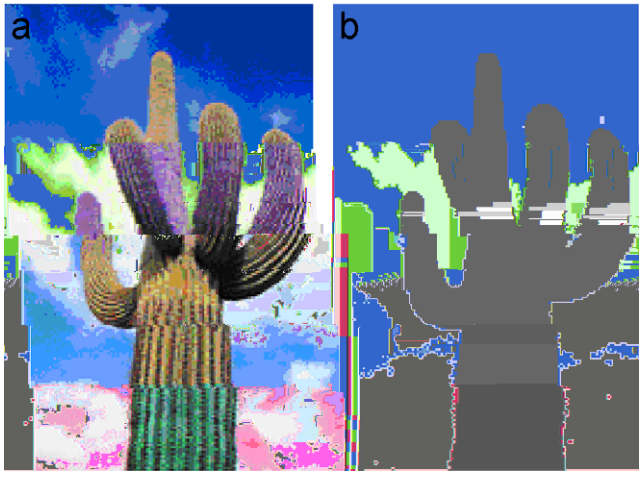


Fig. 7. The segmentation result on the color image of cactus. (a) The original color image of cactus; (b) the segmented image via the BYY annealing learning algorithm (after 16 iterations).

As shown in Figs. 6–8(a), three typical color images of house, cactus, and jellies, are selected for the segmentation experiments. Each pixel in the image is represented by a three-dimensional point that correspond to the coding of the RGB system. In our experiments, for the ease of segmentation we regularized all the three coordinates of the pixels in each color image via dividing them by 128 so that the regularized coordinates are within a reasonable interval. Upon such a preprocessing, we implemented the BYY annealing learning algorithm on the three color images, respectively, by letting $k = 6$ with a simplified stopping criterion: $\sum_{j=1}^{k} \|$

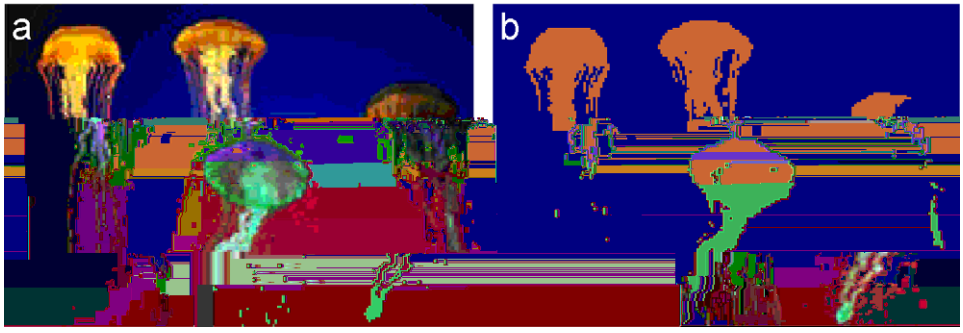


Fig. 8. The segmentation result on the color image of jellies. (a) The original color image of jellies; (b) the segmented image via the BYY annealing learning algorithm (after 22 iterations).

match the actual object boundaries in the image. Also, we found from the experiments that the mixing proportions $\alpha_j$ of those extra objects could be quickly reduced to below the threshold $T$ and be discarded in the algorithm. That is, the BYY annealing learning algorithm can detect the number of actual objects correctly in these color images. Moreover, the segmentation results of the BYY annealing learning algorithm are better than those of the generalized competitive clustering (GCC) algorithm [23] (based on the fuzzy clustering theory). Actually, in comparison with the segmented result of the cactus color image from the web http://www-rocq.inria.fr/~boujemaa/Partielle2.html, we found that the BYY annealing learning algorithm obtains a more accurate segmentation on the contours of the objects in the same cactus color image.

## 4. Conclusions

We have proposed a Bayesian Ying–Yang (BYY) annealing learning algorithm for data clustering or partition with automated model selection. The algorithm is derived to search for the global maximum of the harmony function on a specific B-architecture of the BYY system in a simulated annealing way, such that the posteriori probabilities of the Gaussian mixture gradually change from the soft version to the final hard-cut version. Our algorithm can be considered as a kind of simulated annealing EM algorithm for the Gaussian mixture, but it outperforms the deterministic annealing EM algorithm [11] with a new feature that the model selection can be performed automatically along with parameter learning. The simulation experiments have shown that the BYY annealing learning algorithm can automatically and efficiently determine the number of clusters or Gaussians for learning a parametric mixture model of a sample data set. Moreover, the BYY annealing learning algorithm succeeds in two real-life unsupervised learning tasks, including of Iris data classification and color image segmentation.

## Acknowledgments

This work was supported by the Natural Science Foundation of China for Projects 60071004, 60471054. The authors thank Prof. Taijun Wang for his simulation supports.

## References

[1] P.A. Devijver, J. Kittter, Pattern Recognition: A Statistical Approach, Prentice-Hall, Englewood Cliffs, NJ, 1982.
[2] R.A. Render, H.F. Walker, Mixture densities, maximum likelihood and the EM algorithm, SIAM Rev. 26 (2) (1984) 195–239.
[3] J.A. Hartigan, Distribution problems in clustering, in: J. Van Ryzin (Ed.), Classification and Clustering, Academic Press, New York, 1977, pp. 45–72.
[4] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control AC-19 (1974) 716–723.
[5] H. Bozdogan, Model selection and Akaike's information criterion: the general theory and its analytical extensions, Psychometrika 52 (3) (1987) 345–370.
[6] M.D. Escobar, M. West, Bayesian density estimation and inference using mixtures, J. Am. Stat. Assoc. 90 (430) (1995) 577–588.
[7] S. Recgardson, P.J. Green, On Bayesian analysis of mixtures with an unknown number of components, J. R. Stat. Soc. B 59 (4) (1997) 731–792.
[8] T. Hofmann, J.M. Buhmann, Pairwise data clustering by deterministic annealing, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1) (1997) 1–14.
[9] M. Kloppenburg, P. Tavan, Deterministic annealing for density estimation by multivariate normal mixtures, Phys. Rev. E 55 (3) (1997) R2089–R2092.
[10] K. Rose, Deterministic annealing for clustering, compression, classification, regression, and related optimization problem, Proc. IEEE 86 (11) (1998) 2210–2239.
[11] N. Ueda, R. Nakano, Deterministic annealing EM algorithm, Neural Networks 11 (1998) 271–282.
[12] S.J. Robert, R. Everson, I. Rezek, Maximum certainty data partitioning, Pattern Recognition 33 (2000) 833–839.
[13] N. Ueda, Z. Ghahramani, Bayesian model search for mixture models based on optimizing variational bounds, Neural Network 15 (2002) 1223–1241.
[14] 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), 30 October–3 November 1995, vol. 2, pp. 977–988.
[15] L. Xu, Bayesian Ying–Yang machine, clustering and number of clusters, Pattern Recognition Lett. 18 (1997) 1167–1178.
[16] 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, Int. J. Neural Syst. 11 (1) (2001) 43–69.
[17] L. Xu, BYY harmony learning, structural RPCL, and topological self-organizing on mixture modes, Neural Networks 15 (2002) 1231–1237.
[18] J. Ma, T. Wang, L. Xu, A gradient BYY harmony learning rule on Gaussian mixture with automated model selection, Neurocomputing 56 (2004) 481–487.

[19] J. Ma, B. Gao, Y. Wang, Q. Cheng, Conjugate and natural gradient rules for BYY harmony learning on Gaussian mixture with automated model selection, Int. J. Pattern Recognition Artif. Intell. 19 (2005) 701–713.
[20] J. Ma, L. Wang, BYY harmony learning on finite mixture: adaptive gradient implementation and a floating RPCL mechanism, Neural Processing Lett 24 (1) (2006) 19–40.
[21] L. Xu, A. Krzyak, E. Oja, Rival penalized competitive learning for clustering analysis, RBF net, and curve detection, IEEE Trans. Neural Networks 4 (4) (1993) 636–649.
[22] C.L. Blake, C.J. Merz, UCI repository of machine learning databases ⟨http://www.ics.uci.edu/∼mlearn/MLRepository.html⟩, University of California, Irvine, Department of Information and Computer Science, 1998.
[23] 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.

**About the Author**—JINWEN MA received the Master of Science degree in applied mathematics from Xi'an Jiaotong University in 1988 and the Ph.D. degree in probability theory and statistics from Nankai University in 1992. From July 1992 to November 1999, he was a lecturer or associate professor at Department of Mathematics, Shantou University. From December 1999, he became a full professor at Institute of Mathematic, Shantou University. From September 2001, he has joined the Department of Information Science at the School of Mathematical Sciences, Peking University, where he is currently a full professor and Ph.D. advisor. During 1995 and 2003, he also visited several times the Department of Computer Science & Engineering, the Chinese University of Hong Kong as a Research Associate or Fellow. He also worked as Research Scientist at Amari Research Unit, RIKEN Brain Science Institute, Japan from September 2005 to August 2006. He has published over 80 academic papers on neural networks, pattern recognition, learning theory and algorithm, and information theory.

**About the Author**—JIANFENG LIU received the Bachelor of Science from the Department of Information Science at the School of Mathematical Sciences, Peking University in 2004. Recently, he is a graduate student at the same department. His interests includes pattern recognition, learning theory and algorithm, and bioinformatics.