# Harmonious Competition Learning for Gaussian Mixtures

GuoJun Liu and XiangLong Tang

Harbin Institute of Technology, China

**Abstract.** This paper proposes a novel automatic model selection algorithm for learning Gaussian mixtures. Unlike EM, we shall further increase the negative entropy of the posterior of latent variables to exert an indirect effect on model selection. The increase of negative entropy can be interpreted as a competition, which corresponds to an annihilation of those components with insufficient data to support. More importantly, this competition only depends on the data itself. Additionally, we seamlessly integrate parameter estimation and model selection into a single algorithm, which can be applied to any kind of parametric mixture model solved by an EM algorithm. Experiments involving Gaussian mixtures show the efficiency of our approach on model selection.

**Keywords:** Harmonious competition learning, Gaussian mixture model, Model selection, Expectation maximization.

## 1 Introduction

Gaussian mixtures as a flexible probabilistic modeling tool play an important role in many fields, such as machine learning, pattern recognition, bioinformatics, computer vision, signal and image analysis. Typically, Gaussian mixtures consists of $K$ components. Supposed that each observation has been produced by exactly one of $K$ components, to identify Gaussian mixtures, three levels of inference need to be solved, inferring which component produce each observation, i.e., inferring the parameters of each one of $K$ components, and inferring the number of components, i.e., the value of $K$. The former two lead to a clustering of the set of observations, the last one is an important issue, also known as model selection or model comparison, which assigns a preference to a set of alternative statistical models with differing complexities. However, until now, there is little agreement on what on earth the best approach of model selection is.

Technically, the underlying mixture model is often not the one that fits the data best due to over-fitting. When the number of components $K$ is fixed, maximum likelihood (ML) has proven to be an effective method of parameter estimation [1]. Nevertheless, if the value of $K$ itself also needs to be estimated, maximum likelihood tends to be greedy and results in those over-parameterized models.

In this several decades, a great number of model selection methods have been proposed to avoid over-fitting, these methods can be broadly divided into four categories.

C. Sun et al. (Eds.): IScIDE 2013, LNCS 8261, pp. 385–392, 2013.

First, some methods attempt to indirectly compensate for the loss of the upper relation $\mathcal{M}_k \to \boldsymbol{\theta}$ by the addition of a penalty term $\mathcal{P}(\mathcal{M}_k)$ to the best-fit loglikelihood $\log p(\mathbf{X}|\boldsymbol{\theta}_{\mathrm{ML}})$, such as cross-validation (CV) based criteria and the Akaike information criterion (AIC) [2]. Second, a constraint on the relation $\mathcal{M}_k \to \boldsymbol{\theta}$ is directly introduced by choosing a reasonable prior $p(\boldsymbol{\theta}|\mathcal{M}_k)$, the goal is to maximize $\log p(\mathbf{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{M}_k)$, which is called maximum a posteriori (MAP) estimation in Bayesian approach. Similarly, in devising two-part coding schemes, both minimum description length (MDL) and minimum message length (MML) employ different approaches of parameter truncation to such a Bayesian situation. Third, the primary aim is to maximize the log marginal likelihood $\log p(\mathbf{X}|\mathcal{M}_k) = \log \int p(\mathbf{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{M}_k)\, \mathrm{d}\boldsymbol{\theta}$ by integrating out nuisance parameters. Unfortunately, in many cases, this Bayesian integral is generally difficult to compute, therefore, we have to resort to approximation schemes, such as Laplace's method used in Bayesian information criterion (BIC) [3], and variational approximation employed in variational Bayes (VB) [4]. Last, the competitive learning methods [5,6] have attracted more and more attentions for the ability of simultaneously dealing with both the parameter estimation and model selection. A important feature is that it can automatically perform component annihilation [7] , that is to say, the too weak component unsupported by data is simply annihilated by an explicit or heuristic competitive learning rule. However, there are still some problems and limitations for above methods.

In this paper, a novel automatic model selection algorithm is proposed to learn Gaussian mixtures. Unlike EM, we shall further increase the negative entropy of the posterior of latent variables to exert an indirect effect on model selection. The increase of negative entropy is virtually a transition from disorder to order, and also be interpreted as a competition. More importantly, this competition only depends on the data itself.

The rest of paper is organized as follows: in Section 2, we derive the harmonious competition learning on the basis of EM. In Section 3, we give a more detailed solution of harmonious competition function as a constrained optimization problem as well as its important properties. Section 4 reports experimental results on model selection for Gaussian mixtures and Section 5 ends the paper by presenting some concluding remarks.

## 2 Derivation of Harmonious Competition Learning Based on EM

In this section, we derive the harmonious competition learning on the basis of EM [8]. The EM algorithm [9] is an elegant and powerful technique to find maximum likelihood solutions for probabilistic models with latent nuisance variables $\mathbf{Z}$, it is an iterative optimization method to estimate some unknown parameters $\boldsymbol{\theta}$, in the light of the observed variables $\mathbf{X}$. The goal is to maximize the posterior probability of the parameters $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{Z}} p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{X})$. Equivalently, we can maximize the logarithm of the joint distribution which is proportional to the posterior:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \tag{1}$$

However, maximizing Eq. (1) inevitably involves the logarithm of a sum, which is difficult to deal with. Fortunately, by the Jensen's inequality, we can construct a tractable lower bound $\mathrm{B}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{old}}) \triangleq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z},\boldsymbol{\theta})}{q(\mathbf{Z})}$ in order to simply transform the log of a sum into a sum of logs $\mathrm{B}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{old}}) \leq \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z},\boldsymbol{\theta})}{q(\mathbf{Z})}$ where $q(\mathbf{Z})$ is an arbitrary probability distribution over the space of latent variables $\mathbf{Z}$.

In E-step, the optimal bound at a guess $\boldsymbol{\theta}^{\mathrm{old}}$ can be obtained by maximizing $\mathrm{B}(\boldsymbol{\theta}^{\mathrm{old}}; \boldsymbol{\theta}^{\mathrm{old}})$ with respect to the distribution $q(\mathbf{Z})$. Meanwhile, introducing a Lagrange multiplier $\lambda$ to enforce the constraint $\sum_{\mathbf{Z}} q(\mathbf{Z}) = 1$, so we obtain

$$q(\mathbf{Z}) = \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}^{\mathrm{old}})}{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}^{\mathrm{old}})} = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \tag{2}$$

Subsequently, in M-step, we require to maximize $\mathrm{B}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{old}})$ with respect to $\boldsymbol{\theta}$, and rewrite it as

$$\begin{aligned} \mathrm{B}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{old}}) &\triangleq \mathbb{E}_q \left[ \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \right] + \mathrm{H}_q \\ &= \mathbb{E}_q \left[ \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right] + \log p(\boldsymbol{\theta}) + \mathrm{H}_q \end{aligned} \tag{3}$$

where $\mathbb{E}_q[\cdot]$ denotes the expectation with respect to the distribution of $q(\mathbf{Z})$, $p(\boldsymbol{\theta})$ is the prior of the parameters $\boldsymbol{\theta}$, and $\mathrm{H}_q$ is the entropy of the distribution of $q(\mathbf{Z})$.

Generally, after the E-step, EM algorithm would fix $q(\mathbf{Z})$ at the value of $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})$ as Eq. (2), thereby the entropy $\mathrm{H}_q$ does not depend on $\boldsymbol{\theta}$. Maximizing the bound $\mathrm{B}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{old}})$ with respect to $\boldsymbol{\theta}$ is up to the first two terms only:

$$\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} \{ \mathcal{Q}(q, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \} \tag{4}$$

In particular, we must pay more attention to the first term $\mathbb{E}_q \left[ \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right]$ in Eq. (3) rewritten as $\mathcal{Q}(q, \boldsymbol{\theta})$ by us, instead of $\mathcal{Q}(\boldsymbol{\theta})$ as the convention of EM. Note that $\mathcal{Q}(q, \boldsymbol{\theta})$ is not only a function of the parameters $\boldsymbol{\theta}$, but also a functional of the distribution $q(\mathbf{Z})$, which means that we can further tune the $q(\mathbf{Z})$ on the basis of the fixed value $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})$ in order to increase $\mathcal{Q}(q, \boldsymbol{\theta})$ before the change of the parameters $\boldsymbol{\theta}$ in M-step.

Therefore, a plug-in step, called harmonious competition step or C-step, is able to be inserted between E-step and M-step. In this step, the parameters $\boldsymbol{\theta}$ is still kept at the fixed value $\boldsymbol{\theta}^{\mathrm{old}}$, then we have

$$\begin{aligned} \mathcal{Q}\left(q, \boldsymbol{\theta}^{\mathrm{old}}\right) &= \mathbb{E}_q \left[ \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{\mathrm{old}}) \right] + \log p(\boldsymbol{\theta}^{\mathrm{old}}) \\ &= \mathbb{E}_q \left[ \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \right] + \log p(\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \end{aligned} \tag{5}$$

Increasing $\mathcal{Q}(q, \boldsymbol{\theta}^{\mathrm{old}})$ with respect to $q(\mathbf{Z})$ only depends on the first term in Eq. (5), this is equivalent to further increase the negative entropy of $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})$. It leads to

a new distribution $\hat{q}(\mathbf{Z}) = \mathcal{C}(p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}))$ to ensure that $\mathcal{Q}(\hat{q}, \boldsymbol{\theta}^{\text{old}}) \geq \mathcal{Q}(q, \boldsymbol{\theta}^{\text{old}})$, that is to say,

$$\mathbb{E}_{\hat{q}} \left[ \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \right] \geq \mathbb{E}_{q} \left[ \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \right] \tag{6}$$

where $\mathcal{C}(\cdot)$ denotes harmonious competition function which is considered as a constrained optimization detailed in Section 3.3.

Last but not least, after C-step, $\hat{q}$ as a new responsibility has taken the place of the old one $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ in M-step, then update and get new parameters $\boldsymbol{\theta}^{\text{new}}$. Note that the mixture weight as a subset of $\boldsymbol{\theta}^{\text{new}}$ has been updated by using $\hat{q}$ instead of $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$, in other words, the result of harmonious competition has produced an effect on the mixture weight. More importantly, the increase of negative entropy is able to force the mixture weight of some components to tend to 0. Subsequently, we require another step called component annihilation. In this step, annihilate one or more components whose mixture weight is less than the predefined threshold $\epsilon$, then remove the corresponding parameters from the set of parameters and normalize the mixture weights once more, at the same time, update $K$ to be the number of the survived components. i.e, automatic model selection.

## 3 Harmonious Competition Learning

### 3.1 The Relation of Negative Entropy and Competition

Negative entropy is viewed as a mathematical synonym for *order* in an entropic sense, this term comes from Nobel laureate Erwin Schrödinger's famous booklet *What is life?*.

For a probability distribution, with the increase of negative entropy, it will transition gradually from disorder or chaos to order. Geometrically, it can be interpreted as a collapse from a high-dimensional space to a lower dimensional subspace. For example, suppose that a change of a probability distribution with 3 elements likes that $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \rightarrow (\frac{1}{2}, \frac{1}{2}, 0) \rightarrow (1, 0, 0)$, the corresponding change of negative entropy is $-\log 3 < -\log 2 < 0$. Once the value of some element is equal to 0, in geometry, it means that the corresponding spatial dimension plays no role in describing the current probability distribution. Consequently, it forms a collapse into a subspace.

### 3.2 The Probability Simplex

Simplex is an important family of polyhedra. Specifically, a $(n-1)$-dimensional simplex is the convex hull of its $n$ vertices, e.g., a 0-dimensional simplex is a single point, a 1-dimensional simplex is a line segment, a 2-dimensional simplex is a triangle, and a 3-dimensional simplex is a tetrahedron.

### 3.3 Harmonious Competition Function

The negative entropy of a discrete probability distribution is defined by $\hbar(\mathbf{x}) = \sum_{i=1}^{n} x_i \log x_i$, where $\mathbf{x} \in \mathbb{S}^n$. To increase negative entropy $\hbar(\mathbf{x})$, the gradient based method is available to get $\mathbf{g}$ as following

$$g_i = x_i + \eta \nabla_{x_i} \hbar = x_i + \eta \left(1 + \log x_i\right) \tag{7}$$

where $\eta > 0$, $\eta$ is a learning rate and also called a competition intensity. Note that the vector $\mathbf{g}$ may be not in the probability simplex any more, i.e., $\mathbf{g} \notin \mathbb{S}^n$. Therefore, to turn it into a probability distribution, we just need to project it onto the probability simplex and find its corresponding projection $\mathbf{y} \in \mathbb{S}^n$. This is equivalent to solve a convex optimization problem, we consider it in the standard form.

$$\begin{aligned} &\text{minimize} && f_0(\mathbf{y}) = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - g_i\right)^2 \\ &\text{subject to} && \mathbf{y} \succeq \mathbf{0},\ \mathbf{1}^{\mathsf{T}} \mathbf{y} = 1 \end{aligned} \tag{8}$$

Introducing Lagrange multipliers $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ for the inequality constraints $\mathbf{y}^* \succeq \mathbf{0}$ and a multiplier $\nu^* \in \mathbb{R}$ for the equality constraint $\mathbf{1}^{\mathsf{T}} \mathbf{y} = 1$, We define the Lagrangian $\mathcal{L}$ associated with the problem as

$$\mathcal{L}(\mathbf{y}, \boldsymbol{\lambda}^*, \nu^*) = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - g_i\right)^2 + \sum_{i=1}^{n} \lambda_i^* \left(-y_i^*\right) - \sum_{i=1}^{n} \nu^* \, y_i^* \tag{9}$$

These above equations satisfy the KKT conditions and can be solved directly to find $\mathbf{y}^*$, $\boldsymbol{\lambda}^*$, and $\nu^*$. Thus we have

$$y_i^* = \begin{cases} \nu^* + g_i & \nu^* > -g_i \\ 0 & \nu^* \leq -g_i \end{cases} \tag{10}$$

or, put more simply, $y_i^* = \max\{0, \nu^* + g_i\}$. Substituting it into the second condition $\mathbf{1}^{\mathsf{T}} \mathbf{y}^* = 1$, we obtain

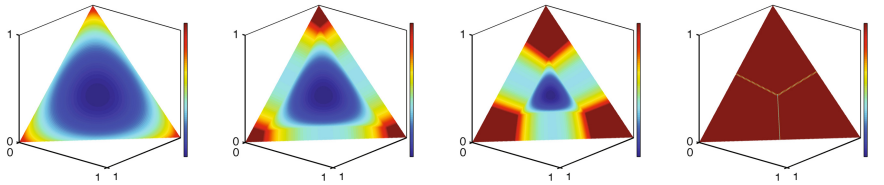
$$\sum_{i=1}^{n} \max\left\{0, \nu^* + g_i\right\} = 1 \tag{11}$$

This solution method is called water-filling. The left-hand side is a piecewise-linear increasing function of $\nu^*$, with breakpoints at $-g_i$. Therefore, it is solvable and has a unique solution.

For convenience, we shall give a definition of the above method and call it harmonious competition function.

**Definition 1 (Harmonious competition function).** *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be two* $n$*-dimensional vectors in the probability simplex, i.e.,* $\mathbf{x}, \mathbf{y} \in \mathbb{S}^n$*, the harmonious competition function* $\mathcal{C} : \mathbf{x} \mapsto \mathbf{y}$ *is defined by*

$$y_i = \mathcal{C}(x_i) = \max\left\{0, x_i + \triangle_i + v\right\},\ i \in \{1, \ldots, n\} \tag{12}$$



(a) $\eta = 0$ (b) $\eta = 0.1$ (c) $\eta = 0.5$ (d) $\eta = 100$

**Fig. 1.** The change of negative entropy before and after the harmonious competition function. The value of negative entropy ranges from $-\log 3$ to 0, correspondingly, it is described by the jet colormap which ranges from blue to red, and passes through the colors cyan, yellow, and orange.

*where* $\triangle \overset{\text{def}}{=} f(\mathbf{x})$*, $f$ is a monotonically increasing function and* $\triangle \in \mathbb{R}^n$*. $v$ is a single variable and chosen such that* $\sum_{i=1}^{n} \max\{0, x_i + \triangle_i + v\} = 1$*.*

Subsequently, we shall give a quantitative analysis of how to increase negative entropy with the different value of competition intensity, as illustrated in Fig. 1. Suppose that $\mathbf{x} \in \mathbb{S}^3$, then calculate $\hbar(\mathbf{x})$, as shown in Fig. 1(a), where $\eta = 0$ such that $\hbar(\mathbf{y}) = \hbar(\mathbf{x})$ by Eq. (7). Let $\mathbf{y} = \mathcal{C}(\mathbf{x})$ for every $\mathbf{x}$, and redraw $\hbar(\mathbf{y})$ onto the same probability simplex with different $\eta$, as illustrated in Fig. 1(b)-(d). Here, $\eta$ governs the intensity of the competition. For example, in Fig. 1(d), for almost all vectors in the domain, the value of negative entropy is approximately equal to 0, it means that an competition is so intense to make the probability of one element equal to 1 and the probability of the other two elements equal to 0. In other words, the competition reassigns the probability of each element. When $\eta$ tends to infinity, harmonious competition will degenerate to a winner-take-all manner of $K$-means.

## 4 Experiments

Although the proposed method can be applied for any kind of mixture model, our experiments focus only on Gaussian mixtures, which are by far the most common model. To compare our algorithm (HCL) with those traditional methods referred in Section 1, we chose AIC, BIC and VB, as the most commonly used model selection criterion. For all experiments, we set the candidate model of AIC and BIC range from $K_{min} = 1$ to $K_{max} = 10$, where $K$ denotes the number of mixture components. In addition, the Dirichlet distribution as the conjugate prior of multinomial distribution is often used in VB, the value of its hyperparameter acts as a prior knowledge and has an important effect on model selection, therefore, let it equal to a small value, i.e., an uninformative prior. Last, we set the initial number of mixture components $K_{init}$ be large enough, e.g., $K_{init} = 15$ in VB and our method.

There are two groups of experiments, for every data set, all methods run 100 times respectively, then compare the results with the true number of mixtures components, and obtain the percentage of success of various methods.

**Table 1.** Percentage of success of various methods of two real data sets

| DATA SET | AIC | BIC | VB | HCL |
|---|---|---|---|---|
| Old faithful | 0 | 100 | 93 | 100 |
| Iris data | 0 | 2 | 33 | 65 |

In the first example, we consider the two well-known real data sets, one is Old Faithful, a 272 2-dimensional bimodal data set, the other is Iris data set, 150 4-dimensional points from three classes, 50 per class. The results are shown in Table 1. For a large sample low-dimensional data set, BIC, VB and our method have a good performance. Notwithstanding a sharp decline in the correct model selections with the dimensional increase and the decrease of sample number, our approach is still superior to other methods.

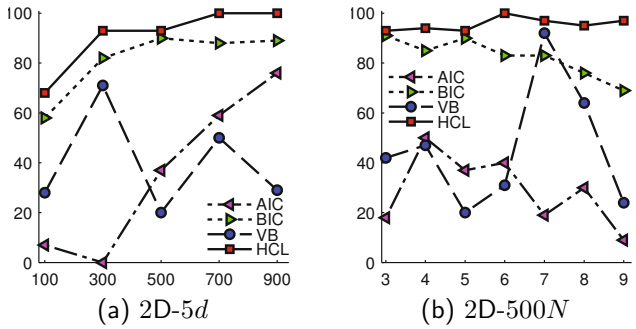



**Fig. 2.** Percentage of success of various methods using 2-dimensional synthetic data with different $N$ and $d$ respectively

Second, we use $N$ samples from a 5-component bivariate mixture, the mixture weight of each component is equal to $1/5$, mean vectors at $[0, 0]^{\mathsf{T}}$, $[0, d]^{\mathsf{T}}$, $[0, -d]^{\mathsf{T}}$, $[d, 0]^{\mathsf{T}}$, $[-d, 0]^{\mathsf{T}}$ where $d$ denotes the distance from the origin, and equal covariance matrices $diag\{2, 0.2\}$. In Fig. 2(a), we fix $d = 5$ and draw different number of samples from above Gaussian mixtures, $N$ ranges from 100 to 900. Next, we fix $N = 500$, then use different $d$ to generate samples.

As illustrated in Fig. 2, the performance of those traditional methods is unstable, worse for most cases and better only for some special cases which seem suitable for necessary approximation conditions, such as AIC, BIC. As for VB, it naturally embodies many features of Bayesian inference, so it automatically makes the trade-off between fitting the data and model complexity, but most importantly, that which one is more comprise-inclined in practice is not clear. In contrast, the harmonious competition only depends on data itself, instead of some approximation techniques or heuristic rules, therefore, our method is more robust.

## 5 Conclusion

This paper proposes a novel automatic model selection algorithm for learning Gaussian mixtures. The novelty in our approach is that harmonious competition is able to make the mixture weight of those components with insufficient data to support tend to zero, more importantly, it only depends on the data itself. Furthermore, we seamlessly integrate parameter estimation and model selection into a single algorithm, which can be applied to any kind of parametric mixture model solved by an EM algorithm. Experiments involving Gaussian mixtures show the efficiency of our approach on model selection.

**Acknowledgments.** This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. HIT.NSRIF.2014069) and National Natural Science Foundation of China (Grant No. 61173087).

## References

1. Lanterman, A.D.: Schwarz, wallace, and rissanen: Intertwining themes in theories of model selection. International Statistical Review 69(2), 185–212 (2001)
2. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (1974)
3. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6(2), 461–464 (1978)
4. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Machine Learning 37(2), 183–233 (1999)
5. Xu, L., Krzyzak, A., Oja, E.: Rival penalized competitive learning for clustering analysis, RBF net, and curve detection. IEEE Transactions on Neural Networks 4(4), 636–649 (1993)
6. Xu, L.: Bayesian Ying-Yang system, best harmony learning, and five action circling. Frontiers of Electrical and Electronic Engineering in China 5(3), 281–328 (2010)
7. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 381–396 (2002)
8. Neal, R.M., Hinton, G.E.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan, M.I. (ed.) Learning in Graphical Models, 1st edn., pp. 355–368. MIT Press, Cambridge (1998)
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39(1), 1–38 (1977)