International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 18, No. 8 (2004) 1501–1512

ENTROPY PENALIZED AUTOMATED MODEL SELECTION ON GAUSSIAN MIXTURE*

JINWEN MA
Department of Information Science, School of Mathematical Sciences & LMAM, Peking University, Beijing, 100871, P. R. China
jwmamath.pku.edu.cn

TAIJUN WANG
Department of Wireless Engineering, Southeast University, Nanjing, 210018, P. R. China

Gaussian mixture modeling is a powerful approach for data analysis and the determination of the number of Gaussians, or clusters, is actually the problem of Gaussian mixture model selection which has been investigated from several respects. This paper proposes a new kind of automated model selection algorithm for Gaussian mixture modeling via an entropy penalized maximum-likelihood estimation. It is demonstrated by the experiments that the proposed algorithm can make model selection automatically during the parameter estimation, with the mixing proportions of the extra Gaussians attenuating to zero. As compared with the BYY automated model selection algorithms, it converges more stably and accurately as the number of samples becomes large.

Keywords: Gaussian mixture; model selection; maximum-likelihood estimation; Shannon entropy; penalty.

1. Introduction

Since Gaussian mixture populations are very common in various elds of practical applications, Gaussian mixture modeling is a powerful approach for data analysis. Actually, considerable e ort has been made on Gaussian mixture modeling as well as its applications for clustering analysis on a sample data set.[15] Although there have been various statistical or unsupervised competitive learning methods to do such a task, e.g. EM algorithm[17] for maximum-likelihood, -means algorithm[14] for the Mean Square Error (MSE) clustering and the self-organizing network for hyperellipsoidal clustering (HEC),[9] it is usually assumed that the number of Gaussians, or clusters, in the data set is pre-known. However, in many instances this key information is not available and then the selection of an appropriate number of Gaussians must be made before or during the estimation of the parameters in the

*This work was supported by the Natural Science Foundation of China for Project 60071004.

mixture. Since the number of Gaussians is just the scale of the mixture model, its determination is actually the problem of model selection for Guassian mixture modeling. In fact, it is a rather di cult problem.[8]

The earliest possible method for solving this model selection problem may be to choose the optimal number * of Gaussians (or clusters) by the Akaike's information criterion[2] or its extensions.[4,5,19] Similarly, several indexes of cluster validity have been also proposed to choose the optimal number * (e.g. Refs. 3, 16 and 21). But 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 number of di erent values of , even though such a process is attempted to be organized in a more e cient way, e.g. embedding the checking of the criterion value within clustering as in ISODATA.

From the respect of neural networks, competitive learning (CL) has been developed for clustering analysis and vector quantization.[6] However, the conventional competitive learning algorithms such as the classical competitive learning algorithm[18] and the frequency sensitive competitive learning algorithm,[1] can be only considered as adaptive versions of -means algorithm and thus they are unable to solve this model selection problem. In order to do so, Xu, Krzyzak and Oja proposed a new kind of competitive learning algorithm, called rival penalized competitive learning (RPCL) algorithm,[22] which has the ability of automatically allocating an appropriate number of weight vectors for a sample data set, with the other extra ones being pushed far away from the sample data. Therefore, the RPCL algorithm can be used to solve the above model selection problem by choosing a number of the weight vectors which is surely larger than the number of actual clusters in the sample data set. In some extended versions of the RPCL algorithm,[24] the weight vectors have been generalized to Gaussian densities, which makes it more suitable for the RPCL algorithm to be applied to solving the model selection problem for Gaussian mixture. Recently, Ma, Wang and Xu proposed a cost function by which a generalized RPCL algorithm, called distance sensitive RPCL algorithm, has been proposed with a more powerful performance and easier implementation.[10]

Proposed in 1995[23] and systematically developed in the past years,[25–27] Bayesian Ying–Yang (BYY) harmony learning acts as a general statistical learning framework not only for understanding several existing major learning approaches but also for tackling the learning problem on a set of nite samples with a new learning mechanism that makes model selection to be implemented automatically during parameter learning via a new class of model selection criteria. Speci cally, the BYY harmony learning can be applied to the model selection problem for Gaussian mixture modeling. In fact, Ma, Wang and Xu have already implemented this new mechanism on a BI-architecture and a B-architecture of the BYY system via a gradient, iterative and two annealing algorithms to solve the model selection problem automatically during the parameter learning.[11–13,20] It has been demonstrated by simulation experiments that the number of Gaussians can be automatically selected for the sample data set, with the mixing proportions of the extra Gaussians attenuating to zero.

In the current paper, based on the fact that reducing the Shannon entropy of the xing proportions in a Gaussian mixture can lead to some of these proportions to tenuate to zero, i.e. to eliminate the corresponding Gaussians from the mixture, proposed an entropy penalized maximum-likelihood estimation (MLE) iterative gorithm which can also make the parameter estimation with automated model ection. It is further demonstrated by the simulation experiments.

In the sequel, the entropy penalized maximum-likelihood estimation (MLE) rative algorithm is derived in Sec. 2. In Sec. 3, several simulation experiments e conducted to demonstrate the proposed algorithm, with a comparison between and those BYY model selection algorithms. A brief conclusion is given in Sec. 4.

Entropy Penalized MLE Iterative Algorithm

e consider the following Gaussian mixture model:

$$(x\mathrm{j}\ ) = \sum_{j=1}^{K}\ _j\ (x\mathrm{j}\ _j\ \ _j) \qquad _j\ \ 0\ \sum_{j=1}^{K}\ _j = 1 \tag{1}$$

here

$$(x\mathrm{j}\ \ _j\ \ _j) = \frac{1}{(2\ )^{\frac{d}{2}}\mathrm{j}\ \ _j\mathrm{j}^{\frac{1}{2}}}\ ^{-\frac{1}{2}(x-m_j)^T\Sigma_j^{-1}(x-m_j)} \tag{2}$$

and where is the number of Gaussians, x denotes a sample vector and is the dimensionality of x The parameter vector consists of the mixing proportions $_j$, the mean vectors $_j$, and the covariance matrices $_j = (\ ^{(j)}_{pq})_{d\times d}$ which are assumed symmetric and positive de nite.

In this mixture model, $_j$ is the proportion or probability of Gaussian (component) in the mixture population. Then, the Shannon entropy of these mixing proportions, i.e. $(\ ) = \sum_{j=1}^{K}\ _j \log\ _j$, represents the uncertainty of the distribution of these mixing proportions. Here and hereafter, the base of the logarithm is always the natural number. According to the property of Shannon entropy, as increases, $(\ )$ generally increases. Oppositely, as $(\ )$ decreases, some mixing proportions may be forced to be zero or a small number. That is, the corresponding Gaussians are essentially eliminated from the mixture. Thus, the number of actual Gaussians is decreased. According to this characteristic, we can penalize the MLE by adding certain minus Shannon entropy of the mixing proportions to the log maximum likelihood (ML) function so that the entropy penalized MLE is able to make the parameter estimation with automated model selection. In the following, we will realize this idea by proposing an entropy penalized ML objective function and establishing an iterative algorithm to maximize it.

Given a sample data set $\mathrm{S} = \mathrm{f}x_t\mathrm{g}_{t=1}^{N}$ from a Gaussian mixture with * components and *, the log likelihood function on the mixture model $(x\mathrm{j}\ )$ is given by

$$L(\ ) = \sum_{t=1}^{N} \log\ \ (x_t\mathrm{j}\ ) \tag{3}$$

The entropy penalized ML objective function can be constructed by

$$(\quad) = L(\quad) \qquad (\quad) \tag{4}$$

where 0 is the penalty factor which is a constant.

According to this objective function ( ), we have an entropy penalized maximum-likelihood estimate:

$$\hat{} = \arg\max_{\Theta} (\quad) \tag{5}$$

By Lagrange's method of multipliers for ^, we begin to de ne the -function:

$$L(\quad) = (\quad) + \left(1 \quad \sum_{j=1}^{K} \quad _j \right) \tag{6}$$

Using the general methods for matrix derivatives,[7] we are led to the derivatives of L( ) with respect to each $_j$ $_j$ $_j$ and as follows:

cording to Eqs. (12) and (15), we further have

$$= \sum_{j=1}^{K}\sum_{t=1}^{N} {}_j(\ ) + \sum_{j=1}^{K}(\ {}_j + \ {}_j \log\ {}_j)$$

$$= \ + \ (1 \ (\ ))$$

$$= \ (1 + \ (1 \ (\ ))) \qquad (16)$$

hen, from the Eqs. (12){(14), we obtain the following iterative algorithm for ^:

$${}_j^{+} = \frac{1}{(1 + \ (1 \ (\ )))}\left(\sum_{t=1}^{N} {}_j(\ ) + \ {}_j(1 + \log\ {}_j)\right); \qquad (17)$$

$${}_j^{+} = \frac{1}{\sum_{t=1}^{N} {}_j(\ )}\sum_{t=1}^{N} {}_j(\ )x_t; \qquad (18)$$

$${}_j^{+} = \frac{1}{\sum_{t=1}^{N} {}_j(\ )}\sum_{t=1}^{N} {}_j(\ )(x_t \ {}_j)(x_t \ {}_j)^T \qquad (19)$$

As compared with the EM algorithm for Gaussian mixture, this iterative algorithm implements a penalty mechanism on the mixing proportions during the iterations, which leads to the automated model selection. The penalty factor can be selected by experience.

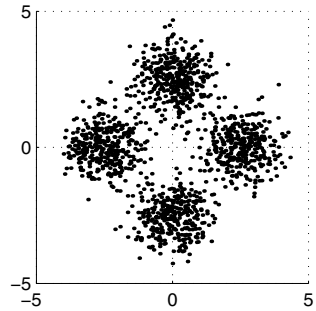
3. Simulation Results

In this section, simulation experiments are carried out to demonstrate the entropy penalized MLE iterative algorithm for both model selection and parameter estimation on a set of sample data from a Gaussian mixture. Moreover, this iterative algorithm is compared with the BYY model selection algorithms.

3.1. *The sample data*

We begin with a description of the seven sets of sample data used for our simulation experiments. We conducted four Monte Carlo experiments in which samples were drawn from a mixture of four or three bivariate Gaussians densities (i.e. = 2).

As shown in Fig. 1, each data set of samples are generated with a certain degree of overlap among the Gaussians (clusters) in the mixture. They are four typical sets of sample data from Gaussian mixtures. The Gaussians in S_1 are sphere-shaped, with equal number of samples. But those in S_2 are ellipse-shaped, with di erent numbers of samples. Moreover, S_3 consists of three very at Gaussians and S_4 has a small number of samples, with similar structure as S_2 The detailed parameters for these four sets of sample data are given in Table 1 where $_i \ _i = [\ {}_{jk}^{i}]_{2\times 2} \ _i$ and $_i$ denote the mean vector, covariance matrix, mixing proportion and the number of samples of the th cluster (Gaussian), respectively.

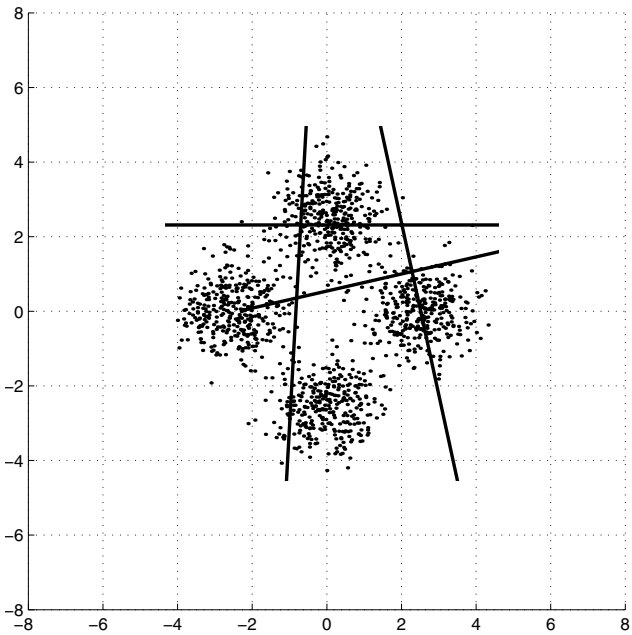

5
0
−5
−5
0
5

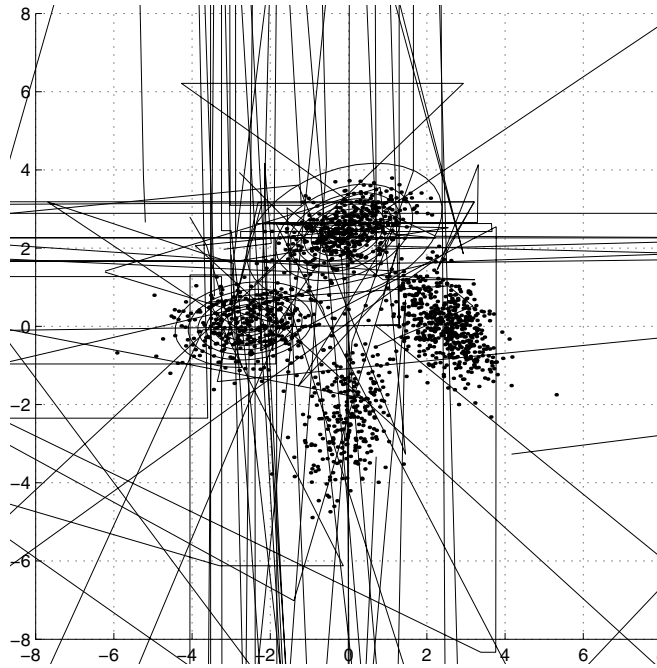
2. *Automated model selection*

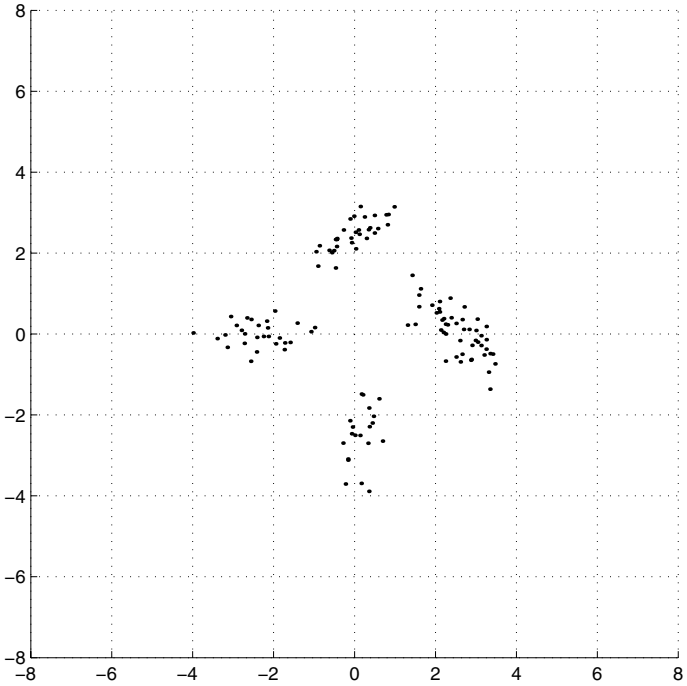
e implement the entropy penalized MLE iterative algorithm on those four sample ta sets always with *, respectively. The parameters are initialized ndomly within certain intervals, satisfying the required constraints. The penalty ctor is selected to be 0.15. In all the experiments, the algorithm is stopped when $(^{\text{new}}_{k})$ $(^{\text{old}}_{k})$| 10^{-7}

The experimental results on S_1 and S_2 are given in Figs. 2 and 3, respectively, th case = 8 and * = 4 We observe that four Gaussians are nally located curately, while the mixing proportions of the other four Gaussians were reduced 0.002 or below, i.e. these Gaussians are extra and can be discarded. That is, the rrect number of the Gaussians has been automatically determined on these data s.

Moreover, the experiment has been made on S_3 with case = 8 * = 3 As own in Fig. 4, the actual Gaussians are very at. However, three Gaussians are still cated accurately, while the mixing proportions of the other ve extra Gaussians e no more than 0.0012. That is, the correct number of the Gaussians can still determined on such a special data set. Furthermore, the iterative algorithm is o implemented on S_4 with = 8 * = 4 As shown in Fig. 5, each cluster has small number of samples, the correct number of Gaussians can still be detected, th the mixing proportions of other four extra Gaussians reduced below 0.008.




8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8


8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8

3.3. *Parameter estimation*

eferences

S. C. Ahalt, A. K. Krishnamurty, P. Chen and D. E. Melton, Competitive learning algorithm for vector quantization, **Neural Networks 3**(3) (1990) 277–291.

H. Akaike, A new look at the statistical model identification, **IEEE Trans. Autom. Contr. AC-19** (1974) 716–723.

J. C. Bezdek and N. R. Pal, Some nex indexes of cluster validity, Contr.

21. X. L. Xie and G. A. Beni, A validity method for fuzzy clustering, **IEEE Trans. Patt. Anal. Mach. Intell. 13**(8) (1991) 843–847.
22. L. Xu, A. Krzyzak and E. Oja, Rival penalized competitive learning for clustering analysis, RBF net, and curve detection, **IEEE Trans. Neural Networks 4** (1993) 636–648.
23. L. Xu **et al.**, A unified learning scheme: Bayesian–Kullback YING–YANG machine, in **Advances in Neural Information Processing Systems**, eds. D. S. Touretzky **et al.** (MIT Press, Cambridge, MA, 1996), Vol. 8, pp. 444–450; A part of its preliminary version on **Proc. 1995 Int. Conf. Neural Information Processing**, pp. 977–988.
24. L. Xu, Rival penalized competitive learning, finite mixture, and multisets clustering, in **Proc. 1998 IEEE Int. Joint Conf. Neural Networks**, Vol. 3, pp. 251–253.
25. 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) (2002) 43–69.
26. L. Xu, BYY harmony learning, structural RPCL, and topological self-organzing on mixture modes, **Neural Networks 15** (2002) 1231–1237.
27. L. Xu, Ying–Yang learning, in **The Handbook of Brain Theory and Neural Networks**, ed. M. A. Arbib, 2nd edn. (The MIT Press, 2002), pp. 1231–1237.

Jinwen Ma received the M.S. degree in applied mathematics from Xi'an Jiaotong University in 1988 and the Ph.D. in probability theory and statistics from Nankai University in 1992. From July 1992 to November 1999, he was a Lecturer and Associate Professor at the Department of Mathematics, Shantou University. He has been a full Professor at the Institute of Mathematics, Shantou University since December, 1999. In September 2001, he was transferred to the Department of Information Science at the School of Mathematical Sciences, Peking University. During 1995 and 2003, he also made several visits to the Department of Computer Science and Engineering, the Chinese University of Hong Kong as a Research Associate or Fellow.

He has published over 40 academic papers on neural networks, pattern recognition, artificial intelligence and information theory.

Taijun Wang received the M.S. degree in signal and information processing from Southeast University, China, in 1982. He has been an Associate Professor at the Department of Radio Engineering of Southeast University since 1988.

His research interests are related to digital signal processing, pattern recognition, artificial neural networks, statistical learning theory and scientific visualization.