International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 27, No. 6 (2013) 1350016 (18 pages)

DOI: 10.1142/S021800141350016X

VARIATIONAL BAYES AND LOCALIZED FEATURE SELECTION FOR STUDENT'S t-MIXTURE MODELS

HUI ZHANG[*,†,‡], Q. M. JONATHAN WU[†,§]
and THANH MINH NGUYEN[†]

[*]*School of Computer & Software*
Nanjing University of Information Science & Technology, P. R. China

[†]*Department of Electrical and Computer Engineering*
University of Windsor, 401 Sunset Ave.
Windsor, ON, N9B 3P4, Canada

[‡]*Jiangsu Engineering Center of Network Monitoring*
Nanjing University of Information Science & Technology, P. R. China
[§]*jwu@uwindsor.ca*

Received 6 November 2012
Accepted 21 June 2013
Published 6 August 2013

In this paper, we propose a novel algorithm for feature selection and model detection using Student's t-distribution based on the variational Bayesian (VB) approach. First, our method is based on the Student's t-mixture model (SMM) which has heavier tail than the Gaussian distribution and is therefore less sensitive to small numbers of data points and consequent precision-estimates of the components number. Second, the number of components, the local feature saliency and the parameters of the mixture model are simultaneously estimated by Bayesian variational learning. Experimental results using synthetic and real data demonstrate the improved robustness of our approach.

Keywords: Bayesian approach; feature selection; Student's t-distributions; variational learning.

1. Introduction

Variational methods have their origins in the 18th century with the work of Euler, Lagrange, and other scientists focus on the calculus of variations.[4] To deal with the intractable integrations appearing in the Bayesian approach, the use of the variational approximation has been proposed.[3,9,14,22] This general optimization method called variational Bayes (VB) is a family of approximation technique which uses global criteria and which has been widely employed in a number of recent works. According to the variational methodology, an approximation to the posterior of the hidden variables given the observations is adopted. Based on this approximation,

[§]Corresponding author.

Bayesian inference is possible by maximizing a lower bound of the likelihood function with respect to the model parameters.[25] Attias employs Gaussian mixture model (GMM) and maximum likelihood approach with various applications of variational methods.[2] Corduneanu and Bishop use GMM and lower bound marginal likelihood to fit mixture numbers.[10] Student's t-mixture model (SMM) proposed by Peel and Mclachlan recently as an alternative to GMM has provided high robustness against noises.[20] The significant tolerance of SMM to training data has been experimentally depicted in many recent articles where it has been shown that SMM can model the hidden patterns of data significantly well.[6,7] Svensen and Bishop propose a Bayesian approach to SMM which is more robust to VB for GMM (VBGMM). VBGMM is a special case of VB for SMM (VBSMM)[23] which is less sensitive to outliers and can obtain robust estimates for the mean of a set of data points.

Methods mentioned above consider all available features with equal weight of the dataset. In practice, this cannot be always true as some features can be irrelevant. An effective solution is feature selection which chooses the best feature for clustering. Traditional feature selection methods for unsupervised learning can be divided into two categories: filters and wrappers.[17] The filter approaches consider the preselect features as an input without regarding the subsequent learning algorithm. Methods proposed by Dash *et al.*[11] and Mitra *et al.*[18] belong to this class. In Ref. 11, the clustering tendency of each feature is assessed by an entropy measure which is low if the data has distinct clusters and high otherwise. Then the relevant feature subset is classified by this measure. In Ref. 18, Mitra *et al.* introduces a feature similarity measure based on a maximum information compression index to detect redundant features. Other wrapper approaches use the clustering algorithm to evaluate the candidate feature subsets for feature searching and selection.[15] Feature selection method in Ref. 12 and its improvement[13] proposed by Dy and Brodley belong to the second class. In Ref. 12, the expectation-maximization clustering is used for multiple feature sets selection in two steps. The first step classifies the query according to the class labels of the images using features that best discriminate the classes. In the second step, it retrieves the most similar images using the features customized to distinguish "subclasses" within the predicted class. In Ref. 13, two different criteria: normalizing scatter separability (for k-means clustering) and maximum likelihood (for EM clustering) are used to evaluate the quality of clusters. Then, the cross-projection criterion normalization scheme is used to choose the feature selection criterion. The "optimal" number of clusters is found by Bayesian information criterion (BIC) algorithm.

Recently, Law *et al.*[15] defines feature saliency as a probability and assumes that these features are independent and are following a common distribution with Gaussian mixture models for clustering. The complement of this probability is that the measure of feature saliency and the estimation is carried out by the maximum likelihood (ML) or maximum *a priori* (MAP) with the EM algorithm. The number of components is determined by the minimal message length (MML) algorithm.

Constantinopoulos *et al.*[8] and Li *et al.*[16,17] adopt the same GMM as in Ref. 15 to describe the feature saliency, but the former employs VB algorithm instead of MML method and the latter proposes localized feature saliency measure instead of global feature approaches.

In this paper, a new VBSMM is introduced with the consideration of localized feature selection methods for unsupervised learning. The SMM is more robust than GMM for heavier tail and the distribution for each component in the SMM is embedded in a wider class of elliptically symmetric distributions with an additional parameter called the number of degrees of freedom. Moreover, selected features have found wide applications in areas ranging from business application, handwriting classification to texture analysis and image retrieval.[5,19,24,26] Based on previous work,[23] we incorporate feature selection into VBSMM to improve the performance of classifier by ignoring the noisy features. Feature selection is important for machine learning due to the fact that noisy features can degrade the performance of the classification algorithms. Feature selection can also improve the performance of classifiers learned from limited amounts of data. It leads to more economical (both in storage and computation) classifiers and, in many cases, it may lead to interpretable models.[21] In Sec. 2, we briefly review the mathematic background of necessary. The VBSMM for localized feature selection, model detection and parameter estimation are presented in Sec. 3. The experimental results are described in Secs. 4 and 5 concludes the paper.

2. Feature Selection for VBSMM

The Student's t-distribution provides a heavy-tailed alternative to the Gaussian distribution for potential outliers. The probability density function (pdf) of a Student's t-distribution with mean μ, precision τ and degrees of freedom v is[4]

$$t(x;\mu,\tau,v) = \frac{\Gamma(v/2+1/2)}{\Gamma(v/2)}\left(\frac{\tau}{\pi v}\right)^{1/2}\left[1+\frac{\tau(x-\mu)^2}{v}\right]^{-v/2-1/2}, \tag{1}$$

where $\Gamma(s)$ is the Gamma function,

$$\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt. \tag{2}$$

It can be shown that the Student's t-distribution $t(\mu,\tau,v)$ is equivalent to a Gaussian distribution $N(\mu,u\tau)$ with the precision scaling factor u which is given as:

$$x \,|\, u \sim N(\mu, u\tau), \tag{3}$$

where the scaling factor u is a Gamma-distributed latent variable, depending on the degree of freedom v, defined as:

$$u \sim G\left(\frac{v}{2},\frac{v}{2}\right), \tag{4}$$

the pdf of the Gamma distribution $G(u;\alpha,\beta)$ is given by

$$G(u;\alpha,\beta)=\frac{1}{\Gamma(\alpha)}\beta^{\alpha}u^{\alpha-1}e^{-\beta u}. \tag{5}$$

We now consider VB method as SMM with localized feature saliency through the binary latent variables z_{jn}, where $z_{jn}\in\{0,1\}$ and $\sum_{j=1}^{J}z_{jn}=1$. If a given data point x_{in} is generated from component j then the value of z_{jn} is one; otherwise, it is zero. The saliency of localized feature is expressed through the hidden variables s_{ijn}, where $s_{ijn}=1$ indicates that for data x_n, feature i is associated with component j, and $s_{ijn}=0$ otherwise. Then $w_{ij}=\Pr(s_{ijn}=1)$ is the probability that feature i is relevant to component j for all observed data X. The likelihood function is given as follows:

$$p(X|\Phi)=\prod_{n=1}^{N}\prod_{j=1}^{J}\left[\prod_{i=1}^{d}t(x_{in};\mu_{ij},\tau_{ij},v_{ij})^{s_{ijn}}t(x_{in};\varepsilon_{ij},\gamma_{ij},\eta_{ij})^{1-s_{ijn}}\right]^{z_{jn}}. \tag{6}$$

The sets $\{\mu_{ij}\}$, $\{\tau_{ij}\}$ and $\{v_{ij}\}$ denote the mean, precision and degrees of freedom of the "useful" subcomponents. Correspondingly, $\{\varepsilon_{ij}\}$, $\{\gamma_{ij}\}$ and $\{\eta_{ij}\}$ are the sets of parameters for the "noise" subcomponents. With the latent variables u_{ijn} and λ_{ijn}, Student's t-distribution can be equivalent to a Gaussian distribution as follows:

$$p(X|\Phi)=\prod_{n=1}^{N}\prod_{j=1}^{J}\left[\prod_{i=1}^{d}N(x_{in};\mu_{ij},u_{ijn}\tau_{ij})^{s_{ijn}}N(x_{in};\varepsilon_{ij},\lambda_{ijn}\gamma_{ij})^{1-s_{ijn}}\right]^{z_{jn}}. \tag{7}$$

The hidden variable Z is given the distribution governed by the mixing probabilities $\pi=\{\pi_j\}$ and the hidden variable S is given the distribution governed by the probabilities $w=\{w_{ij}\}$ as:

$$p(Z|\pi)=\prod_{n,j=1}^{N,J}\pi_j^{z_{jn}},\quad p(S|w)=\prod_{n,j,i=1}^{N,J,d}w_{ij}^{s_{ijn}}(1-w_{ij})^{1-s_{ijn}}.$$

The model selection is accomplished by introducing conjugate priors over the means, precisions and degrees of freedom for "useful" subcomponents

$$p(\mu)=\prod_{j=1}^{J}\prod_{i=1}^{d}N(\mu_{ij};m_i,c), \tag{8}$$

$$p(T)=\prod_{j=1}^{J}\prod_{i=1}^{d}G(\tau_{ij};\alpha,\beta), \tag{9}$$

$$p(U)=\prod_{n=1}^{N}\prod_{j=1}^{J}\prod_{i=1}^{d}G\left(u_{ijn};\frac{v_{ij}}{2},\frac{v_{ij}}{2}\right). \tag{10}$$

Here, the hyperparameters m_i,c,α,β control the prior distributions over μ_{ij} and are chosen to give broad prior for τ_{ij} with "useful" subcomponents. Moreover, the actual model parameters are not sensitive to these hyperparameters on the near zero

scale. It is noticed that there is no conjugate priority for v_{ij} since no priority is imposed on it. The learning approach used for the proposed model is discussed in the next part.

3. Variational Approximation

In order to obtain the estimation of parameters, we maximize the marginal likelihood $p(X)$ by integrating out the variables as follows:

$$p(X) = \int p(X, \Phi) d\Phi. \tag{11}$$

The integral denotes the joint integration over observed variables and the summation over hidden variables Z and S. Since the integration in the equation above is intractable, an alternative way to solve this problem is VB method which aims to maximize a lower bound L of the logarithmic marginal likelihood:

$$L(Q) = \int Q(\Phi) \log \frac{p(X, \Phi)}{Q(\Phi)} d\Phi \leq \log p(X), \tag{12}$$

where Q is an arbitrary distribution which provides an approximation to the true posterior distribution. We see that the function $L(Q)$ forms a rigorous lower bound on the true log marginal likelihood. Although the computation of original log likelihood function $\log p(X)$ is not tractable, the lower bound $L(Q)$ may be tractable to compute through choosing a suitable form for the Q distribution. The difference between the lower bound $L(Q)$ and the true log likelihood $\log p(X)$ is 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.

By minimizing the KL divergence with respect to all possible function form of Q,] the standard variational approach provides the following general form of the solutions

$$Q(\Phi_i) = \frac{\exp\langle p(X, \Phi)\rangle_{k \neq i}}{\int \exp\langle p(X, \Phi)\rangle_{k \neq i} d\Phi_i}, \tag{13}$$

where $\langle \cdot \rangle_{k \neq i}$ denotes an expectation with respect to the distributions $Q_k(\Phi_k)$ for all $k \neq i$. The factors of the variational posterior are given by calculation of (13) as follows:

$$Q(Z) = \prod_{n=1}^{N} \prod_{j=1}^{J} r_{jn}^{z_{jn}}, \tag{14}$$

$$Q(\mu) = \prod_{j=1}^{J} \prod_{i=1}^{d} N(\mu_{ij}; m_{ij}^{\mu}, c_{ij}^{\mu}), \tag{15}$$

$$Q(T) = \prod_{j=1}^{J}\prod_{i=1}^{d} G(\tau_{ij}; \alpha_{ij}^{\tau}, \beta_{ij}^{\tau}), \tag{16}$$

$$Q(U) = \prod_{n=1}^{N}\prod_{j=1}^{J}\prod_{i=1}^{d} G(u_{ijn}; \alpha_{ijn}^{u}, \beta_{ijn}^{u}), \tag{17}$$

$$Q(S) = \prod_{n=1}^{N}\prod_{j=1}^{J}\prod_{i=1}^{d} \rho_{ijn}^{s_{ijn}}(1-\rho_{ijn})^{1-s_{ijn}}. \tag{18}$$

The variational parameters are given by maximizing and determining the density involved in Q

$$r_{jn} = \frac{\tilde{r}_{jn}}{\sum_{j=1}^{J}\tilde{r}_{jn}}, \tag{19}$$

$$\tilde{r}_{jn} = \pi_j \exp\left(\frac{1}{2}\sum_{i=1}^{d}\langle s_{ijn}\rangle\{\langle\log\tau_{ij}\rangle + \langle\log u_{ijn}\rangle - \langle(x_{in}-\mu_{ij})^2\rangle\langle u_{ijn}\rangle\langle\tau_{ij}\rangle\}\right), \tag{20}$$

$$m_{ij}^{\mu} = \frac{cm_i + \langle\tau_{ij}\rangle\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle\langle u_{ijn}\rangle x_{in}}{c + \langle\tau_{ij}\rangle\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle\langle u_{ijn}\rangle}, \tag{21}$$

$$c_{ij}^{\mu} = c + \langle\tau_{ij}\rangle\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle\langle u_{ijn}\rangle, \tag{22}$$

$$\alpha_{ij}^{\tau} = \alpha + \frac{1}{2}\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle, \tag{23}$$

$$\beta_{ij}^{\tau} = \beta + \frac{1}{2}\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle\langle u_{ijn}\rangle\langle(x_{in}-\mu_{ij})^2\rangle, \tag{24}$$

$$\alpha_{ijn}^{u} = \frac{v_{ij} + \langle z_{jn}\rangle\langle s_{ijn}\rangle}{2}, \tag{25}$$

$$\beta_{ijn}^{u} = \frac{v_{ij} + \langle z_{jn}\rangle\langle s_{ijn}\rangle\langle(x_{in}-\mu_{ij})^2\rangle\langle\tau_{ij}\rangle}{2}, \tag{26}$$

$$\rho_{ijn} = \frac{w_{ij}\tilde{\rho}_{ijn}}{w_{ij}\tilde{\rho}_{ijn} + (1-w_{ij})\xi_{ijn}}, \tag{27}$$

$$\tilde{\rho}_{ijn} = \exp\left\{\frac{1}{2}\langle z_{jn}\rangle[\langle\log\tau_{ij}\rangle + \langle\log u_{ijn}\rangle - \langle(x_{in}-\mu_{ij})^2\rangle\langle u_{ijn}\rangle\langle\tau_{ij}\rangle]\right\}, \tag{28}$$

$$\xi_{ijn} = \exp\left\{\frac{1}{2}\langle z_{jn}\rangle[\log\gamma_{ij} + \log\lambda_{ijn} - (x_{in}-\varepsilon_{ij})^2\lambda_{ijn}\gamma_{ij}]\right\}. \tag{29}$$

The proofs of (19)−(29) are shown in appendix. Since no conjugate priority is imposed on the degrees of freedom, we update v_{ij} with the log-marginal ML estimates,

by setting the corresponding gradient to zero and solving the nonlinear equations as follows:

$$\log\left(\frac{v_{ij}}{2}\right) - \psi\left(\frac{v_{ij}}{2}\right) + 1 + \frac{\sum_n \langle z_{jn}\rangle \langle s_{ijn}\rangle [\langle \log u_{ijn}\rangle - \langle u_{ijn}\rangle]}{\sum_n \langle z_{jn}\rangle \langle s_{ijn}\rangle} = 0, \tag{30}$$

where $\psi(x)$ is the digamma function $\psi(x) = d\log\Gamma(x)/dx$. Degrees of freedom decides the shape of Student's distribution. It tends to a Gaussian distribution for large values of its degrees of freedom and takes in to account the outliers with heavier tails for small value. Compared with previous methods based on GMM, our method needs additional iterative computation for the degrees of freedom v_{ij}.

The expected values in the above formulas are given by:

$$\langle z_{jn}\rangle = r_{jn}, \tag{31}$$

$$\langle \mu_{ij}\rangle = m_{ij}^{\mu}, \tag{32}$$

$$\langle (x_{in} - \mu_{ij})^2\rangle = (x_{in} - m_{ij}^{\mu})^2 + \frac{1}{c_{ij}^{\mu}}, \tag{33}$$

$$\langle \tau_{ij}\rangle = \frac{\alpha_{ij}^{\tau}}{\beta_{ij}^{\tau}}, \tag{34}$$

$$\langle \ln \tau_{ij}\rangle = \psi(\alpha_{ij}^{\tau}) - \log \beta_{ij}^{\tau}, \tag{35}$$

$$\langle u_{ijn}\rangle = \frac{\alpha_{ijn}^{u}}{\beta_{ijn}^{u}}, \tag{36}$$

$$\langle \ln u_{ijn}\rangle = \psi(\alpha_{ijn}^{u}) - \log \beta_{ijn}^{u}, \tag{37}$$

$$\langle s_{ijn}\rangle = \rho_{ijn}. \tag{38}$$

With the variational factors in (14)−(18) and estimated variational parameters in (19)−(38), we can maximize the tight bound of the log marginal likelihood which equals to the minimization of the KL divergence. It is noticed that the variational factors are mutually coupled through their dependence on moments of the other factors, these parameters can be achieved by optimizing each factor in turn, holding the others fixed, with iteratively updating. We first initialize the approximating distributions, cycle round each factor in turn and then replace its current estimate by its optimal solution given the current estimates for the other factors.

It is noticed that the solutions for the factors of the variational posterior, and hence the value of the lower bound, depended on the values π_j, w_{ij}, ε_{ij}, and γ_{ij}. These parameters are obtained by setting the derivative of lower bound L with respect to π_j, w_{ij}, ε_{ij}, and γ_{ij} equal to zero,

$$\begin{aligned} L = {} & \langle \log p(X|\Phi)\rangle + \langle \log p(Z)\rangle + \langle \log p(\mu)\rangle + \langle \log p(T)\rangle + \langle \log p(U)\rangle + \langle \log p(S)\rangle \\ & - \langle \log Q(Z)\rangle - \langle \log Q(\mu)\rangle - \langle \log Q(T)\rangle - \langle \log Q(U)\rangle - \langle \log Q(S)\rangle \end{aligned} \tag{39}$$

which leads to the following updated rules

$$\pi_j = \frac{1}{N}\sum_{n=1}^{N} r_{jn}, \tag{40}$$

$$w_{ij} = \frac{1}{N}\sum_{n=1}^{N} \rho_{ijn}, \tag{41}$$

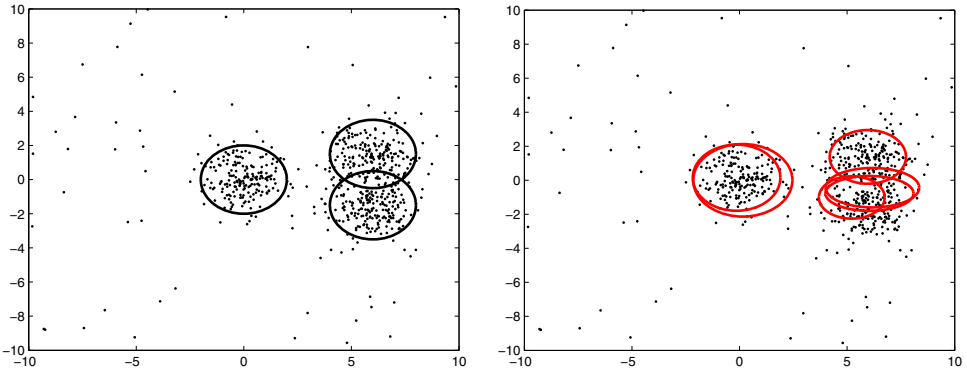
$$\varepsilon_{ij} = \frac{\sum_{n=1}^{N}(1-\rho_{ijn})r_{jn}\lambda_{ijn}x_{in}}{\sum_{n=1}^{N}(1-\rho_{ijn})r_{jn}\lambda_{ijn}}, \tag{42}$$

$$\frac{1}{\gamma_{ij}} = \frac{\sum_{n=1}^{N}(1-\rho_{ijn})r_{jn}\lambda_{ijn}(x_{in}-\varepsilon_{ij})^2}{\sum_{n=1}^{N}(1-\rho_{ijn})r_{jn}}. \tag{43}$$

The proofs of (40)−(43) are shown in appendix. The previous equations are repeated alternatively until convergence, which can be monitored through inspection of the variational lower bound. The lower bound value does not decrease with the increasing of the iterative times. Here, parameters π_j and w_{ij} can be considered as "contribution" parameters for cluster numbers and feature saliencies. The property of VB algorithm guaranties that the components with similar parameters fitting the same Student's t-distributions generating a dominant cluster. Thus, we can start the model with a large number of components; the competition among components finally yields a model which is composed by dominant components and the redundant components are removed. Meanwhile, feature saliency determines the subcomponent to be a "useful" subcomponent with value one or "noise" subcomponent with value zero. The update of the parameters π_j and w_{ij} identify the cluster numbers and feature saliencies simultaneously.

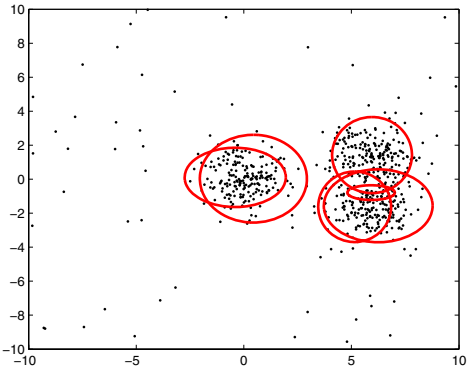
4. Experimental Results

In this section, we experimentally evaluate our method (feature selection and variational Bayes for Student's t-mixture model, FSVB-SMM) in a set of synthetic data and real data. For comparison, we also evaluate the method in Ref. 23 (VBSMM) and the method in Ref. 17 (feature selection and variational Bayes for Gaussian mixture model, FSVB-GMM). The synthetic data set contains 600 data points from a mixture of three Gaussian components, where the component mean is $m_1 = (6, -1.5),\ m_2 = (6, 1.5),\ m_3 = (0, 0)$ and covariance is $\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ (shown in Fig. 1(a)). Here, each component contains 200 data points. These Gaussians are then embedded into subsets of 10D background with standard Gaussian noise. Components can have a different number of features for the feature saliency test. Thus, we select features 1 and 3 from the background data and then replace the first 200 positions with component 1. Similarly, we embed component 2 with features 4 and 5, and component 3 with features 2 and 5 into the background. The test data

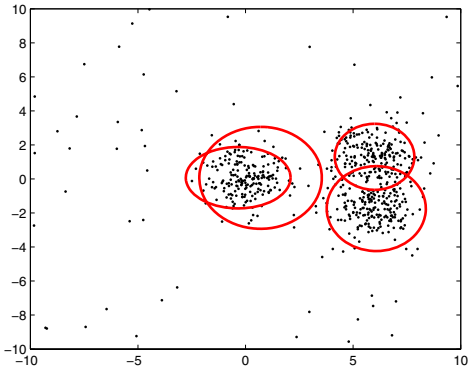


(a) True distribution

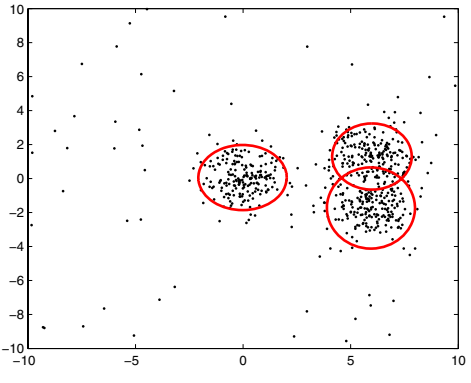
(b) Iteration = 30



(c) Iteration = 60

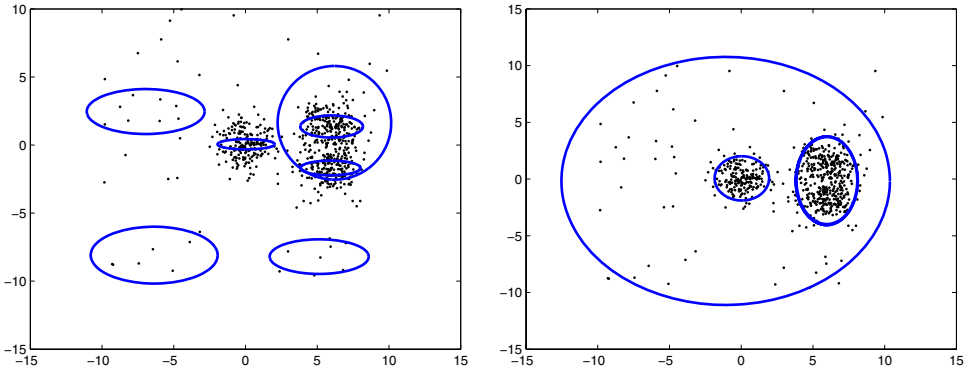


(d) Iteration = 90



(e) Final Step, Iteration = 200

Fig. 1. Configuration for the synthetic data by proposed method. (a) Original data distribution. (b) Proposed method with iteration = 30. (c) Proposed method with iteration = 60. (d) Proposed method with iteration = 90. (e) Proposed method with iteration = 200 (color online).



(a) (b)

Fig. 2. Final configuration for the synthetic data by previous methods. (a) Method in Ref. 23, VBSMM. (b) Method in Ref. 17, FSVB-GMM (color online).

is further augmented by 10% outliers, which are drawn from a uniform distribution on the interval $[-10, 10]$. All 10D data cannot be shown on a 2D display, thus we selected data with "useful" features for effective display. The results for proposed method are given in Figs. 1(b)−1(e). The results for VBSMM and FSVB-GMM are shown in Figs. 2(a) and 2(b), respectively. It can be seen that FSVB-SMM can give the right components and estimate the mean and covariance values well, and is less affected by the "noise" features and outliers. For deep investigation, we give the average mean values estimation in 10 run-times for the three methods in Table 1, where we can see that proposed method outperforms other methods. Our algorithm yields a nearly perfect result, with the estimated mean values almost identical to the actual mean values of the modeled data distributions.

In each iteration, the computational complexity of the proposed algorithm is $O(NJD)$. The total computation time depends on the number of iterations to converge. We compute quantities in (18)−(30) and (40)−(43) in each iteration. The computation complexity to calculate ξ_{ijn} is $O(1)$. There are total (NJD) ξ s so that it requires $O(NJD)$. Similarity, computing ρ_{ijn} and $\tilde{\rho}_{ijn}$ requires $O(NJD)$. Computing m_{ij}^{μ} requires to navigate through all the samples, resulting in the complexity $O(NJD)$. Similar results are obtained to calculate c_{ij}^{μ}, α_{ij}^{τ}, β_{ij}^{τ}, α_{ijn}^{u}, β_{ijn}^{u} and $\tilde{r}_{jn}$. The computation complexity for w_{ij}, ε_{ij} and γ_{ij} is also $O(NJD)$. In summary, the overall computational complexity for one iteration is $O(NJD)$ to calculate these 13

Table 1. Estimated mean value for different methods.

	Real Mean	VBSMM	FSVB-GMM	FSVB-SMM
Component 1	[6 −1.5]	[6.14 −1.18]	[5.97 −0.22]	[5.96 −1.62]
Component 2	[6 1.5]	[6.21 1.14]	[0.74 −0.21]	[6.02 1.43]
Component 3	[0 0]	[−0.01 0.01]	[0.05 0.06]	[0.03 0.05]
CPU time (s)	—	89.37	42.96	116.4

quantities. However, the computational complexity is 10 $O(NJD)$ in FSVB-GMM. Moreover, we need to compute additional Eq. (30) for degrees of freedom v_{ij} compared with FSVB-GMM. Thus, the computational complexity of proposed algorithm is heavier than FSVB-GMM. We provide the computation time for three different methods in this experiment (last line of Table 1). Our experiment has been developed in Matlab R2009b, and is executed on an Intel Pentium Dual-Core 2.2GHZ CPU, 2G RAM. Our algorithm is slowest compared with other two methods. This is the limitation of our algorithm.

The second experiment is implemented with four real data sets, Wine, Heart, WPBC and Yeast, available from the well-known UCI machine learning repository.[1] After the normalization of UCI data, we add standard Gaussian noise (mean = 0, covariance = 0.02) to the available data to make the classification task more challenging. Here, all methods use the same initialization with the common k-means algorithm for 500 runs and start with initial 20 components. The error rate is calculated according to the samples in true class labels, and the estimated samples in evaluated components. In our experiment, it is calculated based on the average of 5 random runs and the lower value indicates better performance. The mean and standard deviation of the error rate, and the class components over 5 runs for different methods are shown in Table 2. Without feature selection, VBSMM is affected from the initial number of components and tends to keep most of them. Therefore, VBSMM cannot estimate the correct component number and thus leads to the higher error rates. With feature selection, FSVB-GMM and our proposed algorithm FSVB-SMM estimate the component number J well. Moreover, it is clear to see that FSVB-SMM outperforms FSVB-GMM, with lower error rates and variances. The high error rate for the Yeast data set is because Proteins are inherently more difficult to cluster. Another reason is that some classes contain only 20 samples, and some even fewer, at five samples of total 1484 samples, which makes the classification more difficult.

In the third experiment, we change the initial number of components for all methods using UCI data sets test. The error rate is calculated based on the average of 10 random runs. The lower error rate value indicates better performance. Figure 3 shows the error rate versus the different initial number of components for all three data sets. It can be seen from the figure that all methods obtain optimal results when the initial component is equal to the real class (Wine is 3, Heart is 2, WPBC is 2 and

Table 2. Average error rate and estimated components.

	VBSMM		FSVB-GMM		FSVB-SMM	
	Error rate (%)	Estimated J	Error rate (%)	Estimated J	Error rate (%)	Estimated J
Wine	85.39 (2.45)	18.28 (0.73)	7.4 (1.96)	5.2 (0.45)	6.2 (0.78)	3.2 (0.45)
Heart	85.56 (3.51)	17.46 (1.31)	43.85 (9.12)	5 (0.71)	37.38 (4.35)	4.8 (0.84)
WPBC	82.52 (2.62)	15.17 (1.27)	38.28 (5.77)	2.8 (0.84)	34.6 (1.97)	2.6 (0.55)
Yeast	68.67 (1.19)	18.58 (0.86)	64.28 (1.50)	5.8 (1.10)	61.32 (4.73)	7.4 (1.14)

Note: The numbers in parenthesis are the standard deviations of the corresponding quantities.

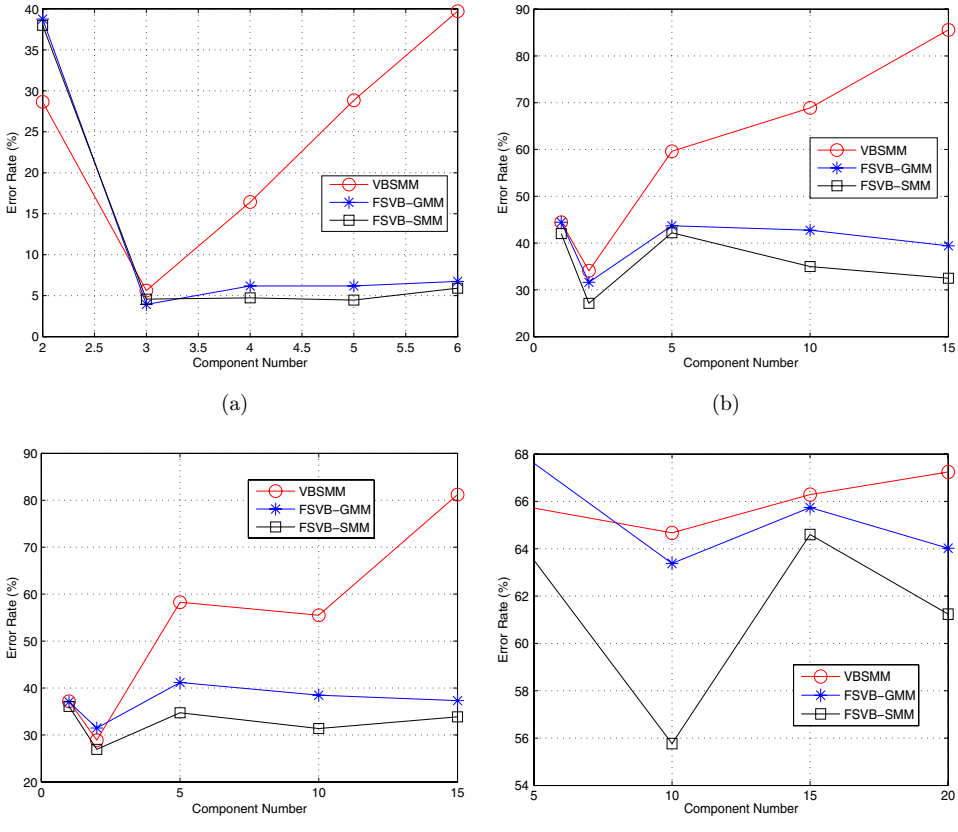


Fig. 3. Error rates with the different initial number of components for all methods. (a) Wine, (b) heart, (c) WPBC, (d) yeast (color online).

yeast is 10). The error rate of all methods increases with the increasing of the initial component. However, FSVB-SMM and FSVB-GMM are more robust than VBSMM. The error rate of VBSMM increases significantly when the initial number of components increases. On the contrary, FSVB-GMM and FSVB-SMM perform more steadily with the variation of initial component numbers. This may be due to the property of localized feature selection. Although this is not obviously observed in Yeast data set due to the high error rate for all methods, the proposed method still performs better than VBSMM for component estimation.

5. Conclusion

In this paper, we propose a new FSVB-SMM algorithm which incorporates the feature selection methodology into traditional VBSMM model. The parameters of SMM, the number of components and the localized feature saliency can be automatically determined by the constructed model. The experiments conducted under synthetic

and real dates demonstrate the discriminative power of the proposed descriptors, especially in noise conditions, which clearly outperform existing methods.

Acknowledgments

The authors would like to thank Dr. Constantinopoulos for offering some Matlab codes. This work was supported in part by the Canada Chair Research Program and the Natural Sciences and Engineering Research Council of Canada. This work was supported in part by the National Natural Science Foundation of China under Grant 61105007 and 61103141. This work was supported in part by PAPD (A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions).

Appendix

With the help of (13), we have

$$\begin{aligned}
\log Q(Z) &= \left\langle \log \left[\prod_{n=1}^{N} \prod_{j=1}^{J} \left[\prod_{i=1}^{d} N(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij})^{s_{ijn}} \right]^{z_{jn}} \pi_j^{z_{jn}} \right] \right\rangle_{\mu,\tau,s,u} + \text{const} \\
&= \sum_{n,j} z_{jn} \Bigg\{ \log \pi_j + \sum_i \frac{1}{2} \langle s_{ijn} \rangle [\langle \log \tau_{ij} \rangle + \langle \log u_{ijn} \rangle \\
&\quad - \langle (x_{in} - \mu_{ij})^2 \rangle \langle u_{ijn} \rangle \langle \tau_{ij} \rangle] \Bigg\} + \text{const} \\
&= \sum_{n,j} z_{jn} \log \tilde{r}_{jn}
\end{aligned}$$

which eventually yields (19) and (20).

Applying (15) to (13), we have

$$\begin{aligned}
\log Q(\mu) &= \left\langle \log \left[\prod_{i=1}^{d} \prod_{j=1}^{J} \left[\prod_{n=1}^{N} N(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij})^{s_{ijn} z_{jn}} \right] \right. \right. \\
&\quad \left. \left. \times N(\mu_{ij}; m_i, c) \right] \right\rangle_{z,\tau,s,u} + \text{const} \\
&= \sum_{i,j} \left\{ \sum_n \langle z_{jn} \rangle \langle s_{ijn} \rangle \langle \log N(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij}) \rangle + \log N(\mu_{ij}; m_i, c) \right\} + \text{const} \\
&= \sum_{i,j} \left\{ -\frac{1}{2} \sum_n \langle z_{jn} \rangle \langle s_{ijn} \rangle \langle u_{ijn} \rangle (x_{in} - \mu_{ij})^2 \langle \tau_{ij} \rangle - \frac{1}{2} (\mu_{ij} - m_i)^2 c \right\} + \text{const} \\
&= \sum_{i,j} \left\{ -\frac{1}{2} (\mu_{ij} - m_{ij}^{\mu})^2 c_{ij}^{\mu} \right\} + \text{const} \\
&= \sum_{i,j} \log N(\mu_{ij}; m_{ij}^{\mu}, c_{ij}^{\mu}).
\end{aligned}$$

Concerning the items with respect to μ_{ij}, the proof of (21) and (22) is completed.

Similarly, applying (16) to (13), we have

$$\begin{aligned}
\log Q(T) &= \left\langle \log \left[\prod_{i=1}^{d} \prod_{j=1}^{J} \left[\prod_{n=1}^{N} N(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij})^{s_{ijn} z_{jn}} \right] \right. \right. \\
&\quad \left. \left. \times G(\tau_{ij}; \alpha, \beta) \right] \right\rangle_{z,\mu,s,u} + \text{const} \\
&= \sum_{i,j} \left\{ \sum_{n} \langle z_{jn} \rangle \langle s_{ijn} \rangle \langle \log N(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij}) \rangle \right. \\
&\quad \left. + (\alpha - 1) \log \tau_{ij} - \beta \tau_{ij} \right\} + \text{const} \\
&= \sum_{i,j} \left\{ \sum_{n} \langle z_{jn} \rangle \langle s_{ijn} \rangle \left[\frac{1}{2} \log \tau_{ij} + \frac{1}{2} \langle \log u_{ijn} \rangle - \frac{1}{2} \langle (x_{in} - \mu_{ij})^2 \rangle \langle u_{ijn} \rangle \tau_{ij} \right] \right. \\
&\quad \left. + (\alpha - 1) \log \tau_{ij} - \beta \tau_{ij} \right\} + \text{const} \\
&= \sum_{i,j} (\alpha_{ij}^{\tau} - 1) \log \tau_{ij} - \beta_{ij}^{\tau} \tau_{ij} + \text{const} \\
&= \sum_{i,j} \log G(\tau_{ij}; \alpha_{ij}^{\tau}, \beta_{ij}^{\tau}).
\end{aligned}$$

Thus α_{ij}^{τ} and β_{ij}^{τ} have the form of (23) and (24), respectively.

With respect to latent variable U, we have

$$\begin{aligned}
\log Q(U) &= \left\langle \log \left[\prod_{i=1}^{d} \prod_{j=1}^{J} \prod_{n=1}^{N} N(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij})^{s_{ijn} z_{jn}} \right. \right. \\
&\quad \left. \left. \times G\left(u_{ijn}; \frac{v_{ij}}{2}, \frac{v_{ij}}{2} \right) \right] \right\rangle_{z,\mu,s,\tau} + \text{const} \\
&= \sum_{i,j,n} \left\{ \langle z_{jn} \rangle \langle s_{ijn} \rangle \langle \log N(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij}) \rangle \right. \\
&\quad \left. + \left(\frac{v_{ij}}{2} - 1 \right) \log u_{ijn} - \frac{v_{ij}}{2} u_{ijn} \right\} + \text{const} \\
&= \sum_{i,j,n} \left\{ \langle z_{jn} \rangle \langle s_{ijn} \rangle \left[\frac{1}{2} \langle \log \tau_{ij} \rangle + \frac{1}{2} \log u_{ijn} - \frac{1}{2} \langle (x_{in} - \mu_{ij})^2 \rangle \langle \tau_{ij} \rangle u_{ijn} \right] \right. \\
&\quad \left. + \left(\frac{v_{ij}}{2} - 1 \right) \log u_{ijn} - \frac{v_{ij}}{2} u_{ijn} \right\} + \text{const}
\end{aligned}$$

$$= \sum_{i,j,n} (\alpha_{ij}^u - 1) \log u_{ijn} - \beta_{ij}^u u_{ijn} + \text{const}$$

$$= \sum_{i,j,n} \log G(u_{ijn}; \alpha_{ij}^u, \beta_{ij}^u)$$

which yields (25) and (26).

For hidden variable S for localized feature saliency, we obtain

$$\begin{aligned}
\log Q(S) &= \left\langle \log \left[\prod_{i=1}^{d} \prod_{j=1}^{J} \prod_{n=1}^{N} \left[N(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij})^{s_{ijn}} N(x_{in}; \varepsilon_{ij}, \lambda_{ijn}\gamma_{ij})^{1-s_{ijn}} \right]^{z_{jn}} \right. \right. \\
&\quad \left. \left. \times w_{ij}^{s_{ijn}} (1 - w_{ij})^{1-s_{ijn}} \right] \right\rangle_{z,\mu,\tau,u} + \text{const} \\
&= \sum_{i,j,n} \langle \log \{ [N(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij})^{z_{jn}} w_{ij}]^{s_{ijn}} \\
&\quad \times [N(x_{in}; \varepsilon_{ij}, \lambda_{ijn}\gamma_{ij})^{z_{jn}} (1 - w_{ij})]^{1-s_{ijn}} \} \rangle + \text{const} \\
&= \sum_{i,j,n} \left\{ \begin{array}{l} s_{ijn} \left[\langle z_{jn} \rangle \dfrac{1}{2} [\langle \log \tau_{ij} \rangle + \langle \log u_{ijn} \rangle \right. \\ \quad \left. - \langle (x_{in} - \mu_{ij})^2 \rangle \langle u_{ijn} \rangle \langle \tau_{ij} \rangle] + \log w_{ij} \right] \\ + (1 - s_{ijn}) \left[\langle z_{jn} \rangle \left[\dfrac{1}{2} \log \gamma_{ij} + \dfrac{1}{2} \log \lambda_{ijn} \right. \right. \\ \quad \left. \left. - \dfrac{1}{2} (x_{in} - \varepsilon_{ij})^2 \lambda_{ijn} \gamma_{ij} \right] + \log(1 - w_{ij}) \right] \end{array} \right\} + \text{const} \\
&= \sum_{i,j,n} s_{ijn} [\log w_{ij} + \log \tilde{\rho}_{ijn}] + (1 - s_{ijn}) [\log(1 - w_{ij}) + \log \xi_{ijn}]
\end{aligned}$$

which yields (27)−(29).

With the help of (19), (ignore irrelative items)

$$\frac{\partial L}{\partial \pi_j} = \frac{\partial \langle \sum_{n,j} z_{jn} \log \pi_j + \varphi(\sum_j \pi_j - 1) \rangle}{\partial \pi_j} = \sum_n \frac{r_{jn}}{\pi_j} + \varphi = 0.$$

Thus, we have

$$\pi_j = \frac{\sum_n r_{jn}}{N}.$$

Derivation of (41):

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial \{ \sum_{n,i,j} [\rho_{ijn} \log w_{ij} + (1 - \rho_{ijn}) \log(1 - w_{ij})] \}}{\partial w_{ij}} = \sum_n \left[\frac{\rho_{ijn}}{w_{ij}} - \frac{(1 - \rho_{ijn})}{(1 - w_{ij})} \right] = 0.$$

Thus, $\sum_n \rho_{ijn} = w_{ij} N$ yields (41).

Considering the "noise" subcomponents ε_{ij} and γ_{ij},

$$\frac{\partial L}{\partial \varepsilon_{ij}} = \frac{\partial \left[\sum_n \langle \log t(x_{in}; \varepsilon_{ij}, \gamma_{ij}, \eta_{ij})^{(1-s_{ijn})z_{jn}} \rangle + \text{const}\right]}{\partial \varepsilon_{ij}}$$

$$= -\sum_n r_{jn}(1-\rho_{ijn})(x_{in}-\varepsilon_{ij})\gamma_{ij}\lambda_{ijn} = 0.$$

With the simplification, we have

$$\sum_n r_{jn}(1-\rho_{ijn})(x_{in}-\varepsilon_{ij})\lambda_{ijn} = 0$$

which yields (42).

Similarity, we have

$$\frac{\partial L}{\partial \gamma_{ij}} = \frac{\partial \left[\sum_n \langle \log t(x_{in}; \varepsilon_{ij}, \gamma_{ij}, \eta_{ij})^{(1-s_{ijn})z_{jn}} \rangle + \text{const}\right]}{\partial \gamma_{ij}}$$

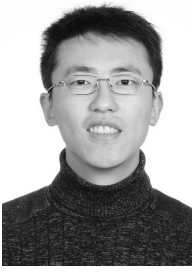
$$= \sum_n r_{jn}(1-\rho_{ijn})\left\{\frac{1}{2\gamma_{ij}} - \frac{1}{2}(x_{in}-\varepsilon_{ij})^2\lambda_{ijn}\right\} = 0,$$

which yields (43).

References

1. A. Asuncion and D. Newman, UCI machine learning repository, Available at http://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007 (last accessed 29 July 2013).
2. H. Attias, Inferring parameters and structure of latent variable models by variational Bayes, in *Proc. 15th Conf. Uncertainty in Artificial Intelligence*, 1999, pp. 21−30.
3. H. Attias, A variational Bayesian framework for graphical models, *Advances in Neural Information Processing Systems 12* (MIT Press, Cambridge, MA, 2000), pp. 209−215.
4. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, NY, 2006).
5. R. C. Blattberg and N. J. Gonedes, A comparison of the stable and student distributions as statistical models for stock prices, *J. Business* **47**(2) (1974) 244−280.
6. S. P. Chatzis, D. I. Kosmopoulos and T. A. Varvarigou, Signal modeling and classification using a robust latent space model based on *t*-distributions, *IEEE Trans. Signal Process.* **56**(3) (2008) 949−963.
7. S. P. Chatzis, D. I. Kosmopoulos and T. A. Varvarigou, Robust sequential data modeling using an outlier tolerant hidden markov model, *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(9) (2009) 1657−1669.
8. C. Constantinopoulos, M. K. Titsias and A. Likas, Bayesian feature and model selection for Gaussian mixture models, *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(6) (2006) 1013−1018.
9. C. Constantinopoulos and A. Likas, Unsupervised learning of Gaussian mixtures based on variational component splitting, *IEEE Trans. Neural Networks* **18**(3) (2007) 745−755.
10. A. Corduneanu and C. M. Bishop, Variational Bayesian model selection for mixture distributions, *Artificial Intelligence and Statistics 2001* (Morgan Kaufmann, San Mateo, CA, 2001), pp. 27−34.

11. M. Dash and H. Liu, Feature selection for clustering, in *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining*, 2000.
12. J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick and A. M. Aisen, Unsupervised feature selection applied to content-based retrieval of lung images, *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(3) (2003) 373–378.
13. J. G. Dy and C. E. Brodley, Feature selection for unsupervised learning, *J. Mach. Learning Res.* **5**(8) (2004) 845–889.
14. Z. Ghahramani and M. Beal, Variational inference for Bayesian mixtures of factor analysers, *Advances in Neural Information Processing Systems 12* (MIT Press, Cambridge, MA, 2000), pp. 449–455.
15. M. H. Law, M. A. T. Figueiredo and A. K. Jain, Simultaneous feature selection and clustering using mixture models, *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(9) (2004) 1154–1166.
16. Y. H. Li, M. Dong and J. Hua, Localized feature selection for clustering, *Pattern Recogn. Lett.* **29**(1) (2008) 10–18.
17. Y. H. Li, M. Dong and J. Hua, Simultaneous localized feature selection and model detection for Gaussian mixtures, *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(5) (2009) 953–960.
18. P. Mitra and C. A. Murthy, Unsupervised feature selection using feature similarity, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(3) (2002) 301–312.
19. P. Paalanen, J. K. Kamarainen, J. Ilonen and H. Kalviainen, Feature representation and discrimination based on Gaussian mixture model probability densities-practices and algorithms, *Pattern Recogn.* **39**(7) (2006) 1346–1358.
20. D. Peel and G. McLachlan, Robust mixture modeling using the *t*-distribution, *Statist. Comput.* **10**(4) (2000) 335–344.
21. S. J. Raudys and A. K. Jain, Small sample size effects in statistical pattern recognition: Recommendations for practitioners, *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(3) (1991) 252–264.
22. M. Sato, On-line model selection based on the variational Bayes, *Neural Comput.* **13**(7) (2001) 1649–1681.
23. M. Svensen and C. M. Bishop, Robust Bayesian mixture modelling, *Neurocomputing* **64** (2005) 235–252.
24. Y. Y. Tang, L.-T. Tu, J. Liu, S.-W. Lee, W.-W. Lin and I.-S. Shyu, Offline recognition of Chinese handwriting by multifeature and multilevel classification, *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(5) (1998) 556–561.
25. D. G. Tzikas, A. C. Likas and N. P. Galatsanos, The variational approximation for Bayesian inference: Life after the EM algorithm, *IEEE Signal Process. Mag.* **19**(6) (2008) 131–146.
26. H. Zhenyu, Y. Xinge and Y. Yuan, Texture image retrieval based on non-tensor product wavelet filter banks, *Signal Process.* **89**(8) (2009) 1501–1510.



Hui Zhang received his B.S. degree in Radio Engineering, his M.S. degree in Biomedical Engineering, and the Ph.D. all from Southeast University, Nanjing, China, in 2003, 2006, and 2010, respectively. He is currently a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, Canada. His research is mainly focused on pattern recognition, machine learning, image segmentation and image processing.

Q. M. Jonathan Wu received his Ph.D. in Electrical Engineering from the University of Wales, Swansea, U.K., in 1990. He was with the National Research Council of Canada for ten years from 1995, where he became a Senior Research Officer and a Group Leader. He is currently a Professor at the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He has published more than 250 peer-reviewed papers in computer vision, image processing, intelligent systems, robotics, and integrated microsystems. His current research interests include 3D computer vision, active video object tracking and extraction, interactive multimedia, sensor analysis and fusion, and visual sensor networks.

Dr. Wu holds the Tier 1 Canada Research Chair in Automotive Sensors and Information Systems. He is an Associate Editor for the IEEE Transactions on Systems, Man, and Cybernetics: Systems, IEEE Transaction on Neural Networks and Learning Systems, and the International Journal of Robotics and Automation. He has served on technical program committees and international advisory committees for many prestigious conferences.

Thanh Minh Nguyen received his B.E. degree with honors in Electrical and Electronic Engineering from the HCM City University of Technology, Vietnam in 2004, his M.S. degree from the Department of Electrical Engineering at the Da-Yeh University, Taiwan in 2006, and his Ph.D. from the Department of Electrical Engineering of the University of Windsor, Canada in 2011. He is currently a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering at the University of Windsor, Canada. His research interests are computer vision and pattern recognition.

Dr. Thanh has consistently excelled in all areas of his research work. The best example of his ingenuity is shown through the project on Autonomous Racing Challenge, which took top honors at the University of Waterloo in two years: 2008, 2009. He is the first author of some excellent journal papers in his research field, such as IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Medical Imaging, IEEE Transactions on Systems, Man, and Cybernetics, Part B, etc. He has also been serving as a reviewer for the most prominent journals of his research field, such as IEEE Transactions on Medical Imaging, IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Systems, Man, and Cybernetics, Part B, etc.