2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Asymmetric Mixture Model with Variational Bayesian Learning

Thanh Minh Nguyen and Q. M. Jonathan Wu, *Senior Member, IEEE*

***Abstract*— Bayesian detection for the symmetric Gaussian mixture model has recently received great attention for pattern recognition problems. However, in many applications, the distribution of the data has a non-Gaussian and non-symmetric form. This study presents a new asymmetric mixture model for model detection. In this paper, the proposed asymmetric distribution is modeled with multiple Student's-t distributions, which are heavily tailed and more robust than Gaussian distributions. Our method has the flexibility to fit different shapes of observed data such as non-Gaussian and non-symmetric. Another advantage is that the proposed algorithm, which is based on the variational Bayesian learning, can simultaneously optimize over the number of the Student's-t distribution that is used to model each asymmetric distribution, and the number of components. The performance of the proposed model is compared to other mixture models, demonstrating the robustness, accuracy, and effectiveness of our method.**

I. Introduction

A mixture model is widely used in areas where statistical modeling of data is needed such as in bioinformatics, pattern recognition, and machine learning. The main advantage of this technique lies in its capability to use prior knowledge to model uncertainty in a probabilistic manner. Among the algorithms based on the Bayesian technique, the Gaussian mixture model (GMM) [1], [2] is a well-known method used for most applications. Although the GMM is a flexible and powerful tool for data analysis, it is sensitive to outliers and may lead to excessive sensitivity to small numbers of data points. Also, for many applied problems, the tail of the Gaussian distribution is shorter than required.

In order to improve the robustness of the algorithm for modeling data with different shapes, the Student's-t mixture model (SMM) has been proposed in [3], [4]. Compared to the GMM, each component of SMM has one more parameter called the degrees of freedom (v). This parameter is viewed as a robustness tuning parameter. For the particular case of v = 1, the Student's-t distribution reduces to the Cauchy distribution. When v tends to infinity, the Student's-t distribution approaches the Gaussian distribution. Also, for many applied problems [5], [6], the tail of the Gaussian distribution is shorter than required. Hence, SMM provides a more powerful and flexible approach for modeling data compared to the GMM. An advantage of GMM and SMM is that they require a small number of parameters for learning. Also, these parameters can be efficiently estimated by adopting the expectation maximization (EM) algorithm [7] to maximize the log-likelihood function.

Thanh Minh Nguyen and Q. M. Jonathan Wu are with the Department of Electrical and Computer Engineering, University of Windsor, 401 Sunset Avenue, Windsor, ON, Canada, N9B-3P4. Tel: +1 (519) 253-3000 Ext. 4862. Email: {nguyen1j, jwu}@uwindsor.ca

This research has been supported in part by the NSERC Discovery grant.

An important issue in mixture modeling is model detection, where the number of components is automatically selected. In order to overcome this issue, a variational Bayesian for the Gaussian mixture model (VB-GMM) is proposed in [8], [9]. In this technique, the mixture model is started with a large number of components. The competition among components finally yields a model which is composed of dominant components, and the redundant components are removed. In order to improve the robustness of the algorithm, a variational Bayesian for the Student's-t mixture model (VB-SMM), which includes VB-GMM as a special case, has been proposed in [10]. Compared to the VB-GMM, the VB-SMM is less sensitive to outliers and can obtain robust estimates for the mean of a set of data points.

All the mixture models with Bayesian detection [8]–[10] are based on the symmetric Gaussian distribution for modeling the underlying distributions. In real applications [11], [12], however, the intensity distribution of each label type of the dataset are not symmetric. Motivated by the aforementioned observations, we introduce in this paper a new asymmetric mixture model for model detection. Our approach differs from those discussed above by the following advantages. First, the Student's-t distribution, which is heavily tailed and more robust than the Gaussian, is used in this paper. Secondly, while the previous models are based on the symmetric distribution, we propose a new distribution that is applied for the data that has the asymmetric distribution. Finally, the proposed algorithm, which is based on the variational Bayesian learning, can simultaneously optimize over the number of the Student's-t distribution that is used to model each asymmetric distribution, and the number of components. We demonstrate through extensive simulations that the proposed model is superior to other mixture models.

The remainder of this paper is organized as follows: section II presents a brief introduction of the mixture model with Bayesian detection, commonly used in the literature; section III describes the proposed method in detail; section IV presents the parameter estimation; section V sets out the experimental results; and section VI presents our conclusions.

II. Related Works

Notations used throughout the paper are as follows. Let x_i denote the i-th observation. The main objective is to cluster a dataset consisting of N real observations into K labels. Labels are denoted by (Ω_1,Ω_2,...,Ω_K). Let us consider the problem of estimating the posterior probability of x_i belonging to label Ω_j. The mixture model [1] assumes the

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

density function $f(x_i|\Omega_j)$ at an observation x_i is given by

$$f(x_i|\Omega_j) = \sum_{j=1}^{K} \pi_j \varphi(x_i|\Omega_j) \tag{1}$$

Each distribution $\varphi(x_i|\Omega_j)$ is called a component of the mixture. Note that, $\varphi(x_i|\Omega_j)$ can be any kind of distribution. In GMM [1], [2], $\varphi(x_i|\Omega_j)$ is the Gaussian distribution $\Phi(x_i|\mu_j, \tau_j)$ with two parameters: the mean μ_j, and the precision τ_j. Although the GMM is a flexible and powerful tool for data analysis, it is sensitive to outliers and may lead to excessive sensitivity to small numbers of data points. Also, for many applied problems, the tail of the Gaussian distribution is shorter than required.

In order to improve the robustness of the model, Student's-t distribution is used in SMM [3], [4]. In this model, $\varphi(x_i|\Omega_j)$ is the Student's-t distribution $\mathcal{S}(x_i|\mu_j, \tau_j, v_j)$ with longer tails and one more parameter compared to the Gaussian distribution $\Phi(x_i|\mu_j, \sigma_j)$. Each Student's-t distribution has its own mean μ_j, precision τ_j, and degree of freedom v_j. Given function $\varphi(x_i|\Omega_j)$, the likelihood function can be written as

$$p(\mathrm{x}) = \prod_{i=1}^{N} \sum_{j=1}^{K} \pi_j \varphi(x_i|\Omega_j) \tag{2}$$

In order to optimize the parameters, we need to maximize the log-likelihood function given in (2) with EM algorithm [7], or by variational Bayesian (VB) approximation [8], [9], [13]. However, for the EM algorithm, in order to estimate the number of components, other criteria such as entropy measure or minimal message length are required. Also, another problem encountered by the EM algorithm is that singular components lead to infinite likelihood, which is not the case with VB.

III. Proposed Method

As shown in section II, the main goal of a mixture model with Bayesian detection is to establish a model that can best describe the statistical properties of the underlying source. The existing mixture models have relied on $\varphi(x_i|\Omega_j)$, which is based on the symmetric distribution, for modeling the underlying distributions. In many applications, the intensity distribution of each label type of the dataset does not exhibit exactly a Gaussian shape and is not symmetric. In order to overcome this problem, we present a new asymmetric mixture model for model detection, which is useful for modeling non-Gaussian data.

First, we define a new non-Gaussian and non-symmetric distribution that is used for the component of our mixture model. Differing from the above-mentioned mixture models, each component density in our model is modeled with multiple Student's-t distributions. The function $\varphi(x_i|\Omega_j)$ in our model is defined as

$$\varphi(x_i|\Omega_j) = \sum_{m=1}^{M_j} \eta_{jm} \mathcal{S}(x_i|\mu_{jm}, \tau_{jm}, v_{jm}) \tag{3}$$

where M_j is the number of the Student's-t distribution that is used to model the label Ω_j. And η_{jm} is called the weighting factor that satisfies the following constraints: $\eta_{jm} \geq 0$ and $\sum_{m=1}^{M_j} \eta_{jm} = 1$.

In (3), $\mathcal{S}(x_i|\mu_{jm}, \tau_{jm}, v_{jm})$ is the Student's-t distribution with the mean μ_{jm}, the precision τ_{jm}, and the degree of freedom v_{jm}.

$$\mathcal{S}(x_i|\mu_{jm}, \tau_{jm}, v_{jm}) = \frac{\Gamma(v_{jm}/2 + 1/2)}{\Gamma(v_{jm}/2)} \left(\frac{\tau_{jm}}{\pi v_{jm}}\right)^{1/2} \times \left[1 + \frac{\tau_{jm}(x_i - \mu_{jm})^2}{v_{jm}}\right]^{-(v_{jm}+1)/2} \tag{4}$$

In (4), $\Gamma(\cdot)$ is the gamma function. The idea to define the distribution in (3) is an extension and improvement on the idea presented in [12]. It is based on the fact that non-symmetric distribution can be approximated by multiple Student's-t distributions. Compared with the original work in [12], the advantage of our method, which is based on the variational Bayesian learning, can simultaneously optimize over the number of the Student's-t distribution that is used to model the label Ω_j, and the number of components.

Given the function $\varphi(x_{il}|\Omega_j)$ in (3), the next objective is to optimize the parameter. We now consider the VB approximation [8], [9], [13] for the proposed model through the hidden variables z_{ij}, where $z_{ij} = \{0, 1\}$ and $\sum_{j=1}^{K} z_{ij} = 1$. If a given data point x_i is generated from the label Ω_j the value of z_{ij} is one; otherwise, it is zero. The weighting factor is expressed through the hidden variables y_{ijm}, where $y_{ijm} = \{0, 1\}$ and $\sum_{m=1}^{M_j} y_{ijm} = 1$. The value of $y_{ijm} = 1$ indicates that the Student's-t distribution $\mathcal{S}(x_i|\mu_{jm}, \tau_{jm}, v_{jm})$ is associated with label Ω_j.

Given the sets of hidden variables $\mathrm{z} = \{z_{ij}\}$, and $\mathrm{y} = \{y_{ijlm}\}$, the data is assumed to be independently drawn from a distribution.

$$p(\mathrm{x}|\mathrm{z}, \mathrm{y}, \mu, \tau, v) = \prod_{i=1}^{N} \prod_{j=1}^{K} \left[\prod_{m=1}^{M_j} \left(\mathcal{S}(x_i|\mu_{jm}, \tau_{jm}, v_{jm})\right)^{y_{ijm}}\right]^{z_{ij}} \tag{5}$$

Note that there is no closed form solution for maximizing the likelihood under a Student's-t distribution. To overcome this problem, the Student's-t distribution in [10] is represented as an infinite mixture of scaled Gaussians. By adopting the idea in [10], with the latent variables $u = \{u_{ijm}\}$, the Student's t distribution in (5) can be equivalent to a Gaussian distribution as follows

$$p(\mathrm{x}|\mathrm{z}, \mathrm{y}, \mu, \tau, u) = \prod_{i=1}^{N} \prod_{j=1}^{K} \left[\prod_{m=1}^{M_j} \left(\Phi(x_i|\mu_{jm}, u_{ijm}\tau_{jm})\right)^{y_{ijm}}\right]^{z_{ij}} \tag{6}$$

The distribution of the hidden variables z given the prior probabilities π_j, and the hidden variables y given the weighting factor η_{jm} is given by

$$p(\mathrm{z}|\pi) = \prod_{i=1}^{N} \prod_{j=1}^{K} \pi_j^{z_{ij}}; \quad p(\mathrm{y}|\eta) = \prod_{i=1}^{N} \prod_{j=1}^{K} \prod_{m=1}^{M_j} \eta_{jm}^{y_{ijm}} \tag{7}$$

The model selection is accomplished by introducing conjugate priors over the means, precisions, and degrees of freedom.

$$p(\mu)=\prod_{j=1}^{K}\prod_{m=1}^{M_j}\Phi(\mu_{jm}|o,c);\quad p(\tau)=\prod_{j=1}^{K}\prod_{m=1}^{M_j}\mathcal{G}(\tau_{jm}|\alpha,\beta)$$
$$p(u)=\prod_{i=1}^{N}\prod_{j=1}^{K}\prod_{m=1}^{M_{jl}}\mathcal{G}(u_{ijm}|v_{jm}/2,v_{jm}/2) \tag{8}$$

In (8), $\mathcal{G}(\cdot)$ is the Gamma distribution. The hyperparameters o, c, α, and β control the distributions over μ_{jm} and τ_{jm}. It is noticed that there is no conjugate prior for v_{jm}; we set their value by optimization as part of the variational procedure [10].

IV. Parameter Learning

To simplify notation, we define $\theta=\{\mathrm{z},\mathrm{y},\mu,\tau,u\}$, the set of random variables. And $\vartheta=\{\pi,\rho,\eta\}$ is the set of parameters. In order to obtain the estimation of parameters, we maximize the marginal likelihood $p(\mathrm{x}|\vartheta)$ by integrating out the variables as follows

$$p(\mathrm{x}|\vartheta)=\int p(\mathrm{x},\theta|\vartheta)d\theta \tag{9}$$

Since the integration in the equation above is intractable, an alternative way to solve this problem is VB approximation [8], [9], [13], which aims to maximize a lower bound of the logarithmic marginal likelihood.

$$L(Q,\vartheta)=\int Q(\theta)\frac{p(\mathrm{x},\theta|\vartheta)}{Q(\theta)}d\theta\le\log p(\mathrm{x}|\vartheta) \tag{10}$$

where Q is an arbitrary distribution, which provides an approximation to the true posterior distribution. We see that the function $L(Q,\vartheta)$ forms a rigorous lower bound on the true log marginal likelihood. Although the computation of original log likelihood function $\log p(\mathrm{x}|\vartheta)$ is not tractable, the lower bound $L(Q,\vartheta)$ may be tractable to compute through choosing a suitable form for the Q distribution. The difference between the lower bound $L(Q,\vartheta)$ and the true log likelihood$\log p(\mathrm{x}|\vartheta)$ is the Kullback-Leibler (KL) divergence. The goal in a variational approach is to choose a suitable form for Q such that the evaluation of the lower bound becomes tractable. For this purpose, we approximate p with optimizing the Q by minimizing the KL divergence. Minimizing the KL divergence with respect to all possible function forms of Q, the standard variational approach provides the following general form of the solutions

$$Q(\theta_i)=\frac{\exp\langle p(\mathrm{x},\theta|\vartheta)\rangle_{k\neq i}}{\int\exp\langle p(\mathrm{x},\theta|\vartheta)\rangle_{k\neq i}d\theta_i} \tag{11}$$

where $\langle\cdot\rangle_{k\neq i}$ denotes an expectation with respect to the distributions $Q_k(\theta_k)$ for all $k\neq i$. The factors of the function Q are given by calculation of (11) as follows

$$\begin{aligned}Q(\mathrm{z})&=\prod_{i=1}^{N}\prod_{j=1}^{K}r_{ij}^{z_{ij}};\quad Q(\mathrm{y})=\prod_{i=1}^{N}\prod_{j=1}^{K}\prod_{m=1}^{M_j}\upsilon_{ijm}^{y_{ijm}}\\ Q(\mu)&=\prod_{j=1}^{K}\prod_{m=1}^{M_j}\Phi(\mu_{jm}|o_{jm}^{\mu},c_{jm}^{\mu})\\ Q(\tau)&=\prod_{j=1}^{K}\prod_{m=1}^{M_j}\mathcal{G}(\tau_{jm}|\alpha_{jm}^{\tau},\beta_{jm}^{\tau})\\ Q(u)&=\prod_{i=1}^{N}\prod_{j=1}^{K}\prod_{m=1}^{M_j}\mathcal{G}(u_{ijm}|\alpha_{ijm}^{u},\beta_{ijm}^{u})\end{aligned} \tag{12}$$

The variational parameters $r_{ij},\upsilon_{jm},c_{jm}^{\mu},o_{jm}^{\mu},\alpha_{jm}^{\tau},\beta_{jm}^{\tau},\alpha_{ijm}^{u},\beta_{ijm}^{u}$ are given by maximizing and determining the density involved in Q. After some algebra, the following equations are obtained

$$r_{ij}=\frac{\pi_j\tilde{r}_{ij}}{\sum_{k=1}^{K}\pi_k\tilde{r}_{ik}} \tag{13}$$

$$\begin{aligned}\tilde{r}_{ij}=\exp\{0.5\sum_{m=1}^{M_{jl}}\upsilon_{ijm}(&-\log 2\pi+\Psi(\alpha_{ijm}^{u})\\ &-\log\beta_{ijm}^{u}+\Psi(\alpha_{jm}^{\tau})-\log\beta_{jm}^{\tau}\\ &-\log\beta_{jm}^{\tau}-\alpha_{ijm}^{u}(\beta_{ijm}^{u})^{-1}\alpha_{jlm}^{\tau}(\beta_{jm}^{\tau})^{-1}\\ &\times((x_{il}-o_{jm}^{\mu})^2+(c_{jm}^{\mu})^{-1}))\}\end{aligned} \tag{14}$$

$$\upsilon_{ijm}=\frac{\eta_{jm}\tilde{\upsilon}_{ijm}}{\sum_{k=1}^{M_{jl}}\eta_{jk}\tilde{\upsilon}_{ijk}} \tag{15}$$

$$\begin{aligned}\tilde{\upsilon}_{ijm}=\exp\{&0.5r_{ij}(-\log 2\pi+\Psi(\alpha_{ijm}^{u})\\ &-\log\beta_{ijm}^{u}+\Psi(\alpha_{jm}^{\tau})-\log\beta_{jm}^{\tau}\\ &-\alpha_{ijm}^{u}(\beta_{ijm}^{u})^{-1}\alpha_{jlm}^{\tau}(\beta_{jm}^{\tau})^{-1}\\ &\times((x_{il}-o_{jm}^{\mu})^2+(c_{jm}^{\mu})^{-1}))\}\end{aligned} \tag{16}$$

$$c_{jm}^{\mu}=c+\alpha_{jm}^{\tau}(\beta_{jm}^{\tau})^{-1}\sum_{i=1}^{N}\alpha_{ijm}^{u}(\beta_{ijm}^{u})^{-1}r_{ij}\upsilon_{ijm} \tag{17}$$

$$o_{jm}^{\mu}=\frac{oc+\alpha_{jm}^{\tau}(\beta_{jm}^{\tau})^{-1}\sum_{i=1}^{N}\alpha_{ijm}^{u}(\beta_{ijm}^{u})^{-1}r_{ij}\upsilon_{ijm}x_i}{c+\alpha_{jm}^{\tau}(\beta_{jm}^{\tau})^{-1}\sum_{i=1}^{N}\alpha_{ijm}^{u}(\beta_{ijm}^{u})^{-1}r_{ij}\upsilon_{ijm}} \tag{18}$$

$$\alpha_{jm}^{\tau}=\alpha+\frac{1}{2}\sum_{i=1}^{N}r_{ij}\upsilon_{ijm} \tag{19}$$

$$\begin{aligned}\beta_{jm}^{\tau}=\beta+\frac{1}{2}\sum_{i=1}^{N}&r_{ij}\upsilon_{ijm}\times\\ &((x_i-o_{jm}^{\mu})^2+(c_{jm}^{\mu})^{-1})\alpha_{ijm}^{u}(\beta_{ijm}^{u})^{-1}\end{aligned} \tag{20}$$

$$\alpha_{ijm}^{u}=\frac{1}{2}(v_{jm}+r_{ij}\upsilon_{ijm}) \tag{21}$$

$$\begin{aligned}\beta_{ijm}^{u}=\frac{1}{2}v_j+\frac{1}{2}&r_{ij}\upsilon_{ijm}\times\\ &((x_i-o_{jm}^{\mu})^2+(c_{jm}^{\mu})^{-1})\alpha_{jm}^{\tau}(\beta_{jm}^{\tau})^{-1}\end{aligned} \tag{22}$$

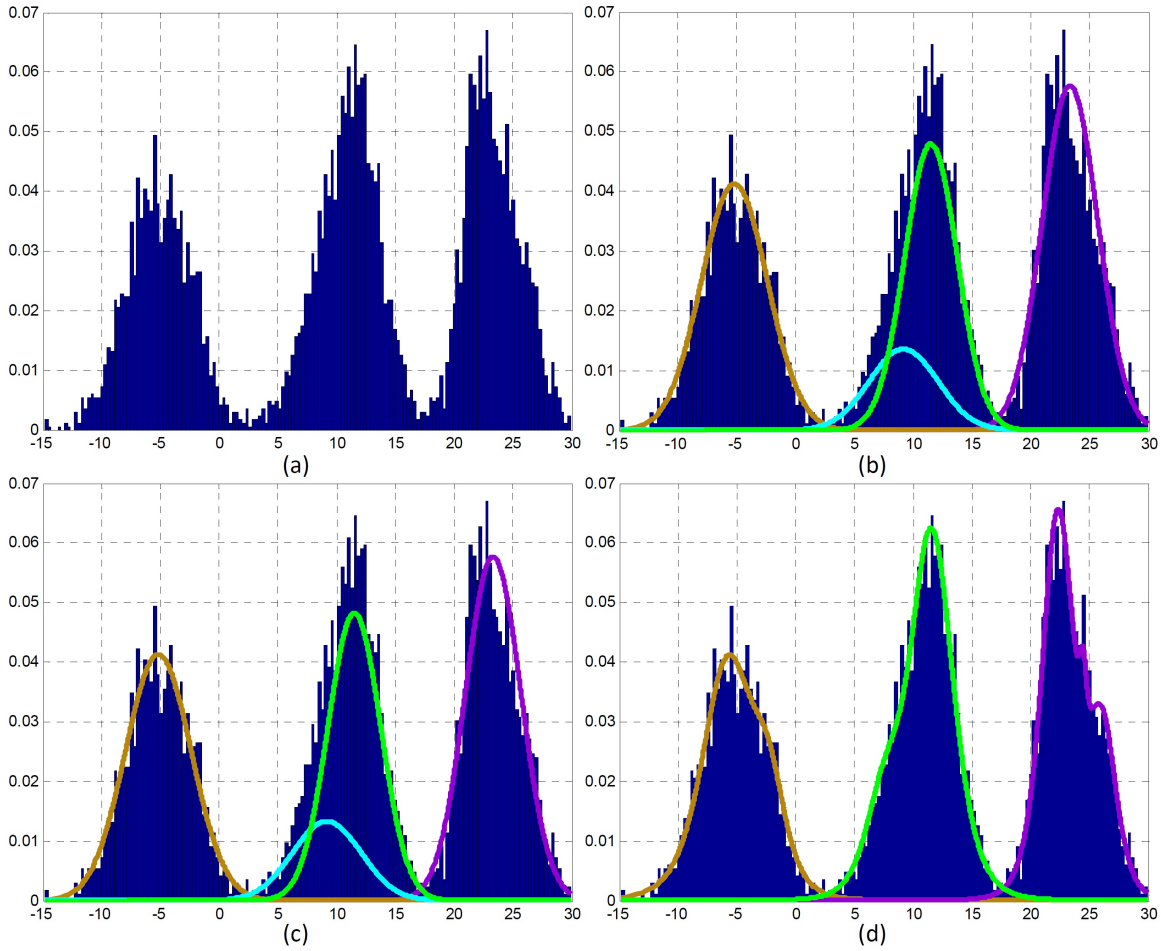

Fig. 1. The first experiments, (a): The original data with three labels, (b): VB-GMM, (c): VB-SMM, (d): Our method.

Since no conjugate prior is imposed on the degree of freedom, we update v_{jm} with the log-marginal maximum likelihood estimates by setting the corresponding gradient to zero and solving the non-linear equations as follows

$$-\Psi(v_{jm}/2) + \log(v_{jm}/2) + 1 + \frac{\sum_{i=1}^{N} r_{ij} v_{ijm} \left(\Psi(\alpha^u_{ijm}) - \log \beta^u_{ijm} - \alpha^u_{ijm} (\beta^u_{ijm})^{-1} \right)}{\sum_{i=1}^{N} r_{ij} v_{ijm}} = 0 \tag{23}$$

where $\Psi(\cdot)$ is the digamma function. In order to estimate π_j and η_{jm} we need to set the derivative of L with respect to these parameters, and, equal to zero, we get the following update rules:

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} r_{ij} \text{ and } \eta_{jm} = \frac{1}{N} \sum_{i=1}^{N} v_{ijm} \tag{24}$$

So far, the discussion has focused on estimating the parameter of the model. The property of the VB algorithm guarantees that the components with similar parameters fitting the same Student's t-distributions generate a dominant cluster. Thus, we can start the model with a large number of components and a large number of numbers of the Student's-t distribution that is used to model the label Ω_j for the l-th feature. The update of the parameters π_j and η_{jm} can simultaneously optimize over the number of the Student's-t distribution that is used to model each asymmetric distribution, and the number of components.

V. Experiments

In this section, the performance of the proposed method is compared to the VB-GMM [9], VB-SMM [10]. These compared methods are initialized by the K-mean algorithm similar to the initialization of the proposed algorithm. For all methods, we start with K = 20 in this paper, and it is run until the convergence of the iteration steps. In our method, for each label, we start with $M_j = 5$. In order to evaluate the segmentation performance quantitatively, variation of information (VI) [14], is employed. Note that, the lower the value of VI, the better the quality of the segmentation.

In the first experiment, in order to explain why the performance of the proposed distribution is better than the regular distribution, we show a sample with 5900 simulated points from three labels. Labels 1, 2, and 3 have 1700, 2200, and 2000 points, respectively. The ground truth distributions of the three labels are shown in Fig. 1(a). As shown in this

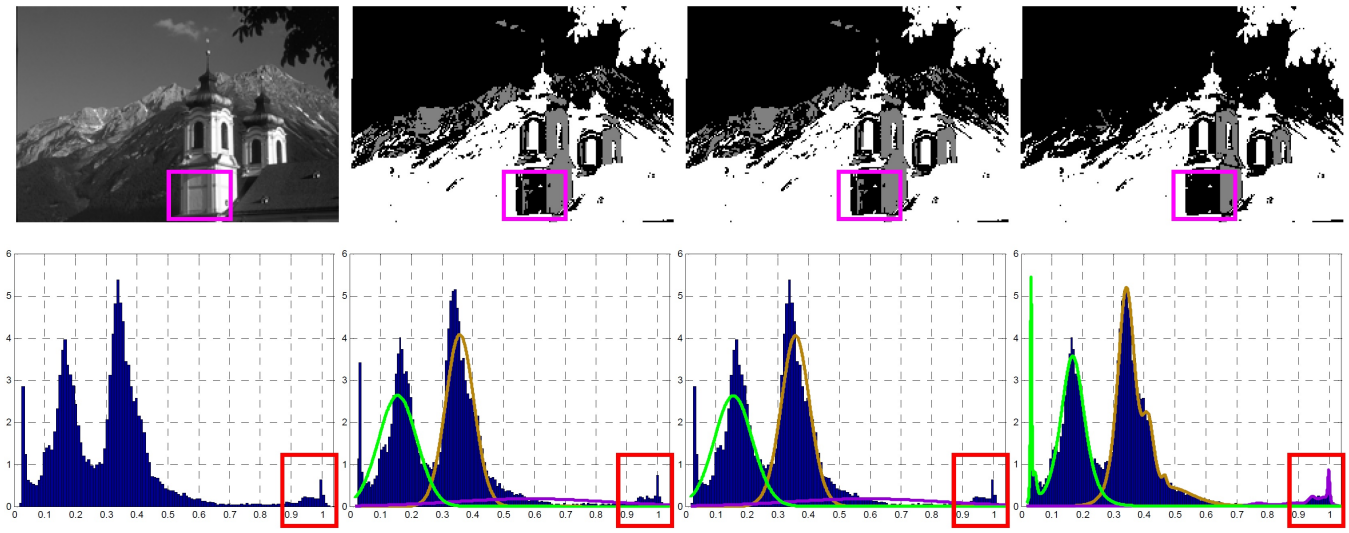

Fig. 2. The second experiments, (1^{st} column): The original image, (2^{nd} column): VB-GMM (VI = 2.6825), (3^{rd} column): VB-SMM (VI = 2.6814), (4^{th} column): Our method (VI = 2.6141).

figure, the intensity distribution of each label type does not exhibit an exact Gaussian shape, and is not symmetric. In Fig. 1(b)–(d), we show the results of VB-GMM, VB-SMM, and our method, respectively. As shown in Fig. 1(b) and (c), the performances of the VB-GMM and VB-SMM are very poor in this situation. The error of the estimated distributions compared to the ground truth distributions remains quite high. Also, the number of the labels detected from these methods is not correct (K = 4). As shown in Fig. 1(d), the proposed method can correctly detect the number of the labels (K = 3), and can better estimate the observed data in comparison to the two previous results.

In the second experiment, a real-world grayscale image from Berkeley Dataset is used to evaluate the performance of the proposed method against VB-GMM and VB-SMM. As shown in Fig. 2, all compared methods obtain the same number of the labels (three labels). However, a closer inspection of the marked box indicates the proposed yields a better result with a smaller value of VI. Moreover, we notice that the estimated distribution of the proposed method obtains a better estimate for the histogram compared to other methods.

A set of real world images are used to evaluate the performance of the proposed method against VB-GMM and VB-SMM methods. Table I contains the cumulative results obtained for all methods, for the given set of real world images. As evident from the results, on average, the proposed method outperforms other methods with a lower VI. Fig. 3 shows some of the other real-world images used for segmentation by employing VB-GMM, VB-SMM and the proposed method, respectively. The first row shows the original images, followed by the corresponding segmented images in the second, third, and the last row. Fig. 3 clearly indicates that our proposed method achieves a better segmentation accuracy.

TABLE I

COMPARISON OF IMAGE SEGMENTATION RESULTS ON BERKELEY'S GRAYSCALE IMAGE SEGMENTATION DATASET: VI.

Image	VB-GMM	VB-SMM	Our method
2092	2.4421	2.4153	1.4377
24063	1.9504	1.9501	1.6892
43051	1.9352	1.9228	1.3241
43070	2.5762	2.5545	1.1735
65084	4.0060	4.0060	4.0899
76002	4.0867	4.0873	4.0154
105019	2.6563	2.6119	0.4751
106025	1.6133	1.6117	1.4401
117025	2.6053	2.5972	2.2322
118035	2.4264	2.4264	2.0681
147021	1.3078	1.3087	0.9477
176039	2.3272	2.3329	2.0127
179084	2.4211	2.5475	2.0231
198087	3.5219	3.5200	3.5803
216041	2.6011	2.5438	1.9428
250087	2.9332	2.9088	2.6510
253036	1.9979	1.9691	0.6044
271031	1.8325	1.8319	1.2910
285022	2.4809	2.4869	2.2793
286092	2.5316	2.5299	2.4353
361084	4.0983	4.0971	3.5025
384022	2.4205	2.4139	2.1683
Mean	**2.5805**	**2.5761**	**2.0629**

VI. CONCLUSIONS

We have presented a new non-symmetric mixture model for model detection in this paper. The distribution of our method has a non-Gaussian and non-symmetric form. Each label is modeled with multiple Student's-t distributions, which are heavily tailed and more robust than Gaussian distributions. The advantage of our method is that it has the flexibility to fit different shapes of observed data such as non-Gaussian and non-symmetric. Besides that, our method can simultaneously optimize over the number of the Student's-t distribution that is used to model each asymmetric distribu-

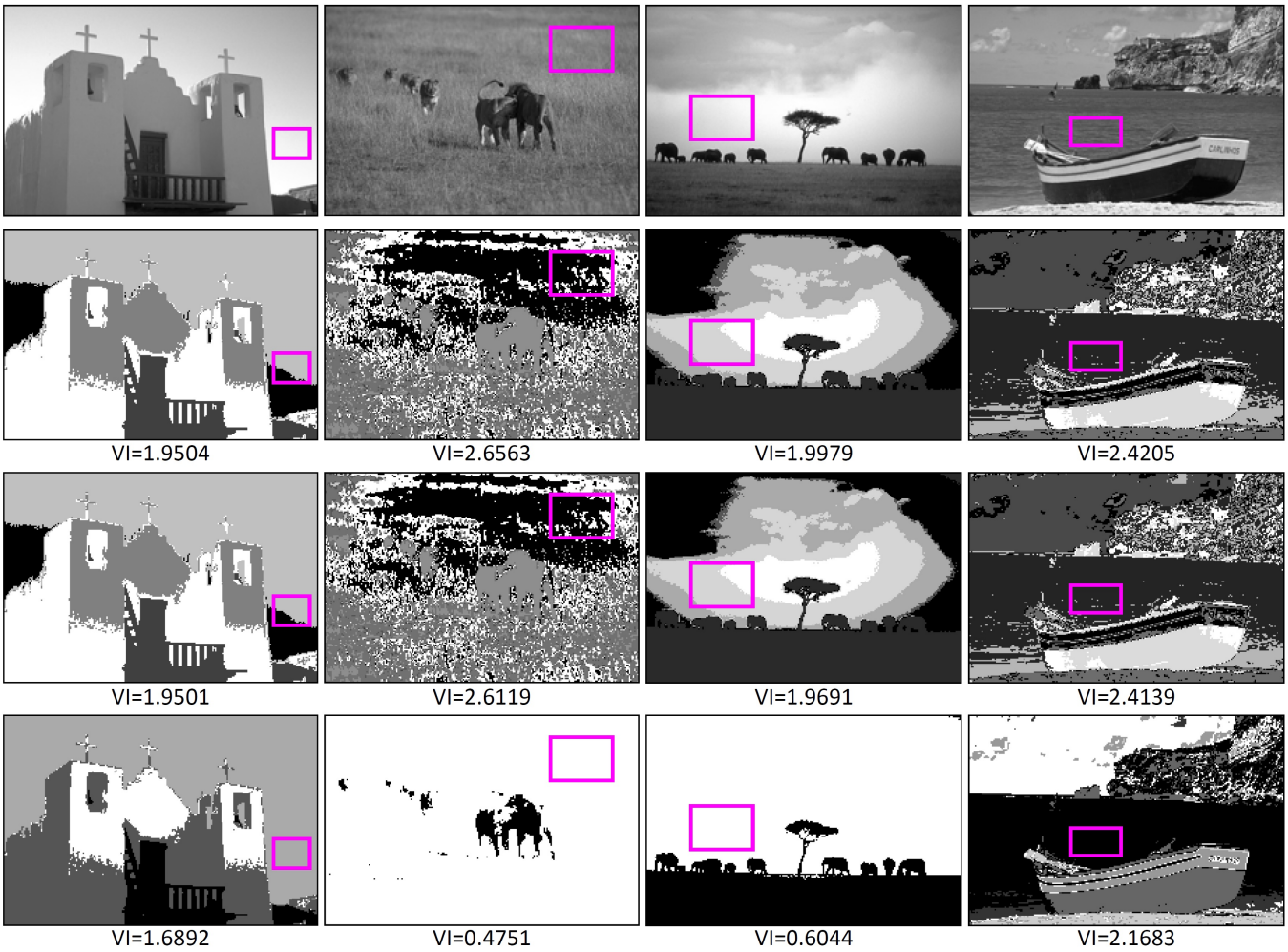

Fig. 3. Image segmentation, (1^{st} row): original image, (2^{nd} row): VB-GMM, (3^{rd} row): VB-SMM, (4^{th} row): Our method.

tion, and the number of components. We demonstrate through extensive simulations that the proposed model is superior to other mixture models.

REFERENCES

[1] McLachlan G., and Peel D., "*Finite Mixture Models*", Wiley, 2000.

[2] Jain A. K., Duin R. P. W., and Mao J., "Statistical Pattern Recognition: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[3] Peel D., and McLachlan G., "Robust Mixture Modeling Using the t Distribution," *Statistics and Computing*, vol. 10, pp. 339–348, 2000.

[4] Liu C., and Rubin D., "ML estimation of the t distribution using EM and its extensions, ECM and ECME," *Statistica Sinica*, vol. 5, no. 1, pp. 19–39, 1995.

[5] Chatzis S. P., Kosmopoulos D., and Varvarigou T., "Robust Sequential Data Modeling Using an Outlier Tolerant Hidden Markov Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1657–1669, 2009.

[6] Thanh M. N., and Wu Q. M. J., "Robust Student's-t Mixture Model with Spatial Constraints and Its Application in Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 103–116, 2012.

[7] Dempster P., Laird N. M., and Rubin D. B., "Maximum likelihood from incomplete data via EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[8] Attias H., "A Variational Bayesian Framework for Graphical Models," *Advances in Neural Information Processing Systems*, vol. 12. 2000.

[9] Bishop C. M., "*Pattern Recognition and Machine Learning*", Springer, 2006.

[10] Svensen M., and Bishop C. M, "Robust Bayesian Mixture Modelling," *Neurocomputing*, vol. 64, pp. 235–252, 2005.

[11] Van L. K. V., Maes F., Vandermeulen D., and Suetens P., "Automated model-based tissue classification of MR images of the brain," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 897–908, 1999.

[12] Thanh M. N., and Wu Q. M. J, "A Nonsymmetric Mixture Model for Unsupervised Image Segmentation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 43, no. 2, pp. 751–765, 2012.

[13] Corduneanu A., and Bishop C. M., "Variational Bayesian Model Selection for Mixture Distributions," *International Conference on Artificial Intelligence and Statistics*, pp. 27–34, 2001.

[14] Meila M., "Comparing Clusterings An Axiomatic View," *International Conference on Machine Learning*, pp. 577–584, 2005.