A Comparative Study on Lagrange Ying-Yang Alternation Method in Gaussian Mixture-Based Clustering

Weijian Long[1], Shikui Tu[1(✉)], and Lei Xu[1,2(✉)]

[1] Department of Computer Science and Engineering, and Center for Cognitive Machines and Computational Health, Shanghai Jiao Tong University, Shanghai, China
weijianlong7@126.com, {tushikui,leixu}@sjtu.edu.cn
[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China

Abstract. Gaussian Mixture Model (GMM) has been applied to clustering with wide applications in image segmentation, object detection and so on. Many algorithms were proposed to learn GMM with appropriate number of Gaussian components automatically determined. Lagrange Ying-Yang alternation method (LYYA) is one of them and it has advantages of no priors as well as the posterior probability bounded by traditional probability space. This paper aims to investigate the performance of LYYA, in comparisons with other methods including Bayesian Ying-Yang (BYY) learning, Rival penalized competitive learning (RPCL), hard-cut Expectation Maximization (EM) method, and classic EM with Bayesian Information Criterion (BIC). Systematic simulations show that LYYA is generally more robust than others on the data generated by varying sample size, data dimensionality and real components number. Unsupervised image segmentation results on Berkeley datasets also confirm LYYA advantages when comparing to the Mean shift and Multiscale graph decomposition algorithms.

Keywords: Gaussian Mixture Model · Lagrange Ying-Yang alternation method · Unsupervised image segmentation · Lagrange coefficient

1 Introduction

Gaussian Mixture Model (GMM) is a classic probabilistic model and has been widely used in clustering analysis, image segmentation, speaker identification [1]. Parameters learning and model selection are two essential parts of conventional

This work was supported by the Zhi-Yuan chair professorship start-up grant (WF220103010) from Shanghai Jiao Tong University.
Shikui Tu was supported by the Tenure-track associate professorship start-up grant from Shanghai Jiao Tong University.

H. Yin et al. (Eds.): IDEAL 2017, LNCS 10585, pp. 489–499, 2017.
https://doi.org/10.1007/978-3-319-68935-7_53

learning in GMM. Parameter learning is often implemented by Expectation-Maximization (EM) algorithm to maximize the likelihood, while determining the number, denoted as k, of Gaussian components is a model selection problem which is traditionally selected by a criterion, e.g., Bayesian Information Criterion (BIC), via a two-stage implementation which runs EM for all possible candidate component numbers.

However, this traditional model selection method is time-consuming. Efforts have been made on automatic model selection. Rival penalized competitive learning (RPCL) [2] is an early attempt on automatic model selection. Proposed in [3], BYY combines parameter learning with model selection and provides the general learning framework as well as specific algorithms. Moreover, automatic model selection can also be implemented via Bayesian approach with proper priors. Minimum message length (MML) [4] and variational Bayes(VB) [5] in GMM learning are two instances of this roadway. Readers can refer to [6] for a detailed analysis and comparison among the three Bayesian approaches.

Further improvements on Ying-Yang two-step alternation algorithm indicate that BYY learning methods may be improved without any help of priors if they can restrict covariance matrices as positive definite matrices [7] which is not considered in traditional BYY method. One recent method given in [8] is called Lagrange Ying-Yang alternation method (LYYA), which ignores influences from priors. It considers the Kullback-Leibler divergence between Ying structure and Yang structure as a Lagrange constraint and uses the coefficient η to control this restriction. There is still lack of a detailed investigation of LYYA in comparisons with other methods.

In this paper, we provide such a detailed investigation. A wide scope of configurations of experiments are considered to generate simulated data sets, with varying sample size, data dimensionality, number of clusters, overlap degree of clusters, and so on. LYYA shows best performance in simulations, comparing with the classic EM with BIC, RPCL, hard-cut EM. Even if repeating 5 times, BIC is still worse than LYYA when processing data with high overlapping degree or high dimensionality. We also study the impact of the coefficient η on model selection and clustering, and suggest an optimal scope of η. Real world applications are also considered. Unsupervised image segmentation results on Berkeley show that LYYA is better than other methods including hard-cut EM, RPCL, BYY and Mean shift algorithm [9].

2 Gaussian Mixture Model and EM Algorithm

For an item $x \in R^n$, Gaussian Mixture Model (GMM) supposes that it comes from a linear combination of k Gaussian distributions:

$$q(x|\theta) = \sum_{j=1}^{k} \alpha_j G(x|\mu_j, T_j), \alpha_j \geq 0, \sum_{j=1}^{k} \alpha_j = 1, \theta = \{\alpha_j, \mu_j, T_j\}_{j=1}^{k}, \quad (1)$$

where $G(x|\mu_j, T_j)$ represents a Gaussian density with mean μ_i and covariance matrix T_i, α_j is the mixing weight of the j-th Gaussian component. GMM can be

treated as a latent variable model by introducing a latent binary vector $y = \{y_1, \ldots, y_k\}$ to mark the Gaussian component the data x belongs to, where $\forall j, y_j \in \{0, 1\}$, $\sum_{j=1}^{k} y_{ij} = 1$. Then, we have

$$q(x|\theta) = \sum_{y} q(x, y|\theta), q(x, y|\theta) = \prod_{j=1}^{k} [\alpha_j G(x|\mu_j, T_j)]^{y_j} \tag{2}$$

Parameters can be estimated from a set of observations $X_N = \{x\}_{i=1}^{N}$ which is assumed to be independently identically distributed (i.i.d.) following GMM, by maximizing the likelihood function, i.e., $\max_\theta q(X_N|\theta) = \prod_{t=1}^{N} q(x_t|\theta)$, with the help of Expectation-Maximization (EM) algorithm, which iterates between the expectation step (E-step) and the maximization step (M-step):
E-step:

$$p_{ij} = p(j|x_i, \theta) = \frac{\alpha_j^{old} G(x_i|\mu_j^{old}, T_j^{old})}{\sum_{j=1}^{k} \alpha_j^{old} G(x_i|\mu_j^{old}, T_j^{old})} \tag{3}$$

M-step:

$$\alpha_j^{new} = \frac{\sum_i p_{ij}}{N}, \mu_j^{new} = \frac{\sum_i x_i p_{ij}}{\sum_i p_{ij}}, T_j^{new} = \frac{\sum_i p_{ij}(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_i p_{ij}}. \tag{4}$$

3 Model Selection Methods

3.1 Traditional Two-Stage Model Selection Method

Maximum likelihood (ML) is not a good principle to determine the number k of Gaussian components in GMM because its value increases as k grows, leading to the overfitting problem. A conventional way is to select the component number according to a model selection criterion such as Bayesian Information Criterion (BIC):

$$k^* = \arg\max_k J_{BIC}(k), J_{BIC}(k) = \ln q(X_N|\hat{\theta}_{ML}) - \frac{1}{2} d_k \ln N, \tag{5}$$

where $\hat{\theta}_{ML}$ is the ML estimate of parameters, d_k is the number of free parameters in GMM, and N is the sample size.

3.2 RPCL and Hard-Cut EM

Model selection by Eq. (5) is time-consuming because it requires running EM for a set of candidate component numbers. Efforts have been made on selecting k automatically during parameter learning. An early attempt is RPCL [2], in which not only the winner (i.e., the one with maximum posterior) is learned but also its rival (i.e., the second winner) is repelled a little bit from the sample

to reduce a duplicated information allocation. Thus, a batch version of RPCL learning is to replace Eq. (3) by:

$$p_{ij}^{new} = \begin{cases} 1 & j = j^*, j^* = \max_j p(j|x_i, \theta) \\ -\gamma & j = r, r = \max_{j \neq j^*} p(j|x_i, \theta) \\ 0 & otherwise \end{cases} \tag{6}$$

where $p(j|x_i, \theta)$ is the posterior probability computed by Eq. (3), and j^*,r represents the winner and the rival respectively, and γ controls the de-learning strength. When $\gamma = 0$, it degenerates to the so called hard-cut EM algorithm, see Eqs. (19) and (20) in [3].

3.3 Ying-Yang Alternation Method in BYY System

Firstly proposed in [3] and systematically developed in the past two decades, Bayesian Ying-Yang (BYY) harmony learning on typical structures leads to a class of algorithms that approach automatic model selection during parameter learning. Readers can refer to [8] for recent systematic introduction about BYY harmony learning. Briefly, BYY considers best harmony between two types of decomposition, namely Yang machine $p(R|X)p(X)$ and Ying machine $q(X|R)q(R)$, where data X is regarded to be generated from its inner representation $R = \{Y, \theta\}$ with latent variables Y and parameters θ. Mathematically, the BYY harmony learning is to maximize the following function, which is called harmony measure [3]:

$$H(p||q) = \int p(R|X)p(X) \ln[q(X|R)q(R)]dXdR \tag{7}$$

For GMM given in Eq. (1), if ignoring prior distributions over parameters, we have

$$H(p||q) = \sum_{j=1}^{k} \sum_{i=1}^{n} p(j|x_i, \theta) \ln[\alpha_j G(x_i|\mu_j, T_j)] \tag{8}$$

Maximizing the Eq. (8), subject to the structure $p(j|x_i, \theta)$ as the posterior distribution, leads to a BYY algorithm, iterating between Ying-Step which is the same as M-step in EM algorithm by Eq. (4) and Yang-Step given by

$$\begin{aligned} p_{ij}^{new} &= p(j|x_i, \theta)(1 + \delta_{ij}(\theta)) \\ \delta_{ij}(\theta) &= \ln[\alpha_j G(x_i|\mu_j, T_j)] - \sum_{j=1}^{k} p(j|x_i, \theta) \ln[\alpha_j G(x_i|\mu_j, T_j)] \end{aligned} \tag{9}$$

where $p(j|x_i, \theta)$ is calculated by Eq. (3) and $\delta_{ij}(\theta)$ is the adjustment on posterior probability $p(j|x_i, \theta)$. When $\delta_{ij}(\theta) > 0$, it will award the effect of the j_{th} component on sample x_i by enhancing the value of p_{ij}. When $\delta_{ij}(\theta) < 0$, it will give a punishment on p_{ij} and reduce the degree that the j_{th} component evolves toward sample x_i.

3.4 Lagrange Ying-Yang Alternation Method

The existing algorithms for maximizing Eq. (7) directly impose the equal-covariance constraint between Ying machine and Yang machine. Posterior probability p_{ij} calculated in Eq. (9) may be negative, which makes learning suffer from local optimum problem and learning instability [8]. To tackle this problem, the equal-covariance constraint can be indirectly considered as a Lagrange constraint [8], i.e.,

$$\begin{aligned} H_L(\theta) &= H(\theta) - \eta KL(p(Y|X)p(X)||q(X|Y,\theta)q(Y|\theta)) \\ H(\theta) &= \int p(Y|X)p(X)\ln[q(X|Y,\theta)q(Y|\theta)]dYdX \end{aligned} \tag{10}$$

where η is a Lagrange coefficient bounded by $\eta \geq 0$.

Maximizing Eq. (10) for GMM in Eq. (1) gives an algorithm called Lagrange Ying-Yang alternation (LYYA), in which Ying step is the same as the M-step in EM algorithm by Eq. (4), while Yang-step is given by:

$$p_{ij} = \frac{[\alpha_j^{old} G(x_i|\mu_j, T_j)]^{\frac{1+\eta}{\eta}}}{\sum_{j=1}^{k}[\alpha_j^{old} G(x_i|\mu_j, T_j)]^{\frac{1+\eta}{\eta}}} \tag{11}$$

when $\eta \to \infty$, $\frac{(1+\eta)}{\eta} \to 1$, Lagrange Ying-Yang alternation method will be equivalent to EM algorithm. When $\eta \to 0$, $\frac{(1+\eta)}{\eta} \to \infty$, Lagrange Ying-Yang alternation method will be extremely closed to hard-cut EM algorithm.

3.5 Rules to Trim a Gaussian Component During Automatic Model Selection

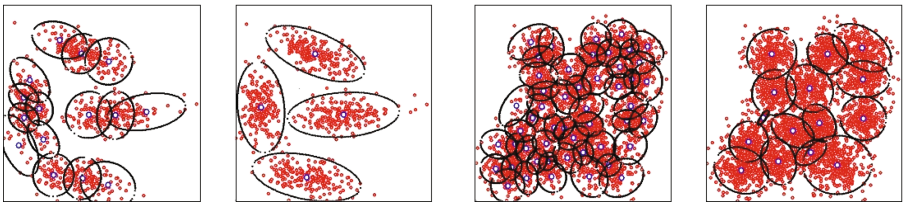
Automatic model selection is achieved during learning with the component being discarded when their mixing weights and determinants of covariance matrices are small enough. In this paper, we adopt the same trimming rule for all algorithms, i.e., for each iteration, among all the components with small enough mixing weights and covariance matrix determinants, the one with least determinant is discarded.

4 Simulation Experiment

4.1 Illustration on 2-D Datasets

To demonstrate how automatic model selection algorithms work, we give two synthetic 2-D datasets as shown in Fig. 1. We run EM+BIC, hard-cut EM, RPCL($\gamma = 0.0001$), BYY and LYYA($\eta = 2$) algorithms on both datasets, for 500 independent trials, respectively. All algorithms are initialized after the first round of K-means algorithm.

We use Rand Index (RI), Normalize Mutual Information (NMI) as well as Correct selection rate (CSR) to evaluate performances of algorithms. CSR is the



(a) Initial state in Dataset 1 (b) Result state in Dataset 1 (c) Initial state in Dataset 2 (d) Result state in Dataset 2

Fig. 1. (a) Dataset 1 is generated by a 4-component GMM with equal weights $\alpha_j^* = 0.25$, with 800 points and the component number is initialized to be $k = 15$. (c) Dataset 2 (taken from http://cs.joensuu.fi/sipu/datasets/) consists of 5000 points in 15 categories and is initialized with $k = 50$. (b) and (d) Successful cases with correct number of components determined. (Red indicates data points, blue represents means and black boundary indicates a contour of density of a Gaussian component.) (Color figure online)

frequency of correct number of clusters obtained by algorithms. All three criteria are at range $[0, 1]$ and the larger value, the better performance. Results in Table 1 show that LYYA gets best result except in CSR column of Dataset 2 where BYY obtains best result. In this two datasets, the performances between LYYA algorithm and BYY algorithm are closed and both are better than the other three algorithms.

Table 1. Performances of BIC, hard-cut EM, RPCL, BYY and LYYA on two datasets, where numbers in bold type indicate the best within columns

Algorithms	Dataset 1			Dataset 2		
	RI	NMI	CSR	RI	NMI	CSR
BIC	0.9895	0.9688	0.6980	0.9469	0.7466	0.0360
hard-cut EM	0.9791	0.9519	0.5300	0.9634	0.7905	0.2240
RPCL	0.9797	0.9530	0.5440	0.9609	0.7882	0.2520
BYY	0.9957	0.9822	**0.9940**	0.9638	0.7949	**0.6400**
LYYA	**0.9961**	**0.9826**	**0.9940**	**0.9640**	**0.7951**	0.2340

4.2 Systematic Comparisons

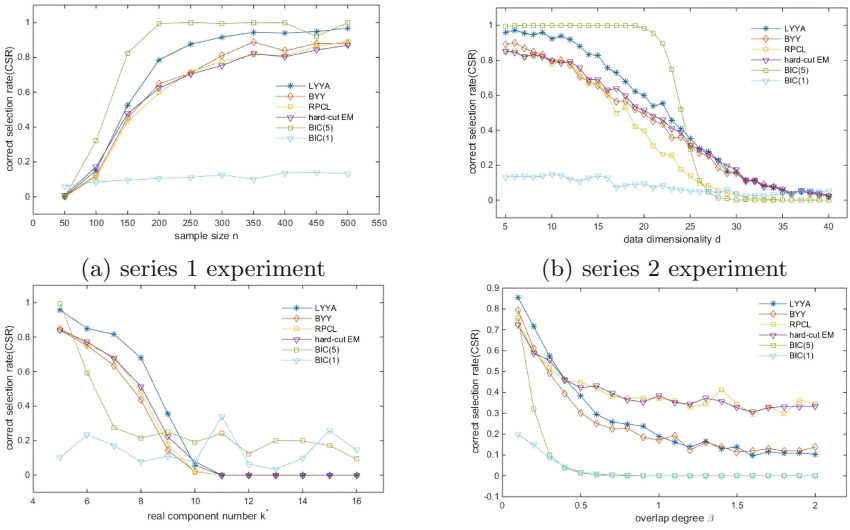
We compare all algorithms with extensive experiments as in [6] which cover a wide scope of conditions by varying sample size n, real components number k^*, data dimensionality d and overlap degree β. The synthetic datasets are generated by GMMs, and their mean vectors μ_j and covariance matrices T_j are randomly generated according to the joint Normal-Wishart distribution $G(\mu_j|m_j, T_j/)W(T_j|\phi, \gamma)$ with $\phi = I$, $\gamma = 50$, $m_j = 0$. The weights of components in all synthetic datasets are equal.

For each configuration $\{n, d, k^*, \beta\}$ in Table 2, all algorithms are run on 500 randomly generated datasets, starting from $k = 20$. The BIC value may not be reliable because EM suffers from local optimum problem. Therefore, we also repeatedly implement EM for 5 times and select the one with the largest likelihood for BIC calculation, denoted as BIC(5). It can be noted from the results in Fig. 2 that BIC(5) is much better than BIC(1) at the cost of 4 times more computation. Two main observations in Fig. 2 can be summarized as follows:

(1) Compared with RPCL, hard-cut EM and BYY algorithms, LYYA algorithm has better performance in all series experiments except for $\beta > 0.4$ in series 4.
(2) Even if ignoring the huge computational cost by BIC(5), LYYA is still more robust than BIC(5) for the cases with the data dimensionality exceeding 25 and the overlapping degree growing high.

Table 2. Four series comparison experiments

Starting cases	$\{n, d, k^*, \beta\} = \{500, 5, 5, 0.02\}$
Series 1	n varies in $[50, 100, 150 \cdots 500]$ with fixed d,k^* and β
Series 2	d varies in $[5, 6, 7 \cdots 40]$ with fixed n,k^* and β
Series 3	k^* varies in $[5, 6, 7 \cdots 16]$ with fixed n,d and β
Series 4	β varies in $[0.1, 0.2, 0.3 \cdots 2]$ with fixed n,k^* and d



(a) series 1 experiment (b) series 2 experiment

(c) series 3 experiment (d) series 4 experiment

Fig. 2. Results from BIC, hard-cut EM, RPCL, BYY, LYYA in four series experiments

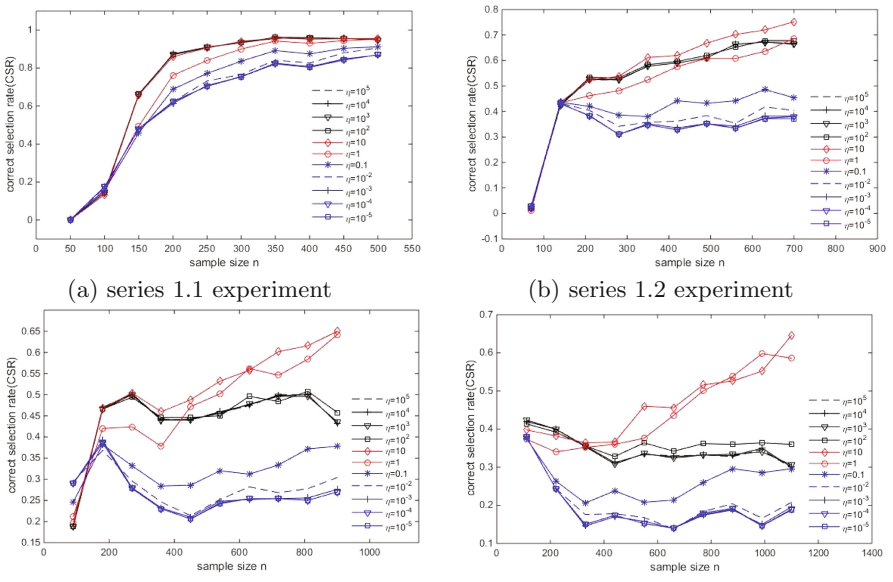
4.3 Investigation on η in LYYA

We choose series 1 experiment in Table 2 to study the change of η in LYYA, which details are illustrated in Table 3. The differences of datasets among various series experiments mainly come from their weights(α^*) of Gaussian components. From series 1.1 to series 1.4, the data volume difference of Gaussian components increases gradually. Each configuration $\{n, d, k^*, \beta\}$ consists of 500 cases and Initializes with 20 components. All cases are processed with 11 different η.

The result is shown in Fig. 3. When datasets share equal weights like series 1.1 experiment in Fig. 3(a), the performance of LYYA algorithm gets better as η grows and remains stable after $\eta = 10$. However, when data volume difference of Gaussian components increases in Fig. 3(b)–(d), $\eta = 1$ and $\eta = 10$ two red lines

Table 3. Four series experiments of $\eta \in \{10^{-5}, \cdots, 1, \cdots, 10^5\}$ in LYYA

Starting cases	$\{n, d, k^*, \beta\} = \{n, 5, 5, 0.02\}$
Series 1.1	$\alpha^* = \{\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\}$, n varies in $[50, 100, \cdots 500]$ with fixed d,k^*,β
Series 1.2	$\alpha^* = \{\frac{1}{7}, \frac{2}{7}, \frac{1}{7}, \frac{2}{7}, \frac{1}{7}\}$, n varies in $[70, 140, \cdots 700]$ with fixed d,k^*,β
Series 1.3	$\alpha^* = \{\frac{1}{9}, \frac{1}{3}, \frac{1}{9}, \frac{1}{3}, \frac{1}{9}\}$, n varies in $[90, 180, \cdots 900]$ with fixed d,k^*,β
Series 1.4	$\alpha^* = \{\frac{1}{11}, \frac{4}{11}, \frac{1}{11}, \frac{4}{11}, \frac{1}{11}\}$, n varies in $[110, 220, \cdots 1100]$ with fixed d,k^*,β



(a) series 1.1 experiment (b) series 1.2 experiment

(c) series 1.3 experiment (d) series 1.4 experiment

Fig. 3. Performances among 11 various η in LYYA from four series experiments

Table 4. Average score of criteria of 100 test images on BSDS300 dataset(k = 30), where numbers in bold type indicate the best within rows

Algorithms	Mean shift	MN-Cut	Hard-cut EM	RPCL	BYY	LYYA
PRI	0.4037	0.4214	0.4806	0.4960	0.4866	**0.5023**
RI	0.7242	0.7327	0.7289	0.7304	0.7282	**0.7359**
NMI	0.5352	**0.5765**	0.4799	0.4801	0.4827	0.4836

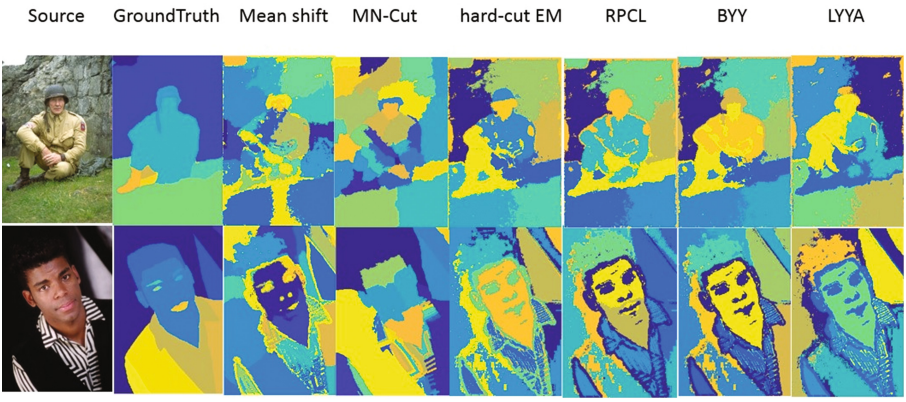


Fig. 4. Segmentation results of two images in BSDS300 dataset among six algorithms

obtain outstanding performances. As a result, we recommend that the value of η should be controlled within $[1, 10]$ when using LYYA method.

5 Application on Image Segmentation

We apply algorithms to unsupervised image segmentation on 100 test images in Berkerly dataset (http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/). We use Blobworld feature [10,11] plus position information to represent information of pixels. Blobworld feature of a pixel is a 6-D vector which consists of its color information from Lab space and 3-D texture information. Position information of a pixel is a 2-D vector of image coordinate and total feature is an 8-D vector per pixel.

We compare LYYA with hard-cut EM, RPCL, BYY, Mean shift [9] and Multiscale graph decomposition(MN-Cut) [12] algorithms on the dataset. The former four algorithms as well as MN-Cut algorithm initialize with 30 components and the bandwidth of Mean shift method is 0.15. Since the real component numbers (k^*) of images are uncertain, we evaluate segmentation results with RI, NMI as well as Probabilistic Rand Index(PRI) [13]. Each value of criteria is the average score of ground truth segmentations per image and the result is shown in Table 4. In this experiment, no data post-processing is applied, and thus the original clustering is kept.

Table 4 shows that LYYA method has best result in PRI as well as RI and MN-Cut algorithm owns the best score of NMI. On the vision of the segmentation examples in Fig. 4, the results of hard-cut EM, RPCL, BYY and LYYA are closed and they have better description than Mean shift as well as MN-Cut algorithms in image details though their dividing lines of different regions tend to be more rough.

6 Conclusion

In this paper, based on GMM, we provide a comparative study on several automatic model selection algorithms including LYYA, BYY, hard-cut EM and RPCL, together with BIC model selection criterion, through systematic experiments. Results indicate that LYYA algorithm is generally more robust than others on the data generated by varying sample size, data dimensionality and real components number. Described in [7], BYY algorithm may be unstable for little constrain on the range of posterior probability. The calculation of p_{ij} can be negative and may be not in the traditional probability space. Different from it, LYYA algorithm is easy computation and doesn't need extra adjustment to solve above problem. We provide an investigation on the Lagrange coefficient η in LYYA algorithm and the result indicates that as the amount of data in Gaussian components is increasingly unbalanced, the choice of various η is increasingly important and the ideal scope of η should be controlled in $[1, 10]$. Moreover, on image segmentation, LYYA outperforms hard-cut EM, RPCL as well as the previous BYY algorithm and also is generally better than Mean shift and MN-Cut algorithms.

References

1. Constantinopoulos, C., Titsias, M.K., Likas, A.: Bayesian feature and model selection for Gaussian mixture models. IEEE Trans. Pattern Anal. Mach. Intell. **28**(6), 1013–1018 (2006)
2. Xu, L., Krzyzak, A., Oja, E.: Rival penalized competitive learning for clustering analysis, RBF net and curve detection. IEEE Trans. Neural Netw. **4**(4), 636–649 (1993)
3. Xu, L.: Bayesian-Kullback coupled Ying-Yang machines: unified learnings and new results on vector quantization. In: Proceedings of International Conference on Neural Information Processing, pp. 977–988 (1995)
4. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 381–396 (2002)
5. Jaakkola, T.S., Jordan, M.I.: Bayesian parameter estimation via variational methods. Stat. Comput. **10**(1), 25–37 (2000)
6. Shi, L., Tu, S., Xu, L.: Learning Gaussian mixture with automatic model selection: A comparative study on three Bayesian related approaches. A special issue on Machine learning and intelligence science: IScIDE2010 (B). J. Front. Electr. Electron. Eng. China **6**(2), 215–244 (2011)
7. Chen, G., Heng, P.A., Xu, L.: Projection-embedded BYY learning algorithm for Gaussian mixture-based clustering. SpringerOpen J. Appl. Inform. **1**(2) (2014)

8. Xu, L.: Further advances on Bayesian Ying-Yang harmony learning. SpringerOpen J. Appl. Inform. 2(5), (2015)
9. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
10. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Mach. Intell. **24**(8), 1026–1038 (2002)
11. Nikou, C., Likas, A.C., Galatsanos, N.P.: A Bayesian framework for image segmentation with spatially varying mixtures. IEEE Trans. Image Process. Publ. IEEE Sig. Process. Soc. **19**(9), 2278–2289 (2010)
12. Cour, T., Bènèzit, F., Shi, J.: Spectral segmentation with multiscale graph decomposition. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. **2**(2), 1124–1131 (2005)
13. Carpineto, C., Romano, G.: Consensus clustering based on a new probabilistic rand index with application to subtopic retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **34**(12), 2315–2326 (2012)