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

World Scientific
www.worldscientific.com

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 fields of practical applications, Gaussian mixture modeling is a powerful approach for data analysis. Actually, considerable effort 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, K-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 K 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 K 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 difficult problem.[8]

The earliest possible method for solving this model selection problem may be to choose the optimal number K^* 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 K^* (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 different values of K, even though such a process is attempted to be organized in a more efficient 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 K-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 K 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 finite 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. Specifically, 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.

The entropy penalized ML objective function can be constructed by

$$f(\ ) = L(\ ) \quad N I(\ ); \tag{4}$$

where > 0 is the penalty factor which is a constant.

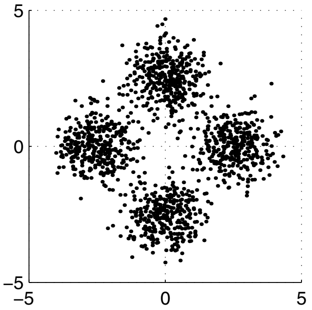
According to this objective function f( ), we have an entropy penalized maximum-likelihood estimate:

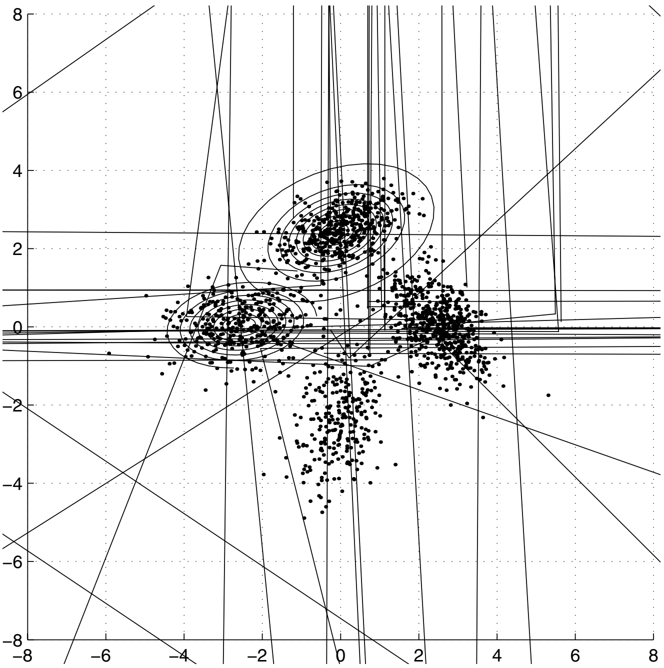
$$\hat{\ } = \arg\max_{\Theta} f(\ ): \tag{5}$$

By Lagrange's method of multipliers for $\hat{\ }$, we begin to define the -function:

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

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


5
0
−5
−5
0
5


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

3.3. *Parameter estimation*

21. X. L. Xie and G. A. Beni, A validity method for fuzzy clustering, *n An M c n* **13**(8) (1991) 843–847.
22. L. Xu, A. Krzyżak and E. Oja, Rival penalized competitive learning for clustering analysis, RBF net, and curve detection, *n* **4** (1993) 636–648.
23. L. Xu , A unified learning scheme: Bayesian–Kullback YING–YANG machine, in *Adv nc n n o n on oc n y n*, eds. D. S. Touretzky (MIT Press, Cambridge, MA, 1996), Vol. 8, pp. 444–450; A part of its preliminary version on *oc n on n o n on oc n* , pp. 977–988.
24. L. Xu, Rival penalized competitive learning, finite mixture, and multisets clustering, in *oc n o n on* , 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, *n y* **11**(1) (2002) 43–69.
26. L. Xu, BYY harmony learning, structural RPCL, and topological self-organzing on mixture modes, **15** (2002) 1231–1237.
27. L. Xu, Ying–Yang learning, in *nd oo o B n o y nd* , 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.

[redacted]

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.