Simultaneous Model Selection and Feature Selection via BYY Harmony Learning

Hongyan Wang and Jinwen Ma*

Department of Information Science, School of Mathematical Sciences & LMAM
Peking University, Beijing, 100871, P.R. China

Abstract. Model selection for Gaussian mixture learning on a given dataset is an important but difficulty task and also depends on the feature or variable selection in practical applications. In this paper, we propose a new kind of learning algorithm for Gaussian mixtures with simultaneous model selection and variable selection (MSFS) based on the BYY harmony learning framework. It is demonstrated by simulation experiments that the proposed MSFS algorithm is able to solve the model selection and feature selection problems of Gaussian mixture learning on a given dataset simultaneously.

Keywords: Gaussian mixtures, Baysian Ying-Yang (BYY) Harmony learning, Model selection, Feature selection, Clustering analysis.

1 Introduction

Finite mixture models [1] are flexible and powerful statistical tools for data analysis and information processing. In fact, they been extensively used in a variety of practical applications such as clustering analysis, image segmentation and speech recognition. Among these applications, the Gaussian mixture model is very popular and very important in theory and practice. In order to solve the problem of Gaussian mixture modeling, several statistical learning methods have been established, such as the EM algorithm [2]-[3]. However, the conventional learning algorithm cannot solve the model selection problem, i.e., to determine the number of Gaussians for a given dataset. When the Gaussian mixture model is applied to clustering analysis, the model selection problem is just to determine the number of clusters for a dataset. Since the number of Gaussians or clusters is not available in the general cases, model selection must be made with the parameter estimation, which is a rather complicated and difficult task [4].

The other crucial problem on Gaussian mixture learning is feature selection. In principle, the more information we have about each individual, the better a learning method is expected to perform. But in practice, some features are noises and may degrade the learning performance, especially in high-dimension circumstances. A genic dataset usually has a limited number of observations with thousands of features. Actually, there are a large number of features which are

* Corresponding author, jwma@math.pku.edu.cn

D. Liu et al. (Eds.): ISNN 2011, Part II, LNCS 6676, pp. 47–56, 2011.

irrelevant to the learning or classification problem. So, feature selection is necessary. In fact, feature selection has been investigated in the context of supervised learning scenarios [5]-[8]. It was shown in [9] that feature selection can improve the performance of a supervised classifier on learning from a limited number of data points. But for unsupervised learning or clustering analysis, because of the lack of labels as guidance, it is rather difficult for a learning method to achieve the feature or variable selection together with the parameter learning.

As the model selection is related to the feature selection on Gaussian mixture learning, it is reasonable to consider the two selection problems simultaneously under a unified framework. In fact, there have been two investigations on this aspect directly for clustering analysis. Martin et al. [10] proposed a simultaneous feature selection and clustering method using mixture models through the concept of feature saliency and the EM algorithm. On the other hand, Li et al. [11] proposed a simultaneous localized feature selection and model detection for Gaussian mixtures by Bayesian variational learning. Now, we try to propose a a simultaneous model selection and variable selection (MSFS) algorithm for Gaussian mixtures based on the Bayesian Ying-Yang (BYY) harmony learning system and theory [12]-[13].

The remainder of this paper is organized as follows. We begin with a brief description of related works on model selection and feature selection in Section 2. In Section 3, we present our simultaneous model selection and feature selection algorithm for Gaussian mixtures. Section 4 contains the experimental results. Finally, we conclude briefly in Section 5.

2 Related Works

2.1 Feature Selection

Feature selection algorithms can be broadly divided into two categories: filters and wrappers. The so-called “filter” approaches select proper features before the learning process or clustering analysis. They evaluate the relevance of each feature to the learning problem using the dataset alone [14]-[15]. Independent selection of the features may influence the effect of learning or clustering. On the other hand, the so-called “wrapper” approaches combine the learning or clustering algorithm with evaluating the quality of each feature. Specifically, a learning algorithm (distance-based [16]-[17] or model-based [18]-[19]) can be implemented for each feature subset. Then this feature subset is evaluated by the performance of learning or clustering. From this point of view, the “wrappers” approaches are usually more computationally demanding since they evaluate all feature subsets.

Intuitionally, feature selection is choosing relevant features, and there are many definitions of feature irrelevancy for supervised learning, such as the correlation or mutual information. Here, we adopt such a definition of feature irrelevancy for unsupervised learning that the i-th variable is irrelevant if its distribution is independent of the class labels. This means that the i-th variable is irrelevant when it comes from a common distribution $p(y_l|\lambda_l)$ which is independent with labels. By contrast, we define the density of a relevant feature l by $p(y_l|\theta_{jl})$ for j-th component of the mixture model. Based on these definitions, if we assume that these variables are independent, the likelihood function can be written as the following form (refer to [10],[11]):

$$p(y|\theta) = \sum_{j=1}^{k} \alpha_j p(y|\theta_j) = \sum_{j=1}^{k} \alpha_j \prod_{l=1}^{D} (\rho_l p(y_l|\theta_{jl}) + (1-\rho_l) q(y_l|\lambda_l)), \qquad (1)$$

where ρ_l is the probability that i-th feature is relevant and θ_{jl} and λ_l are the parameters.

2.2 Model Selection

The traditional approaches to solving the compound Gaussian mixture modeling problem of model selection and parameter learning or estimation are to choose an optimal number k^* of Gaussians as the clusters in the dataset via one of the information, coding and statistical selection criteria such as the famous Akaike's Information Criterion [20], Bayesian Inference Criterion (BIC) [21], Minimum Description Length (MDL) [22], and Minimum Message Length (MML) [23]. Among them, Akaike's information criterion (AIC) and the MML criterion are often used. However, the validating processes of these approaches are computationally expensive because we need to repeat the entire parameter learning process at a large number of possible values of k, i,e, the number of Gaussians in the mixture. Moreover, these existing selection criteria have their limitations.

Since the 1990s, there have appeared some statistical learning approaches to solving this compound modeling problem. The first approach is to utilize certain stochastic simulations to infer the optimal mixture model. Two typical implementations are the methods of Dirichlet processes [24] and reversible jump Markov chain Monte Carlo (RJMCMC) [25]. These stochastic simulation methods generally require a large number of samples through different sampling rules. The second approach is the Bayesian model search based on optimizing the variational bounds [26]-[27]. This approach implements a new selection criterion with the Bayesian variation bound. The third approach is unsupervised learning [28] on finite mixtures (including Gaussian mixture as a particular case) which introduces certain competitive learning mechanism into the EM algorithm such that the model selection can be made adaptively during parameter learning by annihilating the components with very small mixing proportions via the MML criterion. Recently, the Bayesian Ying-Yang (BYY) harmony learning system and theory [12]-[13] haven been developed as a unified statistical learning framework and provided a new statistical learning mechanism that makes model selection adaptively during parameter learning for Gaussian mixtures [29]-[32]. In the following, we will use the BYY harmony learning system as our unsupervised learning framework for Gaussian mixtures.

3 Simultaneous Model Selection and Feature Selection

3.1 BYY Harmony Learning for Gaussian Mixtures

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

$$H(p||q) = \int p(y|x)p(x) \ln[q(x|y)q(y)]dxdy. \tag{2}$$

For the Gaussian mixture model with a given sample dataset $D_x = \{x_t\}_{t=1}^N$, we can utilize the following specific Bi-architecture of the BYY learning system. The inner representation y is discrete in $\mathcal{Y} = \{1, 2, \cdots, k\}$ (i.e., with $m = 1$), while the observation x is continuous from a Gaussian mixture distribution. On the Ying space, we let $q(y = j) = \pi_j \geq 0$ with $\sum_{j=1}^k \pi_j = 1$. On the Yang space, we suppose that $p(x)$ is a latent probability density function (pdf) of Gaussian mixture, with a set of sample data D_x being generated from it. Moreover, in the Ying path, we let each $q(x|y = j) = q(x|m_j, \Sigma_j)$ be a Gaussian probability density with the mean vector m_j and the covariance matrix Σ_j, while the Yang path is constructed under the Bayesian principle by the following parametric form:

$$p(y = j|x) = \frac{\pi_j q(x|m_j, \Sigma_j)}{q(x|\Theta_k)}, \qquad q(x|\Theta_k) = \sum_{j=1}^k \pi_j q(x|m_j, \Sigma_j), \tag{3}$$

where $\Theta_k = \{\pi_j, m_j, \Sigma_j\}_{j=1}^k$ and $q(x|\Theta_k)$ is just a Gaussian mixture model that will approximate the true Gaussian mixture model $p(x)$ hidden in the sample data D_x via the harmony learning on the BYY learning system.

With all these component densities into Eq.(2), we get an estimate of $H(p||q)$ as the following harmony function for Gaussian mixtures with the parameter set Θ_k:

$$J(\Theta_k) = \frac{1}{N} \sum_{t=1}^N \sum_{j=1}^k \frac{\pi_j q(x_t|m_j, \Sigma_j)}{\sum_{i=1}^k \pi_i q(x_t|m_j, \Sigma_j)} \ln[\pi_j q(x_t|m_j, \Sigma_j)]. \tag{4}$$

According to theoretical and experimental results on the BYY harmony learning on the BI-architecture for Gaussian mixtures [29]-[30],[32]-[33], the maximization of the harmony function $J(\Theta_k)$ is able to make model selection adaptively during parameter learning when the actual Gaussians in the sample data are separated in a certain degree. That is, in such a situation, if we set k to be larger than the number k^* of actual Gaussians in the sample data, the maximization of the harmony function can make k^* Gaussians from the estimated mixture match the actual Gaussians, respectively, and force the mixing proportions of $k - k^*$ extra Gaussians to attenuate to zero.

3.2 Proposed BYY Harmony Learning Algorithm

By a transformation, $J(\Theta_k)$ can be divided into two parts:

$$J(\Theta_k) = L(\Theta_k) - O_N(p(y|x)), \tag{5}$$

where the first part is just the log-likelihood function:

$$L(\Theta_k) = \frac{1}{N}\sum_{t=1}^{N}\ln(\sum_{j=1}^{k}(\pi_j q(x_t|m_j, \Sigma_j))), \tag{6}$$

while the second part is the average Shannon entropy of the posterior probability $p(y|x)$ over the sample dataset $\mathcal{D} = \{x_t\}_{t=1}^{N}$:

$$O_N(p(y|x)) = -\frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\ln p(j|x_t). \tag{7}$$

According to Eq.(5), if $-O_N(p(y|x))$ is considered as a regularization term, the BYY harmony learning, i.e., maximizing $J(\Theta_k)$, is a kind of regularized ML learning. This regularization term contributes to avoiding over-fitting and achieving model selection.

If we replace the Likelihood part with (1) and assume that these variables are independent, the maximization of $J(\Theta_k)$ will be able to make model selection and feature selection simultaneously.

$$\begin{aligned} J(\Theta_k) = \frac{1}{N}\sum_{t=1}^{N}\log(\sum_{j=1}^{k}\alpha_j\prod_{l=1}^{D}(\rho_l p(x_{tl}|\theta_{jl}) + (1-\rho_l)q(x_{tl}|\lambda_l))) \\ + \frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\log p(j|x_t), \end{aligned} \tag{8}$$

where θ_{jl}, λ_l are the parameters of Gaussian densities and

$$p(j|x_t) = \frac{\alpha_j \prod_{l=1}^{D}(\rho_l p(x_{tl}|\theta_{jl}) + (1-\rho_l)q(x_{tl}|\lambda_l))}{\sum_{i=1}^{k}\alpha_i\prod_{l=1}^{D}(\rho_l p(x_{tl}|\theta_{il}) + (1-\rho_l)q(x_{tl}|\lambda_l))}. \tag{9}$$

Actually, there have been many learning algorithms to maximize $J(\theta)$. Here, we adopt the fixed-point learning paradigm (refer to [32]) and the learning algorithm can be derived as follows.

Define:

$$\gamma_j(t) = 1 + \log p(j|x_t) - \sum_{i=1}^{k} p(i|x_t)\log p(i|x_t) \tag{10}$$

$$u_{tjl} = \frac{\rho_l p(x_{tl}|\theta_{jl})}{\rho_l p(x_{tl}|\theta_{jl}) + (1-\rho_l)q(x_{tl}|\lambda_l)}; \qquad v_{tjl} = 1 - u_{tjl} \tag{11}$$

By derivation, we have the derivatives of $J(\Theta_k)$ with respect to α_j, θ_{jl} and λ_l, respectively. Letting these derivatives be equal to zero, we get the following fixed-point equations:

$$\hat{\alpha_j} = \frac{\sum_{t=1}^{N} p(j|x_t)\gamma_j(t)}{\sum_{j=1}^{k}\sum_{t=1}^{N} p(j|x_t)\gamma_j(t)} \tag{12}$$

$$\widehat{mean\ in}\ \theta_{jl} = \frac{\sum_{t=1}^{N} p(j|x_t)\gamma_j(t)u_{tjl}x_{tl}}{\sum_{t=1}^{N} p(j|x_t)\gamma_j(t)u_{tjl}} \tag{13}$$

$$\widehat{var\ in}\ \theta_{jl} = \frac{\sum_{t=1}^{N} p(j|x_t)\gamma_j(t)u_{tjl}(x_{tl} - \widehat{mean\ in}\ \theta_{jl})^2}{\sum_{t=1}^{N} p(j|x_t)\gamma_j(t)u_{tjl}} \tag{14}$$

$$\widehat{mean\ in}\ \lambda_l = \frac{\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\gamma_j(t)v_{tjl}x_{tl}}{\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\gamma_j(t)v_{tjl}} \tag{15}$$

$$\widehat{var\ in}\ \lambda_l = \frac{\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\gamma_j(t)v_{tjl}(x_{tl} - \widehat{mean\ in}\ \lambda_l)^2}{\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\gamma_j(t)v_{tjl}} \tag{16}$$

As for ρ_l, since the gradient and Hessian of $J(\Theta_k)$ are very complicated (refer to the appendix), we can use the constrained nonlinear optimization software to find their optimal values in [0,1].

Summarizing the upper results, we obtain a two-step iterated optimization algorithm:

Step 1: Fix ρ_l and estimate α_j, θ_{jl} and λ_l according (12)-(16);

Step 2: Fix α_j, θ_{jl} and λ_l and obtain the optimized ρ_l using the constrained nonlinear optimization software.

In our learning paradigm, k is flexible. However, it should be larger than the true number k^* of Gaussians in the dataset. Then when we repeat these two steps to make the parameter estimation, we cut very weak components (with $\alpha_j < 0.05$) synchronously.

4 Experimental Results

4.1 Three Synthetic Datasets

The first dateset consists of 800 8-dimensional data or points from two classes whose distributions are subject to $\mathcal{N}(m_i, 2.25*I)$, $i = 1, 2$, where I is the identity matrix, $m_1 = (5, 5, 5, 5, 0, 0, 0, 0)$ and $m_2 = (-5, -5, -5, -5, 0, 0, 0, 0)$. In this situation, it is clear that the first four features are informative and the rest four are irrelevant or noises. We check whether our proposed algorithm can distinguish the informative features from the irrelevant ones. The second dataset consists of 1000 10-dimensional data from two classes. For the two classes, their i-th features come from $\mathcal{N}(1/i, 1)$ and $\mathcal{N}(-5/i, 1)$ respectively, for $i = 1, 2, \cdots, 10$. Note that the features are arranged in a descending order of relevance. This design is more challenging for testing the feature selection. The third dataset is an imbalanced one containing three classes with 200, 400 and 200 samples, respectively.

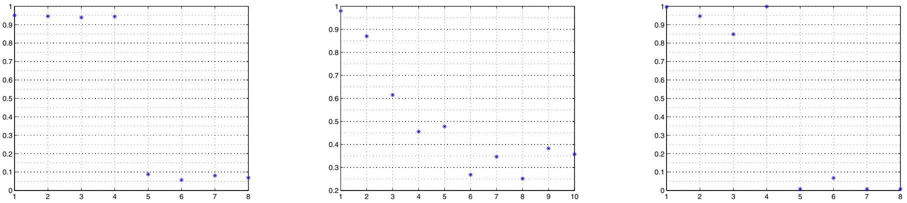

Fig. 1. The sketches of estimated ρ_l for (a). Dataset 1, (b). Dataset 2, (c). Dataset 3.

The Distributions of the three classes are $\mathcal{N}(m_i, 1.414 * I)$, $i = 1, 2, 3$, where I is still the identity matrix, $m_1 = (5, 5, 5, 5, 0, 0, 0, 0)$, $m_2 = (0, 0, 0, 0, 0, 0, 0, 0)$ and $m_3 = (-5, -5, -5, -5, 0, 0, 0, 0)$.

4.2 Simulation Results

We repeat our proposed model selection and feature selection algorithm for Gaussian mixtures on each of these three datasets 50 times with the parameters being randomly initialized. Here, our attention is just focused on the feature selection. So, we only show the average value of ρ_l over 50 experimental results in Fig. 1. It can be seen that our proposed algorithm can successfully distinguish the informative features from the noises, especially for the first dataset on which the last four irrelevant features are found out exactly. The average values of ρ_l are in a descending order just as those features in the second dataset are designed in a descending order of relevance. It can be also noticed that there are some fluctuations along the downtrend. This may be caused by the local optimization of the modified harmony function and can be solved by some global optimization technique.

As for model selection, when k is set to be $2k^*$ (k^* is the true number of Gaussians or classes in the dataset), our proposed algorithm achieves the classification accuracy rates of 66% ,82% and 76% over the three datasets, respectively. Clearly, the model selection result on the first dataset is not so satisfied. In fact, the structure of the first dataset is indeed complicated. For comparison, we implement the RPCL algorithm [34] on the first dataset and generally get a poor clustering result. As for the third dataset, 36 tries out of 50 make the correct model selection and the rest 14 tries lead to 4 clusters. Actually, the largest component is split into two clusters.

5 Conclusions

We have investigated the problem of simultaneous model selection and feature selection for Gaussian mixtures and proposed a new BYY harmony learning algorithm for solving it. The proposed algorithm is constructed in the fixed-point learning paradigm. It is demonstrated by the simulation experiments that the proposed algorithm can simultaneously detect the number of actual Gaussians in the dataset and recognize the informative features accurately.

Acknowledgements

This work was supported by the Natural Science Foundation of China for grant 60771061.

References

1. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 39, 1–38 (1977)
3. Render, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26(2), 195–293 (1984)
4. Hartigan, J.A.: Distribution problems in clustering. In: Garrett, J. (ed.) Classification and Clustering, pp. 45–72. Academic Press, New York (1977)
5. Blum, A., Langley, P.: Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence 97(1-2), 245–271 (1997)
6. Kohavi, R., John, G.H.: Wrapper for feature subset selection. Articial Intelligence 97, 273–324 (1997)
7. Jain, A., Zongker, D.: Feature Selection: Evaluation, Application, and Small Sample Performance. IEEE Trans. Pattern Analysis and Machine Intelligence 19(2), 153–157 (1997)
8. Koller, D., Sahami, M.: Toward optimal feature selection. In: Proc. 13th Int'l Conf. Machine Learning, pp. 284–292 (1996)
9. Raudys, S.J., Jain, A.K.: Small sample size effects in statistical pattern recognition: recommandations for practitioners. IEEE Trans. Pattern Analysis and Machine Intelligence 13(3), 252–264 (1991)
10. Law, M.H.C., Figueiredo, M.A.T., Jain, A.K.: Simultaneous Feature Selection and Clustering Using Mixture Models. IEEE Trans. on Pattern Analysis and Machine Intelligence 26(9), 1154–1166 (2004)
11. Li, Y., Dong, M., Hua, J.: Simultaneous Localied Feature Selection and Model Detection for Gaussian Mixtures. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(5), 953–960 (2009)
12. Xu, L.: Best harmony, unified RPCL and automated model selection for unsupervised and supervised learning on Gaussian mixtures, three-layer nets and ME-RBF-SVM models. International Journal of Neural Systems 11, 43–69 (2001)
13. Xu, L.: BYY harmony learning, structural RPCL, and topological self-organizing on mixture modes. Neural Networks 15, 1231–1237 (2002)
14. Dash, M., Choi, K., Scheuermann, P., Liu, H.: Feature Selection for Clustering - A Filter Solution. In: Second IEEE International Conference on Data Mining (ICDM 2002), p. 115 (2002)
15. Jouve, P.-E., Nicoloyannis, N.: A Filter Feature Selection Method for Clustering. In: Hacid, M.-S., Murray, N.V., Raś, Z.W., Tsumoto, S. (eds.) ISMIS 2005. LNCS (LNAI), vol. 3488, pp. 583–593. Springer, Heidelberg (2005)

16. Fowlkes, E.B., Gnanadesikan, R., Kettering, J.R.: Variable selection in clustering. Journal of Classification 5, 205–228 (1988)
17. Devaney, M., Ram, A.: Eficient feature selection in conceptual clustering. Machine Learning. In: Proceedings of the Fourteenth International Conference, Nashville, TN, pp. 92–97 (1997)
18. Tadesse, M.G., Sha, N., Vannucci, M.: Bayesian Variable Selection in Clustering High-Dimensional Data. Journal of the American Statistical Association 100, 602–617 (2005)
19. Kim, S., Tadesse, M.G., Vannucci, M.: Variable selection in clustering via Dirichlet process mixture models. Biometrika 93, 321–344 (2006)
20. Akaike, H.: A new look at statistical model identification. IEEE Transactions on Automatic Control AC-19, 716–723 (1974)
21. Scharz, G.: Estimating the dimension of a model. The Annals of Statistics 6, 461–464 (1978)
22. Rissanen, J.: Modeling by shortest data description. Automatica 14, 465–471 (1978)
23. Wallace, C., Dowe, D.: Minimum Message Length and Kolmogorov Complexity. Computer Journal 42(4), 270–283 (1999)
24. Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90(430), 577–588 (1995)
25. Recgardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Soceity B 59(4), 731–792 (1997)
26. Ueda, N., Ghahramani, Z.: Bayesian model search for mixture models based on optimizing variational bounds. Neural Networks 15(10), 1123–1241 (2002)
27. Constantinopoulos, C., Likas, A.: Unsupervised learning of Gaussian mixtures based on variational component splitting. IEEE Trans. on Neural Networks 18(3), 745–755 (2007)
28. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(3), 381–396 (2002)
29. Ma, J., Wang, T., Xu, L.: A gradient BYY harmony learning rule on Gaussian mixture with automated model selection. Neurocomputing 56, 481–487 (2004)
30. Ma, J., Wang, L.: BYY harmony learning on finite mixture: adaptive gradient implementation and a floating RPCL mechanism. Neural Processing Letters 24(1), 19–40 (2006)
31. Ma, J., Liu, J.: The BYY annealing learning algorithm for Gaussian mixture with automated model selection. Pattern Recognition 40, 2029–2037 (2007)
32. Ma, J., He, X.: A fast fixed-point BYY harmony learning algorithm on Gaussian mixture with automated model selection. Pattern Recognition Letters 29(6), 701–711 (2008)
33. Ma, J.: Automated model selection (AMS) on finite mixtures: a theoretical analysis. In: Proc. 2006 International Joint Conference on Neural Networks (IJCNN 2006), Vancouver, Canada, July 16-21, pp. 8255–8261 (2006)
34. Ma, J., Wang, T.: A cost-function approach to rival penalized Competitive learning (RPCL). IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics 36(4), 722–737 (2006)

Appendix

Define:

$$h_{tjl} = \frac{p(x_{tl}|\theta_{jl}) - q(x_{tl}|\lambda_l)}{\rho_l p(x_{tl}|\theta_{jl}) + (1-\rho_l) q(x_{tl}|\lambda_l)}.$$

The gradient and Hessian of $J(\Theta_k)$ with respect to ρ_l are:

$$\frac{\partial J(\theta)}{\partial \rho_l} = \frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\gamma_j(t) h_{tjl};$$

if $l \neq m$,

$$\frac{\partial^2 J(\theta)}{\partial \rho_l \partial \rho_m} = \frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\Big[h_{tjm}h_{tjl}\gamma_j(t) + h_{tjm}h_{tjl} - \sum_{i=1}^{k} p(i|x_t) h_{tim}h_{tjl}\gamma_j(t) - \sum_{i=1}^{k} p(i|x_t) h_{til}\gamma_j(t) h_{tjm}\Big].$$

if $l = m$,

$$\frac{\partial^2 J(\theta)}{\partial \rho_l^2} = \frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{k} p(j|x_t)\Big[h_{tkl} - 2\sum_{i=1}^{k} p(i|x_t) h_{til}\gamma_i(t)\Big] h_{til}.$$