RECOGNITION AND CLUSTERING OF DIRICHLET MIXTURES

INTRODUCTION

Proportional data appear naturally in many applications from different domains such as data mining, computer vision, and pattern recognition. Examples include text (or image) clustering using the bag of words (or visual words) approach where each document (or image) can be represented as a normalized histogram of words (or visual words) (1). There is thus a need to develop appropriate statistical models to represent and to cluster these kind of data. The main idea is to discover underlying or hidden structure in these data to explain their generation process. In this case, it is common to consider parametric probability density functions (PDFs). Often the Gaussian PDF, with generally structurally restricted covariance matrix, or a mixture of it is used to describe the probability density of a random vector (2). Unfortunately, the Gaussian assumption is sometimes too restrictive and fails to give good results when the data partitions are clearly non-Gaussian, which is the case of proportional vectors (3). Indeed, it is well known that the key to efficient model-based learning is the choice of the appropriate distributions that reduces the complexity of describing the data drastically. Among numerous distributions, the Dirichlet turns out to be a principled alternative to the Gaussian and of exceptional efficiency for the modeling of proportional data. In this article we will review important issues related to Dirichlet mixture models that allow a formal approach to the clustering of proportional data.

The Dirichlet mixture has been shown to be a powerful tool for many real-world challenging applications such as texture categorization, human skin detection, image databases summarization, and text mining (3–6). A crucial problem when using Dirichlet mixture is the learning of its parameters. Several learning techniques have been proposed in the past to estimate and select (i.e., determine the complexity) finite mixture models (2). In this article we shall present two learning techniques. The first one is based on maximum likelihood estimation performed via the expectation–maximization (EM) algorithm, which is a well-known iterative method that allows finding maximum likelihood estimates in the case of incomplete or missing data (7). The estimation of the number of components, which is of particular importance (8), in this case is carried out using the minimum message length criterion as developed in Reference 9. The second one is Bayesian based on variational approximation and was proposed recently in Reference 1.

The rest of this article is structured as follows. In section "The Dirichlet Mixture Model," we present the Dirichlet mixture. Section "Finite Dirichlet Mixture Learning" discusses two approaches to learn this mixture. Some experimental results are given in section "Experimental Results." We conclude the article in Section "Conclusion."

THE DIRICHLET MIXTURE MODEL

In this section, we will briefly present the finite Dirichlet mixture that has been proposed as a general model for estimating an unknown probability density function. The virtues of the Dirichlet distribution lie mainly in its ease of use and flexibility by allowing different symmetric and asymmetric shapes (9). We say that a D-dimensional random vector $\vec{X} = (X_1, \ldots, X_D)$ has a finite mixture of Dirichlet distributions when its probability density function $p(\vec{X})$ can be written as a finite weighted sum Dirichlet distributions. For M Dirichlet distributions, the mixture model can be written as

$$p(\vec{X}|\Theta) = \sum_{j=1}^{M} \pi_j p(\vec{X}|\vec{\alpha}_j) \tag{1}$$

where $\{\pi_j\}$ represents the set of mixing weights that must be positive and sum to 1 (i.e., $\pi_j > 0$ and $\sum_{j=1}^{M} \pi_j = 1$, for $j = 1, \ldots, M$), $\Theta = \{\{\pi_j\}, \{\vec{\alpha}_j\}\}$ is the complete set of mixture's parameters, and $p(\vec{X}|\vec{\alpha}_j)$ stands for the D-dimensional Dirichlet representing component j:

$$p(\vec{X}|\vec{\alpha}_j) = \mathrm{Dir}(\vec{X}|\vec{\alpha}_j) = \frac{\Gamma(\sum_{l=1}^{D} \alpha_{jl})}{\prod_{l=1}^{D} \Gamma(\alpha_{jl})} \prod_{l=1}^{D} X_l^{\alpha_{jl}-1} \tag{2}$$

where $\sum_{l=1}^{D} X_l = 1$ and $0 \leq X_l \leq 1$ for $l = 1, \ldots, D$. The Dirichlet distribution is the multivariate extension of the two-parameter Beta distribution. Its mean and the variance are given by

$$E(X_l) = \frac{\alpha_l}{|\vec{\alpha}|} \tag{3}$$

and

$$\mathrm{Var}(X_l) = \frac{\alpha_l(|\vec{\alpha}| - \alpha_l)}{|\vec{\alpha}|^2(|\vec{\alpha}| + 1)} \tag{4}$$

where

$$|\vec{\alpha}| = \sum_{l=1}^{D} \alpha_l \tag{5}$$

and the covariance between X_l and X_s is

$$Cov(X_l, X_s) = -\frac{\alpha_l \alpha_s}{|\vec{\alpha}|^2(|\vec{\alpha}| + 1)} \tag{6}$$

It is noteworthy that the Dirichlet distribution is often used as the conjugate prior of the multinomial distribution in Bayesian statistics (10). In the following section, we shall present two different approaches to learn finite Dirichlet mixture models.

FINITE DIRICHLET MIXTURE LEARNING

Mixture models learning has been and is still an active research area, probably because of its wide range of applications. Hence, many different learning algorithms and techniques have been proposed over the last years. Generally, the work that has been carried out on learning mixture models can be grouped roughly into two families of approaches: deterministic or Bayesian. Deterministic approaches aim at optimizing the model likelihood function,

and are generally implemented within the expectation–maximization (7) framework. On the other hand, Bayesian techniques have been proposed to avoid drawbacks related to deterministic techniques such as their suboptimal generalization performance, dependency on initialization, overfitting and noise level underestimation problems of classical likelihood-based inference (11,12). In the following two sections, we are going to introduce two approaches for learning finite Dirichlet mixture models, including the EM algorithm with minimum message length (MML) model selection that we have previously proposed in Reference 9 and the variational inference that we have developed in Reference 1. By learning, we mean both the estimation of the parameters and the selection of the number of components M.

Maximum Likelihood Estimation with MML Model Selection

Most often, the estimation of the parameters is carried out by the maximum likelihood method of which the main goal is maximizing the likelihood of a set of vectors. One of the algorithms generally used for mixture models estimation using maximum likelihood is the EM algorithm (7). EM is a popular method for parameter estimation by providing iterative formulas for the learning of these unknown parameters. Moreover, it has been proven that the EM increases monotonically the likelihood function (13).

Assume that we have a set of independent vectors $\mathcal{X}$. Then, the log-likelihood for the vector Θ of unknown parameters is given by

$$\log p(\mathcal{X}|\Theta) = \log(\prod_{i=1}^{N} p(\vec{X}_i|\Theta)) = \sum_{i=1}^{N} \log p(\vec{X}_i|\Theta) \tag{7}$$

where $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$. In the case of finite Dirichlet mixture model, the log-likelihood corresponding to an M-component, using Lagrange multiplier Λ, is

$$\begin{aligned}\mathcal{L}(\Theta, \mathcal{X}) &= \log \prod_{i=1}^{N} p(\vec{X}_i|\Theta) + \Lambda(1 - \sum_{j=1}^{M} \pi_j) \\ &= \sum_{i=1}^{N} \log \sum_{j=1}^{M} \pi_j p(\vec{X}_i|\Theta) + \Lambda\left(1 - \sum_{j=1}^{M} \pi_j\right)\end{aligned} \tag{8}$$

Then, the maximum likelihood estimate is given by

$$\hat{\Theta}_{\mathrm{ML}} = \arg\max_{\Theta}\{L(\Theta, \mathcal{X})\} \tag{9}$$

Since this maximum likelihood evaluation cannot be found analytically, EM algorithm is required for estimating the involved unknown parameters. In the EM framework, the expectation (E) step is devoted to estimating the probability of each observation to belong to each cluster conditionally on the current parameter estimates. In the maximization (M) step, the model parameters are estimated given the current group membership probabilities.

Apart from the estimation task, a crucial problem when dealing with mixture models is the selection of the appropriate number of components. A wide range of model selection criteria have been proposed in the past and have been deeply discussed in Reference 9, but so far no single method has really emerged as the method of choice. Indeed, it is very difficult to define the meaning of a good clustering structure or an optimal number of clusters. In Reference 9, the authors have shown that several of the previously proposed selection criteria are special cases of the MML criterion (14). Given a data set $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$ governed by the unknown parameters Θ, the message length is defined as minus the logarithm of the posterior probability as

$$\begin{aligned}\mathrm{MessLen}(\Theta, \mathcal{X}) &= -\log(p(\Theta|\mathcal{X})) \\ &= \underbrace{-\log(h(\Theta))}_{\mathrm{MessLen}(\Theta)} \underbrace{-\log(p(\mathcal{X}|\Theta))}_{\mathrm{MessLen}(\mathcal{X}|\Theta)}\end{aligned} \tag{10}$$

where $h(\Theta)$ is the prior probability and $p(\mathcal{X}|\Theta)$ is the likelihood.

An MML-based framework has been proposed in Reference 9 for learning finite Dirichlet mixture model. Within this framework, model parameters are determined using the EM algorithm where model selection (i.e., estimating the number of mixture components) is performed through the incorporation of the MML criterion. More specifically, the MML criterion for the finite Dirichlet mixture model with M components is calculated by

$$\begin{aligned}\mathrm{MessLen} &= \mathrm{MML}(M) \\ &= \sum_{j=1}^{M-1} -\log(j) + 6MD + D\sum_{j=1}^{M} \log(|\vec{\hat{\alpha}}_j|) \\ &\quad - \sum_{j=1}^{M}\sum_{l=1}^{D} \log(\hat{\alpha}_{jl}) \\ &\quad + \frac{\log(N)}{2} - \frac{1}{2}\sum_{j=1}^{M} \log(\pi_j) \\ &\quad + \frac{1}{2}\sum_{j=1}^{M} \log\Big|1 - \Psi'(|\vec{\hat{\alpha}}_j|)\sum_{l=1}^{D} \frac{1}{\Psi'(\hat{\alpha}_{jl})}\Big| \\ &\quad + \frac{1}{2}D\sum_{j=1}^{M} \log(n_j) + \frac{1}{2}\sum_{j=1}^{M}\sum_{l=1}^{D} \log(\Psi'(\hat{\alpha}_{jl})) \\ &\quad - \log(p(\mathcal{X}|\Theta)) - \frac{N_p}{2}\log(12) + \frac{N_p}{2}\end{aligned} \tag{11}$$

where $\Psi'(\cdot)$ is the trigamma function, $\vec{\hat{\alpha}}_j$ is the estimated vector by the EM algorithm, N_p is the dimension of Θ (i.e. the number of parameters to be estimated), and n_j is the number of elements in cluster j. Then, the optimal model M^* is selected as

$$M^* = \arg\min_{M} \mathrm{MML}(M) \tag{12}$$

Although this approach is effective for learning Dirichlet mixture models, it is clearly time consuming since it has to evaluate the criterion for several numbers of mixture components. In the next section, a more efficient and elegant algorithm, namely, variational inference for learning

Dirichlet mixtures as we proposed in Reference 1 shall be described.

Variational Learning

A disadvantage of maximum likelihood estimation is that it exhibits generally local maxima or even get stuck on saddle points in the worst case. It is crucial then to be able to escape from such local maxima or saddle points. Several approaches have been proposed in the past to achieve this goal. Successful approaches have been generally based on Bayesian learning (12,15). In Bayesian estimation learning, it is assumed that there is some prior information about the parameters. It requires then the specification of prior distributions. These priors are used with the data likelihood to obtain the posteriors from which the model's parameters are simulated. This simulation is often performed using Markov Chain Monte Carlo (MCMC) techniques such as Gibbs sampling and Metropolis–Hastings, which is time consuming (12,15). Variational learning (also known as *variational Bayes*) (16–21) is an approximation to exact Bayesian learning that is based on approximating the posterior distribution with another simpler distribution. This approximation makes the learning task tractable while preserving several of the benefits of Bayesian learning such as avoiding under- and overfitting. Thus, the variational Bayesian framework has been proposed and has been widely used as an efficient alternative to purely Bayesian learning. The main goal of this section is to show how a variational inference approach that learns the finite Dirichlet mixture model, by simultaneously optimizing both its parameters and its structure (i.e., complexity or the number of mixture components), could be developed.

For the learning problem, we assume that there is a training set $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$ of N independent and identically distributed vectors which is generated from the mixture of Dirichlet distributions, as illustrated in equation 1. Then, the likelihood function of this Dirichlet mixture model is

$$p(\mathcal{X}|\vec{\pi}, \vec{\alpha}) = \prod_{i=1}^{N} \left\{ \sum_{j=1}^{M} \pi_j \mathrm{Dir}(\vec{X}_i|\vec{\alpha}_j) \right\} \tag{13}$$

where $\vec{\pi} = (\pi_1, \ldots, \pi_M)$ and $\vec{\alpha} = (\vec{\alpha}_1, \ldots, \vec{\alpha}_M)$.

Latent Variable $\mathcal{Z}$ and Prior Distribution $p(\vec{\alpha})$. Generally, the finite mixture model in equation 13 is interpreted as a latent variable model. Indeed, it is possible to assign to each vector $\vec{X}_i$, an M-dimensional binary random vector $\vec{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ where $Z_{ij} \in \{0, 1\}$ and $\sum_{j=1}^{M} Z_{ij} = 1$. We then have $Z_{ij} = 1$ if $\vec{X}_i$ belongs to component j; $Z_{ij} = 0$, otherwise. Thus, Z_{ij} can be viewed as the class indicator. Given $\mathcal{Z} = \{\vec{Z}_1, \ldots, \vec{Z}_N\}$, the conditional distribution of $\mathcal{Z}$ given the mixing coefficients $\vec{\pi}$ is defined as

$$p(\mathcal{Z}|\vec{\pi}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \pi_j^{Z_{ij}} \tag{14}$$

Then, the complete likelihood function of the Dirichlet mixture model can be written as

$$p(\mathcal{X}|\mathcal{Z}, \vec{\alpha}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \mathrm{Dir}(\vec{X}_i|\vec{\alpha}_j)^{Z_{ij}} \tag{15}$$

As already mentioned, an important step in Bayesian learning approaches is to introduce priors over random variables. In our case, a conjugate prior $p(\vec{\alpha})$ is placed over the model parameter $\vec{\alpha}$. It is well known that the choice of conjugate priors facilitates the learning procedure (22). A prior is called conjugate when it has the same form as the resulting posterior. It is possible to derive a conjugate prior for any distributions that belongs to the exponential family (23), which is the case of the Dirichlet (24) as follows:

$$p(\vec{\alpha}_j) = f(\nu, \lambda) \left[\frac{\Gamma(\sum_{l=1}^{D} \alpha_{jl})}{\prod_{l=1}^{D} \Gamma(\alpha_{jl})} \right]^{\nu} \prod_{l=1}^{D} \mathrm{e}^{-\lambda_l(\alpha_{jl}-1)} \tag{16}$$

where $f(\nu, \lambda)$ is a normalization coefficient and (ν, λ) are hyperparameters. Unfortunately, this formal conjugate prior for the Dirichlet distribution is intractable, due to the difficulty to evaluate the normalization coefficient.

In this work, we adopt the Gamma distribution $\mathcal{G}(\cdot)$ to approximate the conjugate prior for $\vec{\alpha}$ by assuming that the Dirichlet parameters are statistically independent as

$$p(\vec{\alpha}) = \mathcal{G}(\vec{\alpha}|\vec{u}, \vec{v}) = \prod_{j=1}^{M} \prod_{l=1}^{D} \frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})} \alpha_{jl}^{u_{jl}-1} \mathrm{e}^{-v_{jl}\alpha_{jl}} \tag{17}$$

where $\vec{u} = \{u_{jl}\}$ and $\vec{v} = \{v_{jl}\}$ are hyperparameters, subject to the constraints $u_{jl} > 0$ and $v_{jl} > 0$.

Variational Solutions to Finite Dirichlet Mixture Models. To simplify notation, we define $\Theta = \{Z, \vec{\alpha}\}$ as the set of latent and random variables. The goal of variational learning approach is to determine a probability distribution $q(\Theta)$, which approximates the true posterior distribution $p(\Theta|\mathcal{X}, \vec{\pi})$. Here the mixing coefficients $\vec{\pi}$ are treated as parameters instead of random variables as in the fully Bayesian approach. The Kullback–Leibler (KL) divergence between the approximated distribution $q(\Theta)$ and the true posterior distribution $p(\Theta|\mathcal{X}, \vec{\pi})$ is given by

$$\begin{aligned} \mathrm{KL}(q||p) &= -\int_{\Theta} q(\Theta) \ln \left(\frac{p(\Theta|\mathcal{X}, \vec{\pi})}{q(\Theta)} \right) \mathrm{d}\Theta \\ &= \ln p(\mathcal{X}|\vec{\pi}) - \mathcal{L}(q) \end{aligned} \tag{18}$$

where $\mathcal{L}(Q)$ is defined as

$$\mathcal{L}(q) = \int_{\Theta} q(\Theta) \ln \left(\frac{p(\Theta, \mathcal{X}|\vec{\pi})}{q(\Theta)} \right) \mathrm{d}\Theta \tag{19}$$

Since the KL divergence satisfies $\mathrm{KL}(q||p) \geq 0$ and equals to 0 when $q(\Theta) = p(\Theta|\mathcal{X}, \vec{\pi})$, this leads to $\mathcal{L}(q) \leq \ln p(\mathcal{X}|\vec{\pi})$, which means that $\mathcal{L}(q)$ represents the lower bound on $\ln p(\mathcal{X}|\vec{\pi})$. Therefore, we can obtain the true posterior distribution by maximizing the lower bound $\mathcal{L}(q)$, which is fulfilled when the KL divergence equals to zero (i.e., $q(\Theta) = p(\Theta|\mathcal{X}, \vec{\pi})$).

Factorization assumption for restricting the form of $q(\Theta)$ (25–28) is generally used in the case of variational learning:

$$q(\Theta) = q(\mathcal{Z})q(\vec{\alpha}) = \Big[\prod_{i=1}^{N}\prod_{j=1}^{M} q(Z_{ij})\Big]\Big[\prod_{j=1}^{M}\prod_{l=1}^{D} q(\alpha_{jl})\Big] \tag{20}$$

It is noteworthy that this is the only assumption about the distribution, and no further restriction is placed on the functional forms of the individual factors $q_s(\Theta_s)$. In order to maximize the lower bound $\mathcal{L}(q)$, we need to make a variational optimization of $\mathcal{L}(q)$ with respect to each of the factors $q_s(\Theta_s)$ in turn (29):

$$q_s(\Theta_s) = \frac{\exp\big\langle \ln p(\mathcal{X}, \Theta)\big\rangle_{\neq s}}{\int \exp\big\langle \ln p(\mathcal{X}, \Theta)\big\rangle_{\neq s} \mathrm{d}\Theta} \tag{21}$$

where $\langle\cdot\rangle_{\neq s}$ denotes the expectation with respect to all the factor distributions except for s. Of course, all factors $q_s(\Theta_s)$ need to be suitably initialized first, then each factor is updated in turn with a revised value obtained by equation 21 using the current values of the other factors (30).

By applying equation 21 to each factor of the variational posterior $q(\Theta)$, we then obtain the following variational solutions:

$$q(\mathcal{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{Z_{ij}} \tag{22}$$

$$q(\vec{\alpha}) = \prod_{j=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\alpha_{jl}|u_{jl}^{*}, v_{jl}^{*}) \tag{23}$$

where we have defined

$$r_{ij} = \frac{\rho_{ij}}{\sum_{j=1}^{M}\rho_{ij}}, \qquad u_{jl}^{*} = u_{jl} + \varphi_{jl}, \qquad v_{jl}^{*} = v_{jl} - \vartheta_{jl} \tag{24}$$

$$\rho_{ij} = \exp\left\{\ln \pi_j + \widetilde{\mathcal{R}}_j + \sum_{l=1}^{D}(\bar{\alpha}_{jl} - 1)\ln X_{il}\right\} \tag{25}$$

$$\begin{aligned}\widetilde{\mathcal{R}}_j = {} & \ln\frac{\Gamma(\sum_{l=1}^{D}\bar{\alpha}_{jl})}{\prod_{l=1}^{D}\Gamma(\bar{\alpha}_{jl})} + \sum_{l=1}^{D}\bar{\alpha}_{jl}\left[\Psi\left(\sum_{l=1}^{D}\bar{\alpha}_{jl}\right) - \Psi(\bar{\alpha}_{jl})\right] \\ & \big[\big\langle \ln\alpha_{jl}\big\rangle - \ln\bar{\alpha}_{jl}\big] \\ & + \frac{1}{2}\sum_{l=1}^{D}\bar{\alpha}_{jl}^{2}\left[\Psi'\left(\sum_{l=1}^{D}\bar{\alpha}_{jl}\right) - \Psi'(\bar{\alpha}_{jl})\right]\big\langle(\ln\alpha_{jl} - \ln\bar{\alpha}_{jl})^2\big\rangle \\ & + \frac{1}{2}\sum_{a=1}^{D}\sum_{\substack{b=1 \\ (b\neq a)}}^{D}\bar{\alpha}_{ja}\bar{\alpha}_{jb} \\ & \left\{\Psi'\left(\sum_{l=1}^{D}\bar{\alpha}_{jl}\right)\big(\big\langle\ln\alpha_{ja}\big\rangle - \ln\bar{\alpha}_{ja}\big)\big(\big\langle\ln\alpha_{jb}\big\rangle - \ln\bar{\alpha}_{jb}\big)\right\}\end{aligned} \tag{26}$$

$$\varphi_{jl} = \sum_{i=1}^{N}\big\langle Z_{ij}\big\rangle\bar{\alpha}_{jl}\left[\Psi\left(\sum_{k=1}^{D}\bar{\alpha}_{jk}\right) - \Psi(\bar{\alpha}_{jl})\right.$$

$$\left. + \sum_{k\neq l}^{D}\Psi'\left(\sum_{k=1}^{D}\bar{\alpha}_{k}\right)\bar{\alpha}_{k}\big(\big\langle\ln\alpha_{k}\big\rangle - \ln\bar{\alpha}_{k}\big)\right] \tag{27}$$

$$\vartheta_{jl} = \sum_{i=1}^{N}\big\langle Z_{ij}\big\rangle\ln X_{il} \tag{28}$$

where $\Psi(\cdot)$ is the digamma functions. $\widetilde{\mathcal{R}}_j$ is the approximated lower bound of $\mathcal{R}_j$, where $\mathcal{R}_j = \big\langle \ln\frac{\Gamma(\sum_{l=1}^{D}\alpha_{jl})}{\prod_{l=1}^{D}\Gamma(\alpha_{jl})}\big\rangle$. Since a closed-form expression cannot be found for $\mathcal{R}_j$, we adopt the second-order Taylor series expansion to find its lower bound approximation. The expected values in the above formulas are defined by

$$\big\langle Z_{ij}\big\rangle = r_{ij}, \qquad \bar{\alpha}_{jl} = \big\langle\alpha_{jl}\big\rangle = \frac{u_{jl}}{v_{jl}} \tag{29}$$

$$\big\langle\ln\alpha_{jl}\big\rangle = \Psi(u_{jl}) - \ln v_{jl} \tag{30}$$

$$\big\langle(\ln\alpha_{jl} - \ln\bar{\alpha}_{jl})^2\big\rangle = [\Psi(u_{jl}) - \ln u_{jl}]^2 + \Psi'(u_{jl}) \tag{31}$$

In this framework, the mixing coefficients $\vec{\pi}$ are treated as parameters, and point estimations of their values are evaluated by maximizing the variational likelihood bound $\mathcal{L}(q)$. Setting the derivative of this lower bound with respect to $\vec{\pi}$ to zero gives

$$\pi_j = \frac{1}{N}\sum_{i=1}^{N} r_{ij} \tag{32}$$

For the determination of the number of components M, we start the learning algorithm with a large initial value of M and then remove the redundant components after convergence. The variational inference for finite Dirichlet mixture model is summarized in Algorithm 1.

Algorithm 1 Variational Dirichlet Mixtures

1: Choose the initial number of components M and the initial values for hyperparameters $\{u_{jl}\}$ and $\{v_{jl}\}$.
2: Initialize the value of r_{ij} by K-means algorithm.
3: **repeat**.
4: The variational E-step: update the variational solutions for $q(\mathcal{Z})$ and $q(\vec{\alpha})$ using equations 22 and 23, respectively.
5: The variational M-step: maximize lower bound $\mathcal{L}(q)$ with respect to the current value of $\vec{\pi}$ in equation 32.
6: **until** Convergence criterion is reached (i.e., stabilization of the variational lower bound in equation 19 with the estimated values of hyperparameters).
7: Detect the optimal number of components M by eliminating the components with small mixing coefficients (less than 10^{-5}).

EXPERIMENTAL RESULTS

In this section, we evaluate the effectiveness of the variational inference algorithm (1) and compare it with the MML algorithm (9) for learning finite Dirichlet mixture models through both synthetic data sets and a real application, namely, unsupervised image categorization. We refer to the variational algorithm as *varDM*, and to the MML algorithm as *mmlDM*. In our experiments, we adopt the

Table 1. Parameters of Different Generated Data Sets

	n_j	j	α_{j1}	α_{j2}	α_{j3}	π_j	$\hat{\alpha}_{j1}$	$\hat{\alpha}_{j2}$	$\hat{\alpha}_{j3}$	$\hat{\pi}_j$	$\breve{\alpha}_{j1}$	$\breve{\alpha}_{j2}$	$\breve{\alpha}_{j3}$	$\breve{\pi}_j$
Data set 1	200	1	12	30	45	0.5	12.59	31.29	45.56	0.50	11.31	28.62	45.75	0.493
(N = 400)	200	2	32	50	16	0.5	33.58	50.20	15.64	0.50	33.28	49.14	16.58	0.507
Data set 2	200	1	12	30	45	0.4	13.91	35.40	51.07	0.398	11.44	33.67	49.54	0.449
(N = 500)	200	2	32	50	16	0.4	32.68	51.71	16.81	0.401	30.19	48.87	16.83	0.354
	100	3	55	28	35	0.2	50.43	25.99	31.95	0.201	52.34	30.12	37.47	0.197
Data set 3	200	1	12	30	45	0.25	13.07	31.96	46.63	0.247	13.69	28.21	46.75	0.233
(N = 800)	200	2	25	18	90	0.25	24.02	17.76	85.44	0.253	25.91	17.53	94.18	0.247
	200	3	55	28	35	0.25	54.89	27.73	34.13	0.249	54.72	29.08	33.97	0.265
	200	4	32	50	16	0.25	31.63	48.73	14.45	0.251	33.51	51.25	17.14	0.255
Data set 4	200	1	12	30	45	0.2	11.46	27.97	41.98	0.198	11.37	32.26	46.87	0.192
(N = 1000)	100	2	25	18	90	0.1	25.16	19.23	93.36	0.098	24.29	19.38	86.54	0.123
	300	3	55	28	35	0.3	54.45	28.58	34.40	0.300	55.55	28.79	36.75	0.281
	200	4	32	50	16	0.2	36.23	55.47	18.04	0.198	36.13	54.82	18.22	0.186
	200	5	3	118	60	0.2	3.22	130.15	65.89	0.206	2.87	110.34	57.31	0.218

N denotes the total number of elements, n_j denotes the number of elements in cluster j. α_{j1}, α_{j2}, α_{j3} and π_j are the real parameters. $\hat{\alpha}_{j1}$, $\hat{\alpha}_{j2}$, $\hat{\alpha}_{j3}$, and $\hat{\pi}_j$ are the estimated parameters by *varDM*. $\breve{\alpha}_{j1}$, $\breve{\alpha}_{j2}$, $\breve{\alpha}_{j3}$, and $\breve{\pi}_j$ are the estimated parameters using *mmlDM*.

same initialization setting for *mmlDM* as in Reference 9. For *varDM*, similar to that in Reference 1, we initialize the number of components to 15 with equal mixing coefficients. We also initialize the hyperparameters $\{u_{jl}\}$ and $\{v_{jl}\}$ to 1 and 0.01, respectively.

Synthetic Data

The performance of *varDM* and *mmlDM*, in terms of estimation and selection, on four three-dimensional synthetic data sets is generated using different mixture's parameters. Table 1 shows the real and estimated parameters of each data set using both *varDM* and *mmlDM*. According to this table, we can observe that both algorithms were able to estimate unknown parameters accurately for all synthetic data sets, yet the *varDM* algorithm provided slightly better results, especially for estimating mixing coefficients. Figure 1 represents the resulting mixtures with different shapes. For *varDM*, in order to estimate the number of components, we apply directly the algorithm on these data sets (by starting with 15 components). The redundant components have estimated mixing coefficients close to 0 after convergence. By removing these redundant components, we obtain the correct number of components for each generated data set. Figure 2 illustrates the value of the variational likelihood bound in each iteration and shows that the likelihood bound increases at each iteration. As shown in Figure 3, we ran *varDM* by varying the number of mixture

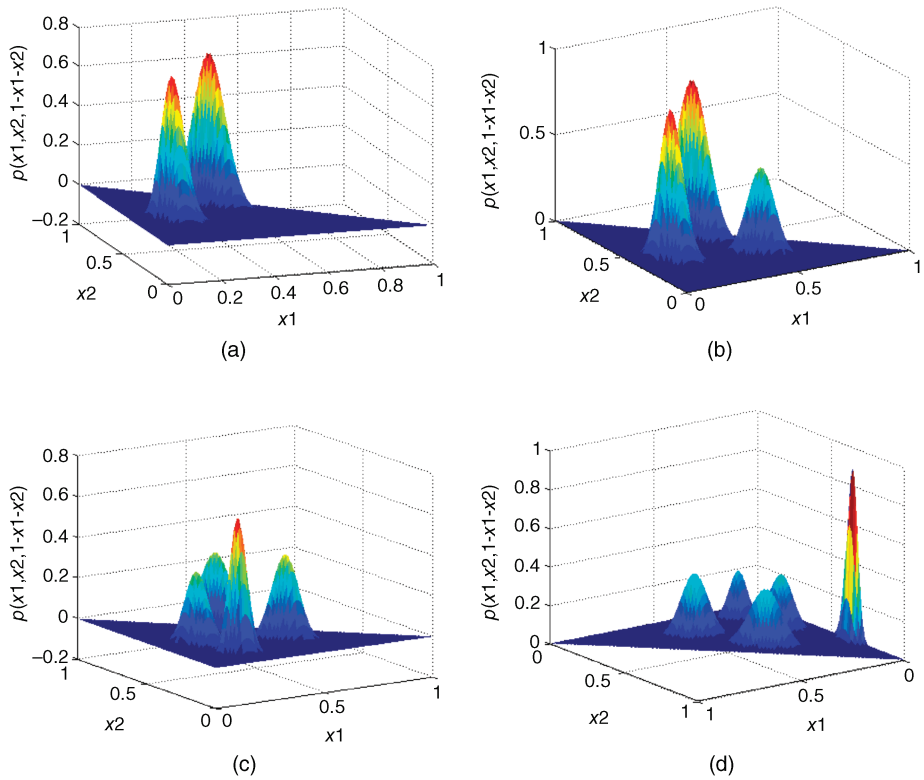

Figure 1. Estimated mixture densities for the artificial data sets using *varDM*. (a) Data set 1. (b) Data set 2. (c) Data set 3. (d) Data set 4.

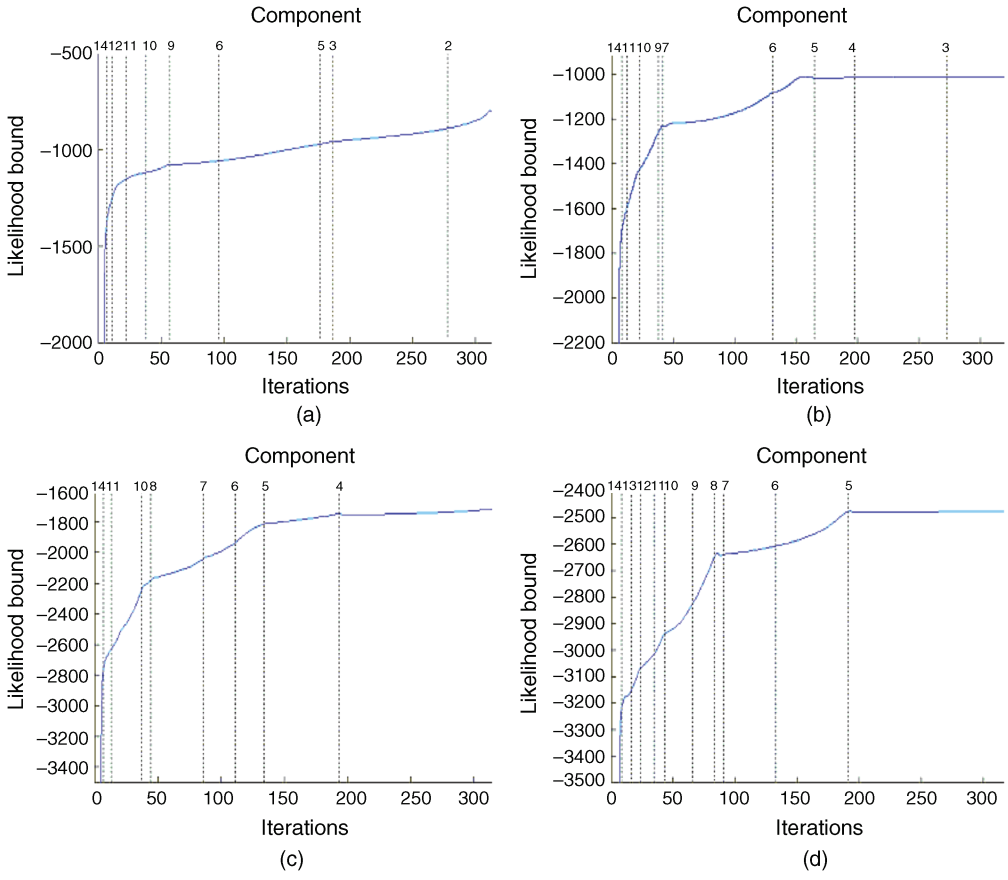

Figure 2. Variational likelihood bound for each iteration for the different generated data sets. The initial number of components is 15. Vertical dashed lines indicate cancellation of components. (a) Data set 1. (b) Data set 2. (c) Data set 3. (d) Data set 4.

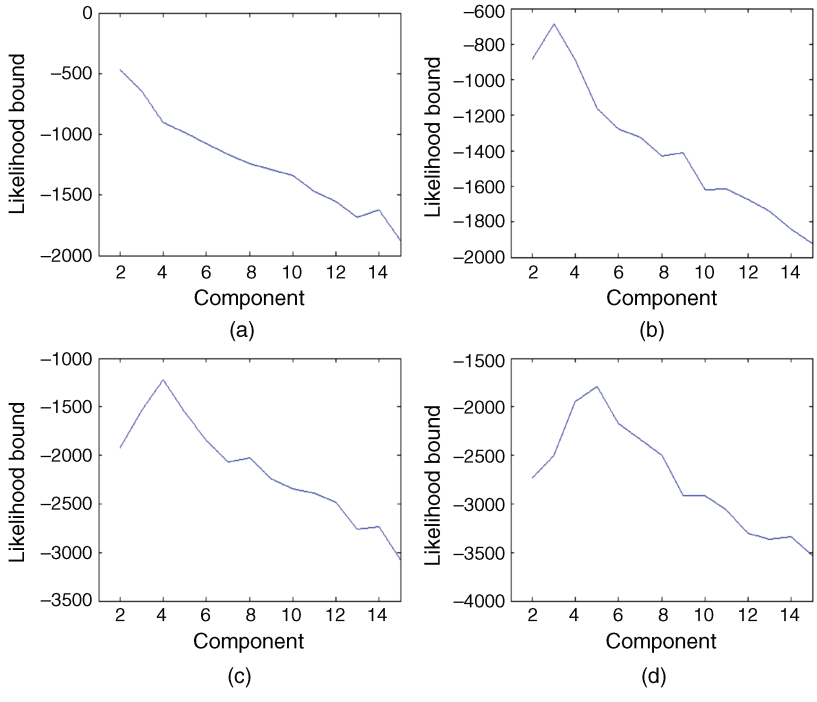

Figure 3. Variational likelihood bound as a function of the fixed assumed number of mixture components for the different generated data sets. (a) Data set 1; (b) Data set 2; (c) Data set 3; (d) Data set 4.

Table 2. Run Time (in Seconds) and Number of Iterations Required before Convergence for *varDM* and *mmlDM*

	varDM		***mmlDM***	
Data set	Run time	No. of iterations	Run time	No. of iterations
1	4.81	278	10.94	357
2	4.73	269	11.12	368
3	4.08	191	12.21	393
4	3.95	189	12.65	414

components from 2 to 15. According to this figure, it is clear that for each data set, the correct number of components is selected. Furthermore, we performed a comparison between the numerical complexity of *varDM* and *mmlDM*, in terms of overall computational time and the number of iterations before convergence. The corresponding results are shown in Table 2. As we can see from this table, it is obvious that for each data set, the *varDM* algorithm is faster and requires less iterations to converge.

Visual Scenes Categorization

Visual scenes categorization is an important task for many pattern recognition and vision-based applications. An important step to perform this task is to extract adequate descriptors (i.e., discriminative and invariant at the same time) to represent the visual scenes. Local features have been shown to provide good categorization capabilities. In this section, we validate the Dirichlet mixture on the image categorization task using the bag-of-visual words approach (31,32). This approach is based on the following steps. First, scale-invariant feature transform (SIFT) (33) descriptors are extracted from each image using the difference-of-Gaussians (DoG) interest point detectors. Then, a visual vocabulary $\mathcal{V}$ is built by quantizing these SIFT features into visual words V (we choose V=750) using K-means algorithm. As a result, each image is represented as a histogram of visual words. Finally, the probabilistic latent semantic analysis (pLSA) model (34) is applied to these histograms to describe each image as a D-dimensional vector of proportions where D is the number of learnt topics (or aspects) (32). In our experiments, we set $\mathcal{D}$ as 40. Finally, *varDM* and *mmlDM* are deployed as classifiers to categorize a testing image into the scene category that has the highest posterior probability based on Bayes' decision rule. We conducted our experiments using a challenging and publicly available data set, namely, OT data set (35,36). The OT data set includes 2688 images with the size of 256 $\times$ 256 pixels, and is composed of eight urban and natural scene categories: coast (360 images), forest (328 images), highway (260 images), inside city (308 images), mountain (374 images), open country (410 images), street (292 images), and tall building (356 images).

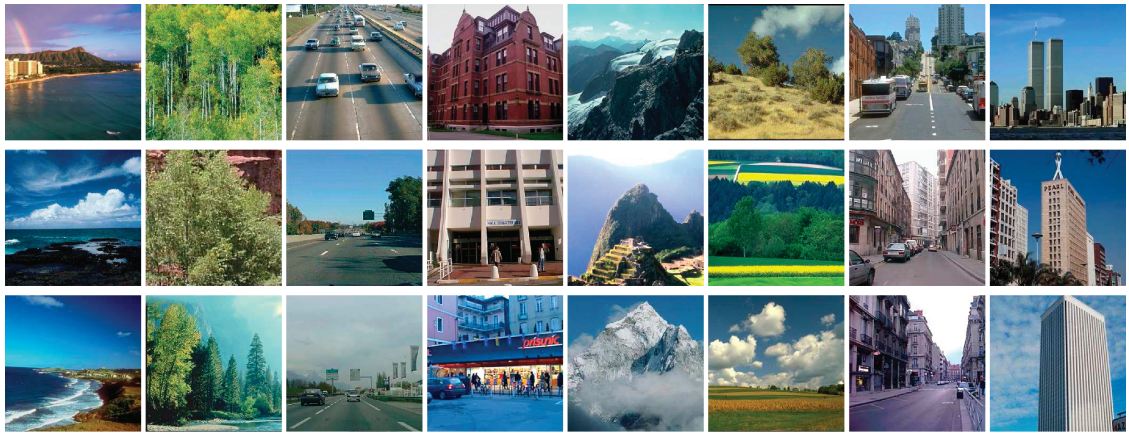

Figure 4. Sample images from the OT data set. (a) Coast. (b) Forest. (c) Highway. (d) Inside city. (e) Mountain. (f) Open country. (g) Street. (h) Tall building.

Table 3. Average Rounded Confusion Matrix for the OT Database Calculated by *mmlDM* Using Different Partitions for Constructing the Visual Vocabulary and for Testing

	C	F	H	I	M	O	S	T	Acc. (%)
Coast (C)	**113**	12	9	2	5	33	5	1	62.8
Forest (F)	0	**148**	0	3	4	5	2	2	90.2
Highway (H)	2	0	**116**	3	1	7	1	0	89.2
Inside city (I)	0	4	1	**114**	5	1	22	7	74.0
Mountain (M)	3	27	9	4	**129**	13	2	0	68.9
Open country (O)	25	31	16	5	6	**115**	4	3	56.1
Street (S)	1	0	5	8	11	6	**112**	3	76.7
Tall building (T)	6	2	8	14	8	3	28	**109**	61.2
							Overall rate		72.4

Table 4. Average Rounded Confusion Matrix for the OT Database Calculated by *varDM* Using Different Partitions for Constructing the Visual Vocabulary and for Testing

	C	F	H	I	M	O	S	T	Acc. (%)
Coast (C)	**115**	17	5	3	3	31	2	4	63.9
Forest (F)	0	**154**	3	2	2	1	0	2	93.9
Highway (H)	3	0	**123**	0	0	1	2	1	94.6
Inside city (I)	3	2	1	**111**	5	6	14	12	72.1
Mountain (M)	6	27	7	4	**134**	7	1	2	71.7
Open country (O)	29	23	13	7	8	**120**	3	2	58.5
Street (S)	0	0	7	9	9	6	**115**	0	78.8
Tall building (T)	8	3	7	12	9	5	22	**112**	62.9
							Overall rate		74.5

Table 5. The Average Classification Accuracy Rate (Acc) (%) and the Number of Categories ($\widehat{M}$) Obtained over 30 Runs Using Different Methods. The Numbers in Parenthesis are the Standard Deviation of the Corresponding Quantities

Method	$\widehat{M}$	Acc (%)
mmlDM	7.19 (0.63)	72.4 (1.41)
varDM	7.38 (0.56)	74.5 (1.36)

Sample images of different categories from the OT data set are shown in Figure 4. We randomly divided the OT data set into two partitions. The first partition is used to construct the visual vocabulary and the other for testing. Tables 3 and 4 present the confusion matrices calculated by *mmlDM* and *varDM*, respectively. The overall categorization accuracy is 72.4% (error rate of 27.6%) using the *mmlDM* algorithm and 74.5% (error rate of 25.5%) using *varDM* for the OT data set. Table 5 illustrates the average number of classes and the average categorization accuracies using both *mmlDM* and *varDM* models.

CONCLUSION

This article has presented the Dirichlet mixture as a good choice for proportional data modeling and clustering. Two approaches have been proposed for the learning of this mixture's parameters. The first approach is based on maximum likelihood estimation, and the second one is variational. Experimental results are presented for generated and real data sets. Although the Dirichlet distribution provides acceptable results, it has a major drawback that is its restrictive negative covariance structure. Recently, other mixture models for proportional data modeling based on the generalized Dirichlet and the Beta-Liouville distributions have been proposed to overcome this limitation (37,38).

BIBLIOGRAPHY

1. W. Fan, N. Bouguila, and D. Ziou. Variational Learning for Finite Dirichlet Mixture Models and Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*(5), pp 762–774.
2. G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, Inc.: New York, 2000.
3. N. Bouguila and D. Ziou. Using Unsupervised Learning of a Finite Dirichlet Mixture Model to Improve Pattern Recognition Applications. *Pattern Recognit. Lett.* **2005**, *26*(12), pp 1916–1925.
4. N. Bouguila and D. Ziou. MML-Based Approach for Finite Dirichlet Mixture Estimation and Selection. In *Lecture Notes in Computer Science*; Perner, P. and Imiya, A., Eds.; *MLDM*, Vol. 3587, Springer, 2005; pp 42–51.
5. N. Bouguila and D. Ziou. A Probabilistic Approach for Shadows Modeling and Detection. In *Proceedings of the International Conference on Image Processing (ICIP)*; IEEE, 2005; pp 329–332.
6. N. Bouguila and D. Ziou. On Fitting Finite Dirichlet Mixture Using ECM and MML. In *Lecture Notes in Computer Science*, Singh, S., Singh, M., Apté, C., and Perner, P., Eds.; *ICAPR (1)*, Vol. 3686, Springer, 2005; pp 172–182.
7. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. B* **1977**, *39*(1), pp 1–38.
8. P. Schlattmann. Estimating the Number of Components in a Finite Mixture Model: The Special Case of Homogeneity. *Comput. Stat. Data Anal.* **2003**, *41*(3-4), pp 441–451.
9. N. Bouguila and D. Ziou. Unsupervised Selection of a Finite Dirichlet Mixture Model: An MML-Based Approach. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*(8), pp 993–1009.
10. N. Bouguila. Count Data Modeling and Classification Using Finite Mixtures of Distributions. *IEEE Trans. Neural Netw.* **2011**, *22*(2), pp 186–198.
11. S. Waterhouse, D. MacKay, and T. Robinson. Bayesian Methods for Mixtures of Experts. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*; MIT Press, 1995; pp 351–357.
12. C.P. Robert. *The Bayesian Choice*. Springer, 2001.
13. J. G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, 1997.
14. C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, 2005.
15. P. Congdon. *Applied Bayesian Modelling*. John Wiley & Sons, Inc.: New York, 2003.
16. H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proceedings of the Conference on Uncertainity in Artificial Intelligence (UAI)*; Morgan Kaufman, 1999; pp 21–30.
17. H. Attias. A Variational Bayes Framework for Graphical Models. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*; MIT Press, 1999; pp 209–215.
18. M. J. Beal and Z. Ghahramani. The Variational Bayesian EM Algorithm for Incomplete Data: With Application to Scoring

Graphical Model Structures. In *Bayesian Statistics*; Bernardo, J. M., et al., Ed., Oxford University Press, 2003; pp 453–464.

19. A. E. Teschendorff, Y. Wang, N. L. Barbosa-Morais, J. D. Brenton, and C. Caldas. A Variational Bayesian Mixture Modelling Framework for Cluster Analysis of Gene-Expression Data. *Bioinformatics* **2005**, *21*(13), pp 3025–3033.
20. B. Wang and D. M. Titterington. Convergence Properties of a General Algorithm for Calculating Variational Bayesian Estimates for a Normal Mixture Model. *Bayesian Anal.* **2006**, *1*(3), pp 625–650.
21. B. Wang and D. M. Titterington. Convergence and Asymptotic Normality of Variational Bayesian Approximations for Exponential Family Models with Missing Values. In *Proceedings of the Conference on Uncertainity in Artificial Intelligence (UAI)*; Morgan Kaufman, 2004; pp 577–584.
22. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
23. P. Diaconis and D. Ylvisaker. Conjugate Priors for Exponential Families. *Ann. Stat.* **1979**, *7*, pp 269–281.
24. E. Castillo, A. S. Hadi, and C. Solares. Learning and Updating of Uncertainty in Dirichlet Models. *Mach. Learn.* **1997**, *26*(1), pp 43–63.
25. C. M. Bishop, N. Lawrence, T. Jaakola, and M. I. Jordan. Approximating Posterior Distributions in Belief Networks Using Mixtures. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*; MIT Press, 1997.
26. G. Parisi. *Statistical Field Theory*. Addison-Wesley, 1988.
27. N. D. Lawrence, C. M. Bishop, and M. I. Jordan. Mixture Representations for Inference and Learning in Boltzmann Machines. In *Proceedings of the 15th Conference on Uncertainity in Artificial Intelligence (UAI)*; Morgan Kaufman, 1998; pp 320–327.
28. M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Mach. Learn.* **1999**, *37*, pp 183–233.
29. A. Corduneanu and C. M. Bishop. Variational Bayesian Model Selection for Mixture Distributions. In *Proceedings of the 8th International Conference on Artificial Intelligence and Statistics (AISTAT)*; Morgan Kaufman, 2001; pp 27–34.
30. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
31. G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual Categorization with Bags of Keypoints. In *Workshop on Statistical Learning in Computer Vision, 8th European Conference on Computer Vision (ECCV)*; Springer Science + Business Media, 2004.
32. A. Bosch, A. Zisserman, and X. Munoz. Scene Classification via pLSA. In *Proceedings of the 9th European Conference on Computer Vision (ECCV)*; Springer Science + Business Media, 2006; pp 517–530.
33. D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*(2), pp 91–110.
34. T. Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Mach. Learn.* **2001**, *42*(1/2), pp 177–196.
35. A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. J. Comput. Vis.* **2001**, *42*, pp 145–175.
36. OT database is available at http://cvcl.mit.edu/database.htm.
37. W. Fan and N. Bouguila. Learning Finite Beta-Liouville Mixture Models via Variational Bayes for Proportional Data Clustering. In *The 2013 International Joint Conference on Neural Networks*; Rossi, F., Ed.; IEEE, 2013.
38. W. Fan, N. Bouguila, and D. Ziou. Unsupervised Hybrid Feature Extraction Selection for High-Dimensional Non-Gaussian Data Clustering with Variational Inference. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*(7), pp 1670–1685.

WENTAO FAN
Huaqiao University, Quanzhou, China

NIZAR BOUGUILA
Concordia University, Montréal Quebec, Canada