Cluster Number Selection for a Small Set of Samples Using the Bayesian Ying–Yang Model

Ping Guo, C. L. Philip Chen, and Michael R. Lyu

***Abstract*—One major problem in cluster analysis is the determination of the number of clusters. In this paper, we describe both theoretical and experimental results in determining the cluster number for a small set of samples using the Bayesian–Kullback Ying–Yang (BYY) model selection criterion. Under the second-order approximation, we derive a new equation for estimating the smoothing parameter in the cost function. Finally, we propose a gradient descent smoothing parameter estimation approach that avoids complicated integration procedure and gives the same optimal result.**

***Index Terms*—Bootstrap, cluster number selection, data smoothing, SEM algorithm, small number sample set, smoothing parameter estimation.**

I. Introduction

IN INTELLIGENT statistical data analysis or unsupervised classification, cluster analysis is to determine the cluster number or cluster membership of a set of given samples, $\{\mathbf{x}_i\}_{i=1}^N$ [1], [2], [3], [27], by its mean vector, $\{\mathbf{m}_y\}_{y=1}^k$. In most cases, the first step of the clustering is to determine the cluster number. The second step is to design a proper clustering algorithm. In recent years, several clustering analysis algorithms have been developed to partition samples into several clusters, in which the number of clusters is *predetermined*. The most notable approaches are, for example, the mean square error (MSE) clustering and finite mixture model algorithms.

The MSE clustering algorithm typically is implemented by the well-known k-mean algorithm [1], [27]. This method requires specifying the number of clusters, k, in advance. If k is correctly selected, then it can produce a good clustering result; otherwise, data sets cannot be grouped into appropriate clusters. However, in most cases the number of clusters is unknown in advance. Because it is difficult to select appropriate number of clusters, some heuristic approaches have been used to tackle this problem. The rival penalized competitive learning (RPCL) [4] algorithm has demonstrated a very good result in finding the cluster number. However, there is still no appropriate theory being developed [5], [6].

Manuscript received September 20, 2000; revised October 10, 2001. This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administration Region (Project CUHK 4222/01E).

P. Guo is with the Department of Computer Science, Beijing Normal University, Beijing, 100875, P.R. China, and the Department of Computer Science and Engineering, Chinese University of Hong Kong Shatin, NT, Hong Kong, P.R. China.

C. L. P. Chen is with the Department of Computer Science and Engineering, Wright State University Dayton, OH 45435 USA (e-mail: pchen@cs.wright.edu).

M. R. Lyu is with the Department of Computer Science and Engineering, Chinese University of Hong Kong Shatin, NT, Hong Kong, P.R. China.

Publisher Item Identifier S 1045-9227(02)04445-4.

In the mixture model cluster analysis, the sample data are viewed as two or more mixtures of normal (Gaussian) distribution in varying proportion. The cluster is analyzed by means of mixture distribution. The likelihood approach to the fitting of mixture models has been utilized extensively [7]–[11]. However, the determination of the appropriate cluster number still remains one of the most difficult problems in cluster analysis [12].

The Bayesian–Kullback Ying–Yang (BYY) learning theory has been proposed in [13]. The BYY learning is a unified algorithm for both unsupervised and supervised learning which provides us a reference for solving the problem of selecting cluster number. The experimental results worked very well for a large set of samples when the smoothing parameter $h \to 0$ [14], [15]. However, for a relatively small set of samples, the maximum likelihood (ML) method with the expectation-maximization (EM) algorithm [16] for estimating mixture model parameters will not adequately reflect the characteristics of the cluster structure. In this way, the selected cluster number is incorrect. To solve the problem for the small set of samples, the BYY theory for data smoothing is developed in [17] is approach considers the nonparametric density estimation and the smoothing factor in the Parzen window.

In this paper, we investigate the problem of determining the smoothing parameter and the model selection in clustering. With this approach, the performance of the BYY model selection criterion for determining cluster number is greatly improved. Finally, we propose an efficient gradient descent smoothing parameter estimation approach that not only reduces the complicated computation procedure but also gives the optimal result.

II. Preliminary

First, we briefly review the finite mixture model and the BYY theory for model selection [14], [19].

A. The Finite Mixture Model

Let us consider a Gaussian mixture model. The joint probability density that consists of k Gaussians is

$$p(\mathbf{x}, \Theta) = \sum_{y=1}^{k} \alpha_y G(\mathbf{x}, \mathbf{m}_y, \mathbf{\Sigma}_y)$$

with

$$\alpha_y \geq 0, \quad \text{and} \quad \sum_{y=1}^{k} \alpha_y = 1 \tag{1}$$

where

$$G(\mathbf{x}, \mathbf{m}_y, \mathbf{\Sigma}_y) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_y)^T\mathbf{\Sigma}_y^{-1}(\mathbf{x}-\mathbf{m}_y)\right]}{(2\pi)^{d/2}|\mathbf{\Sigma}_y|^{1/2}} \quad (2)$$

is a general multivariate Gaussian density function, $\mathbf{x}$ denotes a random vector, d is the dimension of $\mathbf{x}$, and parameter $\Theta \equiv \{\alpha_y, \mathbf{m}_y, \mathbf{\Sigma}_y\}_{y=1}^k$ is a set of finite mixture model parameter vectors. Here, α_y is the *prior* probability, $\mathbf{m}_y$ is the mean vector, and $\mathbf{\Sigma}_y$ is the covariance matrix of the yth component. Based on a given data set $D = \{\mathbf{x}_i\}_{i=1}^N$, these parameters can be estimated by maximum likelihood learning with the EM algorithm [7], [16].

B. The BYY Theory for Finite Mixture Model and EM Algorithm

As mentioned in [19], [20], unsupervised learning problems can be summarized into the problem of estimating joint distribution $P(\mathbf{x}, \mathbf{y})$ of patterns in the input space $\mathbf{X}$ and the representation space $\mathbf{Y}$. By the Bayesian Kullback Ying–Yang theory, we have the following Kullback–Leibler divergence [19]:

$$KL(M_1, M_2) = \iint P_{M_1}(y|\mathbf{x})P_{M_1}(\mathbf{x}) \cdot \ln \frac{P_{M_1}(y|\mathbf{x})P_{M_1}(\mathbf{x})}{P_{M_2}(\mathbf{x}|y)P_{M_2}(y)}\, d\mathbf{x}\, dy. \quad (3)$$

where M_1 and M_2 are two different models.

The minimization of $KL(M_1, M_2)$ can be implemented by the alternative minimization procedure which alternatively minimizes one model while keeping other models temporarily fixed.

We can obtain a general form of KL function in the Gaussian mixture model case as

$$KL(M_1, M_2) = \iint P(y|\mathbf{x})p_{h_x}(\mathbf{x}) \cdot \ln \frac{P(y|\mathbf{x})p_{h_x}(\mathbf{x})}{\alpha_y G(\mathbf{x}, \mathbf{m}_y, \mathbf{\Sigma}_y)}\, d\mathbf{x}\, dy \quad (4)$$

where p_{h_x} is a nonparametric kernel estimation.

With

$$P(y|\mathbf{x}) = \frac{\alpha_y G(\mathbf{x}, \mathbf{m}_y, \mathbf{\Sigma}_y)}{p_{M_2}(\mathbf{x}, \Theta)}$$

$$p_{M_2}(\mathbf{x}, \Theta) = \sum_{y=1}^k \alpha_y G(\mathbf{x}, \mathbf{m}_y, \mathbf{\Sigma}_y) \quad (5)$$

the KL function becomes

$$KL(k, h, \Theta) = -\int p_{h_x}(\mathbf{x}) \ln p_{M_2}(\mathbf{x}, \Theta)\, d\mathbf{x} + \int p_{h_x}(\mathbf{x}) \ln p_{h_x}(\mathbf{x})\, d\mathbf{x}. \quad (6)$$

For a mixture model parameter learning

$$\Theta = \arg\min_{\Theta} KL(\Theta), \qquad KL(\Theta) = KL(k, h, \Theta). \quad (7)$$

If KL function is minimized with respect to parameter Θ, the EM algorithm [7], [16] can be rederived within the limit of $h \to 0$. The following is the EM algorithm which breaks down into E-step and M-step.

E-Step: Calculate the *posterior* probability $P(y|\mathbf{x}_i)$

$$P(y|\mathbf{x}_i) = \frac{\alpha_y G(\mathbf{x}_i, \mathbf{m}_y, \mathbf{\Sigma}_y)}{\sum_{y=1}^k \alpha_y G(\mathbf{x}_i, \mathbf{m}_y, \mathbf{\Sigma}_y)}. \quad (8)$$

M-Step:

$$\alpha_y^{new} = \frac{1}{N}\sum_{i=1}^N \frac{\alpha_y^{old} G(\mathbf{x}_i, \mathbf{m}_y, \mathbf{\Sigma}_y)}{\sum_{y=1}^k \alpha_y^{old} G(\mathbf{x}_i, \mathbf{m}_y, \mathbf{\Sigma}_y)} = \frac{1}{N}\sum_{i=1}^N P(y|\mathbf{x}_i) \quad (9)$$

$$\mathbf{m}_y = \frac{\sum_{i=1}^N P(y|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^N P(y|\mathbf{x}_i)} = \frac{1}{\alpha_y N}\sum_{i=1}^N P(y|\mathbf{x}_i)\mathbf{x}_i \quad (10)$$

$$\hat{\mathbf{\Sigma}}_y = \frac{1}{\alpha_y N}\sum_{i=1}^N P(y|\mathbf{x}_i)[(\mathbf{x}_i - \mathbf{m}_y)(\mathbf{x}_i - \mathbf{m}_y)^T]. \quad (11)$$

A local minima can be found by iterating these two steps.

C. Model Selection Criterion

The determination of an appropriate number of clusters in a data set is one of the most difficult problems in clustering analysis [12]. In the literature, there are several heuristically proposed information theoretical criteria. Following Akaike's pioneering work [21] in which an information criterion was first proposed for use in selecting the number of clusters in the mixture model cluster analysis. Similar studies include AICB [22], CAIC [23], and SIC [24]. These criteria combine the maximum value of the likelihood with the number of parameters.

The cluster number, k, is actually a structural scale parameter of the BYY system. From the BYY system, the BYY model selection criterion for determining the correct cluster number is derived in [14] as follows:

$$k = \arg\min_k J(k) \quad (12)$$

$$J(k) = \gamma_r H_1(k) + J_2^g(k) \quad (13)$$

$$0 \le \gamma_r \le 1 \quad (14)$$

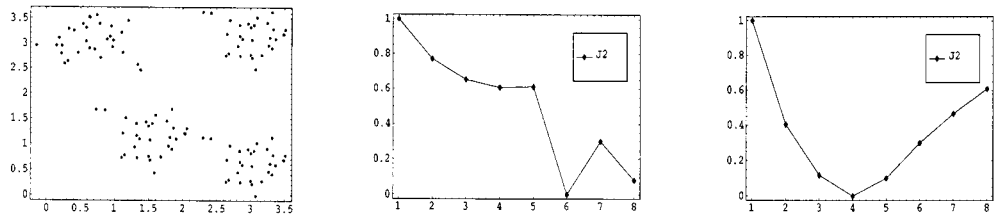
where

$$J_2^g(k) = \sum_{y=1}^k \alpha_y \ln\sqrt{|\Sigma_y|} - \sum_{y=1}^k \alpha_y \ln \alpha_y \quad (15)$$

$$H_1(k) = \frac{1}{N}\sum_{i=1}^N \sum_{y=1}^k P(y|\mathbf{x}_i) \ln P(y|\mathbf{x}_i). \quad (16)$$

In practice, we start with $k = 1$, estimate the parameter Θ by the EM algorithm based on the given samples, and compute $J(k)$. Then, we proceed to $k \to k+1$, and compute $J(k)$ again. We continue this process after we gather a series of $J(k)$. The appropriate cluster number, k, is selected from the one with minimal $J(k)$.

Although the model selection approach discussed above works well for a good size of data samples, we found out, from several experimental results, that the selected cluster number was not correct for a relatively small set of samples. The results



(a) (b) (c)

Fig. 1. Data smoothing result: The 2-D synthetic data set and the comparison of J versus k. (a) Data set. (b) The result of "without data smoothing" approach. (c) The result of "with data smoothing" approach. The results show that four clusters is the best number.

are also incorrect with other theoretical information criteria mentioned above. The reason is that the MLE with the EM algorithm that estimates mixture model parameters will not reflect the characteristics of cluster structures adequately. As a result, it affects the correctness of determining the cluster number. In order to study the effect of parameter estimation on the BYY model selection, we have incorporated the bootstrap technique with the EM algorithm in the MLE of mixture parameters and obtained a relatively robust performance for determining the cluster number with the BYY criterion and clustering for small set of samples [18].

In the next section, we investigate the BYY data smoothing theory for parameter estimation.

III. BYY Data Smoothing Theory

Under the conditional mean field approximation, minimizing KL function corresponding parameter Θ will lead to the smoothing EM (SEM) algorithm [20], where the updates in E-step and M-step are given as follows.

E-Step:

$$P(y|\mathbf{x}_i) = \frac{\alpha_y G(\mathbf{x}_i, \mathbf{m}_y, \mathbf{\Sigma}_y)}{\sum_{y=1}^{k} \alpha_y G(\mathbf{x}_i, \mathbf{m}_y, \mathbf{\Sigma}_y)}. \tag{17}$$

M-Step:

$$\alpha_y^{new} = \frac{1}{N} \sum_{i=1}^{N} \frac{\alpha_y^{old} G(\mathbf{x}_i, \mathbf{m}_y, \mathbf{\Sigma}_y)}{\sum_{y=1}^{k} \alpha_y^{old} G(\mathbf{x}_i, \mathbf{m}_y, \mathbf{\Sigma}_y)} \tag{18}$$

$$\mathbf{m}_y^{new} = \frac{\sum_{i=1}^{N} P(y|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{N} P(y|\mathbf{x}_i)} = \frac{1}{\alpha_y N} \sum_{i=1}^{N} P(y|\mathbf{x}_i)\mathbf{x}_i \tag{19}$$

$$\hat{\mathbf{\Sigma}}_y^{new} = h^2 \mathbf{I}_d + \frac{1}{\alpha_y N} \sum_{i=1}^{N} P(y|\mathbf{x}_i)[(\mathbf{x}_i - \mathbf{m}_y)(\mathbf{x}_i - \mathbf{m}_y)^T] \tag{20}$$

where $\mathbf{I}_d$ is a $d \times d$ dimensional identity matrix. The SEM algorithm is different from the ordinary EM algorithm in that it employs covariance estimation correction.

According to the principle of minimizing KL function, when $h \neq 0$, the smoothing parameter h should be estimated as

$$h = \arg\min J(h), \qquad J(h) = KL(k^*, \Theta^*, h). \tag{21}$$

IV. Practical Implementation Consideration

The BYY data smoothing is a quite new technique. Two aspects for implementing BYY data smoothing should be discussed. One aspect is that we need to verify if the estimated parameter for determining the cluster number with data smoothing. The other aspect is the selection of a proper smoothing parameter to estimate the mixture parameter.

Without loss of generality, we use a heuristic estimation of smoothing parameter h for fast implementation. For example, we can use $1/N$ of average distance approximation to estimate h value as follows:

$$h^2 = \frac{1}{dN^3} \sum_{i=1}^{N} \sum_{j=1}^{N} ||\mathbf{x}_j - \mathbf{x}_i||^2. \tag{22}$$

A. Data Smoothing Experiments

In order to investigate the data smoothing effect, we first use some synthetic data sets to conduct the experiments.

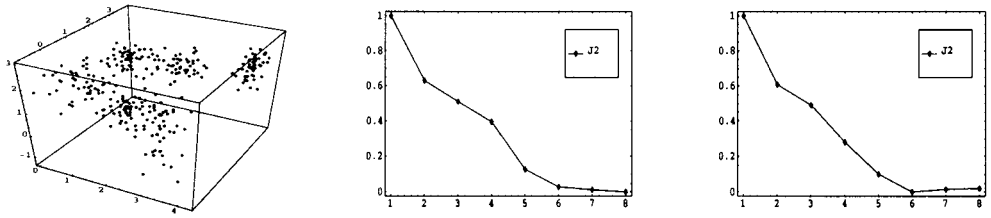
The data sets have been generated under different conditions, such as different Gaussian mixtures, different mean $\mathbf{m}_y$, and different covariance $\mathbf{\Sigma}_y$ of each cluster.

In computer experiments, we randomly generate $30 \times k$ two-dimensional (2-D) samples and $50 \times k$ three-dimensional (3-D) samples, where k is the number of Gaussian mixtures, varying from one to eight. Three data sets and their experimental results are shown in Figs. 1–3.

The cluster number selection criterion is when the cost function $J(k, \Theta)$ versus k reaches its global minimum point at $k = k^*$, where k is the candidate cluster number and k^* is the actual number of Gaussians in the finite Gaussian mixture model.

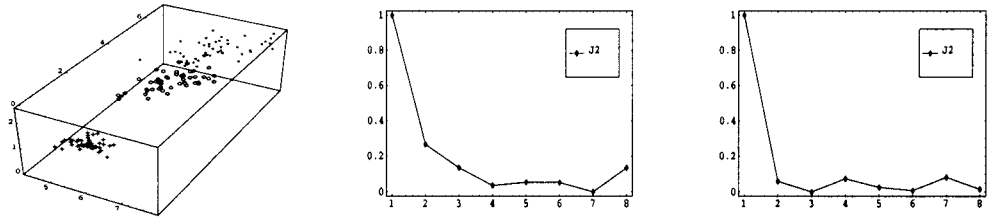
Fig. 1 shows the experimental result of the cost function $J_2(k)$ versus k for two dimensional Gaussian mixture data set. From Fig. 1(b), we find that the ordinary EM algorithm over-estimates the actual cluster number (which gives us six clusters), while the data smoothing SEM algorithm gives a reasonable result. In Fig. 1(c) the best cluster number is four from the J_2 versus k plot. Similarly in Fig. 2(c), the best cluster number is six, while the result of the ordinary EM algorithm shown in Fig. 2(b) is eight. As for large samples case, the experiments show no obvious difference between $h = 0$ and $h \neq 0$ in search of the correct cluster numbers [15].

Another example is the Iris plant dataset [26]. Fig. 3 depicts the results of the Iris dataset. The experimental results show that the correct cluster number is three from Fig. 3(c). We see that with data smoothing, the performance of cluster number selection is improved.



(a) (b) (c)

Fig. 2. Data smoothing result: The 3-D synthetic data set and the comparison of J versus k. (a) Data set. (b) The curve "without data smoothing." (c) The curve "with data smoothing." The results show that six clusters is the best number.



(a) (b) (c)

Fig. 3. The IRIS data set and the comparison of J versus k. (a) IRIS data set in x_1, x_3, x_4 axis view. (b) The curve "without data smoothing." (c) The curve "with data smoothing." The results show that three clusters is the best number.

B. Smoothing Parameter Estimation

According to the principle of minimizing KL function, the optimal smoothing parameter can be obtained from (21). However, the evaluation of integration is computation-expensive. Therefore, we propose an approximation scheme in order to avoid the integration.

In the following, we first review the quantized method which is recommended in [17]; then we derive a new gradient descent approximation for estimating this smoothing parameter h.

The Quantized Method: On each of the quantized levels h_r, $r = 1, 2, \ldots, n_h$, we run the SEM algorithm to obtain a series of mixture parameter Θ^*. We then choose one h_r such that its corresponding value of $KL(\Theta^*, k, h_r)$ is the smallest. This approach is an exhaustive search method and usually is computation-expensive. A gradient descent approach is proposed next.

Gradient Descent Approach: For the gradient descent approach, we need to find an approximation for estimating parameter h. Referring to [17], the smoothing parameter is given as

$$h_x^2 = \frac{1}{d_x N N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \beta_i(\mathbf{x}'_j) ||\mathbf{x}'_j - \mathbf{x}_i||^2 |_{h_x^{old}} \quad \text{[17, eq. (14b)]}$$

where

$$\beta_i(\mathbf{x}) = \frac{G(\mathbf{x}, \mathbf{x}_i, h_x^2 \mathbf{I}_{d_x})}{\sum_{i=1}^{N} G(\mathbf{x}, \mathbf{x}_i, h_x^2 \mathbf{I}_{d_x})}.$$

Note $G(\mathbf{x}, \mathbf{x}_i, h_x^2 \mathbf{I}_{d_x})$ is a Gaussian density function.

Now, let us denote

$$I_1 = \frac{1}{d_x N} \sum_{i=1}^{N} \int \beta_i(\mathbf{x}) ||\mathbf{x} - \mathbf{x}_i||^2 \, d\mathbf{x}$$

$$I_2 = \frac{1}{d_x N} \sum_{i=1}^{N} \int G(\mathbf{x}, \mathbf{x}_i, h_x^2 \mathbf{I}_{d_x}) ||\mathbf{x} - \mathbf{x}_i||^2 \, d\mathbf{x}.$$

Integrate I_2, we get $I_2 = h_x^2$.

Because $\beta_i(\mathbf{x})$ is positive and $\beta_i(\mathbf{x}) \leq G(\mathbf{x}, \mathbf{x}_i, h_x^2 \mathbf{I}_{d_x})$ for $\forall \mathbf{x}$, it leads to

$$\beta_i(\mathbf{x}) ||\mathbf{x} - \mathbf{x}_i||^2 \leq G(\mathbf{x}, \mathbf{x}_i, h_x^2 \mathbf{I}_{d_x}) ||\mathbf{x} - \mathbf{x}_i||^2.$$

This indicates $I_1 \leq I_2 = h_x^2$, no matter how $\mathbf{x}_i$ distributed.

As we know, for any finite number of samples N', the summation value will be less than the integration value when the function is positive, i.e.,

$$(h_x^2)^{new} < I_1 \leq I_2 = (h_x^2)^{old}.$$

From the above inequality, we can see that the approach always finds a h_x^2 regardless the data distribution and initialization. Because h_x^2 is nonnegative, the value of h_x^2 will approach to zero eventually.

In order to cope with the above-mentioned efficiency, we derive a new equation for estimating smoothing parameter h based on Kullback–Leibler divergence.

Rewrite (6) in the following form:

$$KL(\Theta) = \int p_{h_x}(\mathbf{x}) g(\mathbf{x}, \Theta) \, d\mathbf{x} + \int p_{h_x}(\mathbf{x}) \ln p_{h_x}(\mathbf{x}) \, d\mathbf{x} \equiv J_0 + J_h \tag{23}$$

where

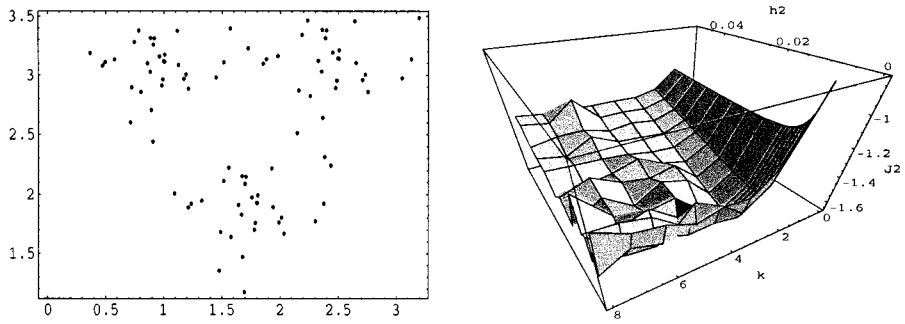
$$J_0 \equiv \int p_{h_x}(\mathbf{x}) g(\mathbf{x}, \Theta) \, d\mathbf{x} \tag{24}$$

$$J_h \equiv \int p_{h_x}(\mathbf{x}) \ln p_{h_x}(\mathbf{x}) \, d\mathbf{x} \tag{25}$$

$$g(\mathbf{x}, \Theta) \equiv -\ln p_{M_2}(\mathbf{x}, \Theta). \tag{26}$$

If we use Gaussian kernel density

$$p_{h_x}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} G(\mathbf{x}, \mathbf{x}_i, h^2 \mathbf{I}_d) \tag{27}$$



(a) (b)

Fig. 4. The quantized method for the synthetic data set with three clusters. (a) Data plot (b) The 3-D view of J_2 versus h, and k. A local minima occurs at $k = 3$ and h^2 is around 0.006.

then we obtain

$$J_0 = \int p_{h_x}(\mathbf{x}) g(\mathbf{x}, \Theta)\, d\mathbf{x} = \frac{1}{N} \sum_{i=1}^{N} \int G(\mathbf{x}, \mathbf{x}_i, h^2\mathbf{I}_d) g(\mathbf{x}, \Theta)\, d\mathbf{x}. \quad (28)$$

Because the $G(\mathbf{x}, \mathbf{x}_i, h^2\mathbf{I}_d)$ term is inside the integral in (29), when $\mathbf{x}$ moves away from $\mathbf{x}_i$, the function value becomes very small. So we can use Taylor expansion for $g(\mathbf{x}, \Theta)$ at $\mathbf{x} = \mathbf{x}_i$. When h is small, we can omit the higher order terms and only keep the first-order term. By doing this, we have the following approximation of J_0 (detailed derivations are given in the Appendix):

$$J_0(\mathbf{x}, \Theta, h) \approx J_{01}(\mathbf{x}_i, \Theta) + h^2 \frac{1}{2N} \cdot \sum_{i=1}^{N} \operatorname{trace}[\nabla\nabla g(\mathbf{x}_i, \Theta)] = J_{01}(\mathbf{x}_i, \Theta) + h^2 J_r(\mathbf{x}_i, \Theta) \quad (29)$$

$$KL(\mathbf{x}, \Theta, h) \approx J_0(\mathbf{x}_i, \Theta) + h^2 J_r(\mathbf{x}_i, \Theta) + \int p_{h_x}(\mathbf{x}) \ln p_{h_x}(\mathbf{x})\, d\mathbf{x} \quad (30)$$

$$\frac{\partial}{\partial h^2} KL(\mathbf{x}, \Theta, h) \approx J_r(\mathbf{x}_i, \Theta) + \frac{\partial}{\partial h^2} \cdot \int p_{h_x}(\mathbf{x}) \ln p_{h_x}(\mathbf{x})\, d\mathbf{x}. \quad (31)$$

We know that

$$\frac{\partial}{\partial h^2} \int p_{h_x}(\mathbf{x}) \ln p_{h_x}(\mathbf{x})\, d\mathbf{x} = \int \ln p_{h_x}(\mathbf{x}) \frac{\partial p_{h_x}(\mathbf{x})}{\partial h^2}\, d\mathbf{x} + \int \frac{\partial p_{h_x}(\mathbf{x})}{\partial h^2}\, d\mathbf{x} \quad (32)$$

where

$$\frac{\partial}{\partial h^2} p_{h_x}(\mathbf{x}) = -\frac{1}{2h^2} d_x p_{h_x}(\mathbf{x}) + \frac{1}{2N(h^2)^2} \cdot \sum_{i}^{N} G(\mathbf{x}, \mathbf{x}_i, h^2\mathbf{I}_d) \|\mathbf{x} - \mathbf{x}_i\|^2 \quad (33)$$

and the last term in (32) can be calculated as

$$\int \frac{\partial p_{h_x}(\mathbf{x})}{\partial h^2}\, d\mathbf{x} = 0. \quad (34)$$

So (32) becomes

$$\frac{\partial}{\partial h^2} J_h = \int \ln p_{h_x}(\mathbf{x}) \frac{\partial}{\partial h^2} p_{h_x}(\mathbf{x})\, d\mathbf{x} \quad (35)$$
$$= -\frac{d}{2h^2} \int p_{h_x}(\mathbf{x}) \ln p_{h_x}(\mathbf{x}) d\mathbf{x} + \frac{1}{2N(h^2)^2} \cdot \int \ln p_{h_x}(\mathbf{x}) \sum_{i}^{N} G(\mathbf{x}, \mathbf{x}_i, h^2\mathbf{I}_d) \|\mathbf{x} - \mathbf{x}_i\|^2\, d\mathbf{x}.$$

From

$$\frac{\partial}{\partial h^2} KL(\mathbf{x}, \Theta, h) = 0 \quad (36)$$

and with mean center approximation (see Appendix), we can obtain the new gradient descent formula for estimating h as

$$(h^2)^{new} = (h^2)^{old} + \eta \delta(h^2) \quad (37)$$

where η is a learning parameter and

$$\delta(h^2) \approx h^2 J_r - \frac{1}{2N} \sum_{j}^{N} (p_{h_x}(\mathbf{x}_j) - 1) \ln p_{h_x}(\mathbf{x}_j) \quad (38)$$

$$J_r(\mathbf{x}_i, \Theta) = \frac{1}{2N} \sum_{i=1}^{N} \left\| \sum_{y=1}^{k} P(y|\mathbf{x}_i)(\mathbf{x}_i - \mathbf{m}_y)^T \mathbf{\Sigma}_y^{-1} \right\|^2. \quad (39)$$

Let $\delta(h^2) = 0$, we obtain the following estimation equation for h^2:

$$h^2 = \frac{\frac{1}{2N} \sum_{j}^{N} [p_{h_x}(\mathbf{x}_j) - 1] \ln p_{h_x}(\mathbf{x}_j)}{J_r(\mathbf{x}_i, \Theta)}. \quad (40)$$

C. Experimental Results and Discussions

Now we present the experimental results for both the quantized and the gradient descent approximations of h.

In the experiments, we vary h^2 value from 0.001 to 0.05 and k from one to eight. From Fig. 4, we see that using the quantized

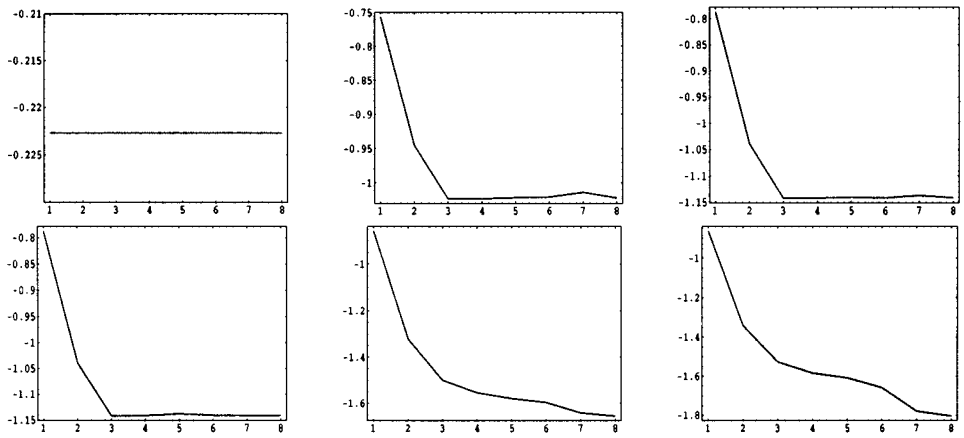


Fig. 5. Different h values and the corresponding $J_1(k)$ curves that are found from the results of the gradient descent approach. If h^2 equals 0.3783, k is underestimated, while for h^2 is less than 0.0024, then k is overestimated. (a) $h^2 = 0.3783$. (b) $h^2 = 0.04814$. (c) $h^2 = 0.03415$. (d) $h^2 = 0.03413$. (e) $h^2 = 0.0024$. (f) $h^2 = 0.0006$.

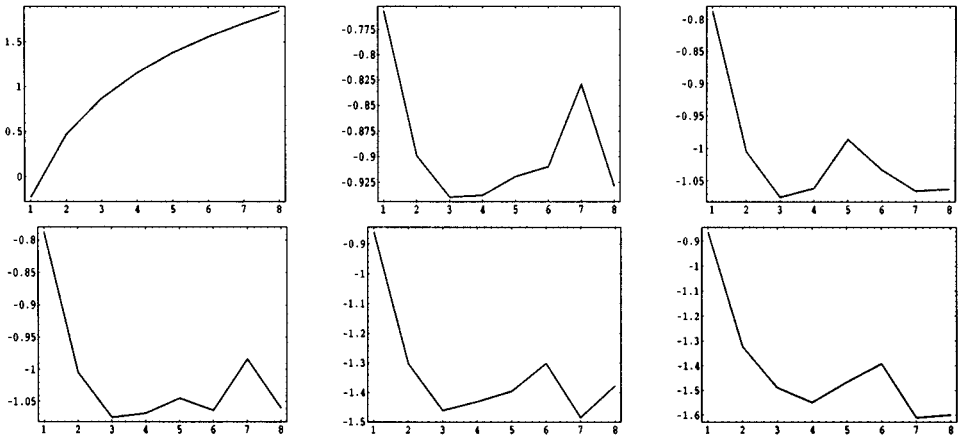


Fig. 6. Different h values and the corresponding $J_2(k)$ curves that are found from the results of the gradient descent approach. (a) $h^2 = 0.3783$. (b) $h^2 = 0.04814$. (c) $h^2 = 0.03415$. (d) $h^2 = 0.03413$. (e) $h^2 = 0.0024$. (f) $h^2 = 0.0006$.

method, k and h can be determined simultaneously. From these results, we obtain h^2 value at minimal $KL(h, k^*, \Theta^*)$.

From these experiments, we know that by using the gradient method, the searching range is limited in a small region of h^2 value compared to the quantized level method. Different k will result in different mixture model parameters, Θ; therefore it produces different h estimations. To find the optimal one, we can use the properties of J_1 and J_2 [17] to analyze the results and to determine k and h. We know that if $h = 0$ or h is too small, k will be over-estimated. If h is too large, the curve will be too smooth and k will be underestimated (see Fig. 5 for the comparison). In most cases, we can determine k and h from simulations easily. For example, in the experiments, the effect of data smoothing is somehow similar to increase of the number of samples. Figs. 5 and 6 show the results of the gradient descent approach. The results also confirm with the theorem in [14], $J_i(k^*) < J_i(k)$ if $k < k^*$, and $J_i(k^*) = J_i(k)$ if $k \geq k^*$, for $i = 1, 2$. From Fig. 6, we can easily find the possible k^* is three. From these figures, we obtain the optimal h^2 values, as 0.048 14, 0.034 13 and 0.034 15, respectively, through the gradient descent approach.

V. Summary

In this paper, we first review the BYY learning theory scheme for data smoothing. For a small set of samples, by combining data smoothing techniques with the SEM algorithm, we obtain a relatively robust performance for determining the cluster number.

The selection of the smoothing parameter h is a crucial problem. In this study, we derive an estimating formula for the smoothing parameter h. Often with the estimated h parameter, we can obtain a correct cluster number. Based on Kullback–Leibler divergence, we derive the gradient descent approach for estimating the smoothing parameter. The experiments indicate that the proposed approach works very well, and it is less computation-intensive compared to the exhausted search methods.

In fact, under the circumstance of different models, different sample sizes, and different data distributions, the determination of an appropriate cluster number using the Gaussian mixture model is very difficult. From our derivations and experiments, the BYY-based model selection criterion can select a reasonable cluster number even in a small set of samples.

APPENDIX
FORMULA OF ESTIMATING SMOOTHING PARAMETER

Here we derive the formula for estimating the smoothing parameter in the Gaussian mixture model case.

In the multidimension case, we have

$$J_r(\mathbf{x}_i, \Theta) = \frac{1}{2Nh} \sum_{i=1}^{N} \int G(\mathbf{x}, \mathbf{x}_i, h^2\mathbf{I}_d)(\mathbf{x}-\mathbf{x}_i)^T \cdot \nabla\nabla g(\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i)\, d\mathbf{x}$$

while $\nabla\nabla g(\mathbf{x}_i)$ is

$$\begin{aligned}\nabla\nabla g(\mathbf{x}_i) &= -\frac{[\nabla\nabla p_{M_2}(\mathbf{x}_i)]p_{M_2}(\mathbf{x}_i) - [\nabla p_{M_2}(\mathbf{x}_i)][\nabla p_{M_2}(\mathbf{x}_i)]}{(p_{M_2}(\mathbf{x}_i))^2} \\ &= \sum_{y=1}^{k} P(y|\mathbf{x}_i)\{\boldsymbol{\Sigma}_y^{-1} - \boldsymbol{\Sigma}_y^{-1}(\mathbf{x}_i-\mathbf{m}_y)(\mathbf{x}_i-\mathbf{m}_y)^T\boldsymbol{\Sigma}_y^{-1}\} \\ &\quad + \left\{\sum_{y=1}^{k} P(y|\mathbf{x}_i)(\mathbf{x}_i-\mathbf{m}_y)^T\boldsymbol{\Sigma}_y^{-1}\right\} \\ &\quad \cdot \left\{\sum_{y=1}^{k} P(y|\mathbf{x}_i)[(\mathbf{x}_i-\mathbf{m}_y)^T\boldsymbol{\Sigma}_y^{-1}]^T\right\}.\end{aligned}$$

Integrating it, we get

$$\begin{aligned}J_r(\mathbf{x}_i, \Theta) &= \frac{1}{2N}\sum_{i=1}^{N} \operatorname{trace}[\nabla\nabla g(\mathbf{x}_i)] \\ &\approx \frac{1}{2N}\sum_{i=1}^{N}\left\|\sum_{y=1}^{k} P(y|\mathbf{x}_i)(\mathbf{x}_i-\mathbf{m}_y)^T\boldsymbol{\Sigma}_y^{-1}\right\|^2.\end{aligned}$$

From (31) and (35), we have

$$h^2 J_r - \frac{d}{2}(h^2)^2 \int p_{h_x}(\mathbf{x}) \ln p_{h_x}(\mathbf{x})\, d\mathbf{x} + \frac{1}{2N}\sum_{i}^{N} \cdot \int \ln p_{h_x}(\mathbf{x}) G(\mathbf{x}, \mathbf{x}_i, h^2\mathbf{I}_d)\|\mathbf{x}-\mathbf{x}_i\|^2\, d\mathbf{x} = 0. \quad (41)$$

For the last term of the above equation, we use the mean center approximation, i.e., $\ln p_{h_x}(\mathbf{x}) \approx \ln p_{h_x}(\mathbf{x}_i)$

$$\frac{1}{2N}\sum_{i}^{N}\int \ln p_{h_x}(\mathbf{x}) G(\mathbf{x}, \mathbf{x}_i, h^2\mathbf{I}_d)\|\mathbf{x}-\mathbf{x}_i\|^2\, d\mathbf{x} \approx \frac{(h^2)^2}{2N}\sum_{i}^{N}\ln p_{h_x}(\mathbf{x}_i). \quad (42)$$

Combining the above equations and with the mean field approximation, we can obtain the following equation:

$$\delta(h^2) \approx (h^2)^{old} J_r - \frac{1}{2N}\sum_{j}^{N}[p_{h_x}(\mathbf{x}_j) - 1]\ln p_{h_x}(\mathbf{x}_j). \quad (43)$$

REFERENCES

[1] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency versus interpretability of classifications," in *Proc. Biometric Soc. Meet.*, Riverside, CA, 1965.

[2] H. H. Bock, "Probability models and hypotheses testing in partitioning cluster analysis," in *Clustering and Classification*. Riverside, CA: World Scientific, 1996, pp. 377–453.

[3] J. A. Hartigen, "Distribution problems in clustering," in *Classification and Clustering*, J. van Ryzin, Ed. New York: Academic, 1977, pp. 45–72.

[4] L. Xu, A. Krzyzak, and E. Oja, "Rival penalized competitive learning for clustering analysis, RBF net and curve detection," *IEEE Trans. Neural Networks*, vol. 4, pp. 636–649, July 1993.

[5] H. Bozdagan, "Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity," in *Proc. 1st US/Japan Conf. Frontiers Statist. Modeling: Inform. Approach*, vol. 2, 1994, pp. 69–113.

[6] E. P. Rosenblum, "A simulation study of information theoretic techniques and classical hypothesis tests in one factor anova," in *Proc. 1st US/Japan Conf. Frontiers Statist. Modeling: Inform. Approach*, vol. 2, 1994, pp. 319–346.

[7] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM Rev.*, vol. 26, pp. 195–239, 1984.

[8] K. E. Basford and G. J. McLachlan, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

[9] D. M. Titterington, "Some recent research in the analysis of mixture distributions," *Statistics*, vol. 21, pp. 619–641, 1990.

[10] J. H. Wolfe, "Pattern clustering by multivariate mixture analysis," *Multivariate Behavioral Res.*, vol. 5, pp. 329–350, 1970.

[11] M. P. Windham and A. Culter, "Information ratios for validating mixture analyzes," *J. Amer. Statist. Assoc.*, vol. 87, pp. 1188–1192, 1992.

[12] I. Mellin and T. Terasvirta, "Model selection criteria and model selection tests in regression models," *Stand. J. Statist.*, vol. 13, pp. 159–171, 1986.

[13] L. Xu, "How many clusters?: A YING–YANG machine based theory for a classical open problem in pattern recognition," in *Proc. IEEE Int. Conf. Neural Networks*, vol. 3, 1996, pp. 1546–1551.

[14] ——, "Bayesian Ying–Yang machine, clustering and number of clusters," *Pattern Recognition Lett.*, vol. 18, no. 11–13, pp. 1167–1178, 1997.

[15] Z. B. Lai, P. Guo, T. J. Wang, and L. Xu, "Comparison on Bayesian YING–YANG theory based clustering number selection criterion with information theoretical criteria," in *Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN'98)*, vol. I, Anchorage, AK, 1998, pp. 725–729.

[16] N. M. Laird, A. P. Dempster, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. B39, pp. 1–38, 1977.

[17] L. Xu, "Bayesian Ying–Yang system and theory as a unified statistical learning approach (VII): Data smoothing," in *Proc. Int. Conf. Neural Inform. Processing (ICONIP'98)*, Kitakyushu, Japan, 1, 1998, pp. 243–248.

[18] P. Guo and L. Xu, "On the study of BKYY cluster number selection criterion for small sample data set with bootstrap technique," in *Proc. 1999 Int. Joint Conf. Neural Networks (IJCNN'99)*, Washington, DC, 1999.

[19] L. Xu, "Bayesian YING–YANG system and theory as a unified statistical learning approach (I): For unsupervised and semi-unsupervised learning," in *Brain-Like Computing and Intelligent Information Systems*, S. Amari and N. Kassabov, Eds. New York: Springer-Verlag, 1997, pp. 241–247.

[20] ——, "Bayesian YING–YANG system and theory as a unified statistical learning approach (II): From unsupervised learning to supervised learning and temporal modeling," in *Theoretical Aspects of Neural Computation: A Multidisciplinary Perspective (TANC'97)*, I. King, K. W. Wong, and D. Y. Yeung, Eds. New York: Springer-Verlag, 1997, pp. 25–42.

[21] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.

[22] H. Bozdogan, "Multiple sample cluster analysis and approaches to validity studies in clustering individuals," Doctoral dissertation, Univ. Illinois, Chicago, 1981.

[23] ——, "Modle selection and Akaike's information criterion: The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[24] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[25] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. London, U.K.: Chapman and Hall, 1993.

[26] University of California. Machine Learning Database, Irvine. [Online]. Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases

[27] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency versus interpretability of classifications," *Biometrics*, vol. 21, no. 3, p. 768.