High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length

Nizar Bouguila, *Member*, *IEEE*, and Djemel Ziou

Abstract—We consider the problem of determining the structure of high-dimensional data without prior knowledge of the number of clusters. Data are represented by a finite mixture model based on the generalized Dirichlet distribution. The generalized Dirichlet distribution has a more general covariance structure than the Dirichlet distribution and offers high flexibility and ease of use for the approximation of both symmetric and asymmetric distributions. This makes the generalized Dirichlet distribution more practical and useful. An important problem in mixture modeling is the determination of the number of clusters. Indeed, a mixture with too many or too few components may not be appropriate to approximate the true model. Here, we consider the application of the minimum message length (MML) principle to determine the number of clusters. The MML is derived so as to choose the number of clusters in the mixture model that best describes the data. A comparison with other selection criteria is performed. The validation involves synthetic data, real data clustering, and two interesting real applications: classification of Web pages, and texture database summarization for efficient retrieval.

Index Terms—Finite mixture models, generalized Dirichlet mixture, EM, information theory, MML, AIC, MDL, MMDL, LEC, data clustering, image database summarization, Web mining.

1 Introduction

Finite mixture models are being increasingly used in statistical inference, providing a formal approach to unsupervised learning [2], [3]. Fields in which mixture models have been successfully applied include image processing, pattern recognition, machine learning, and remote sensing [4]. The adoption of mixture models to clustering has important advantages; for instance, the selection of the number of clusters or a given model can be addressed in a formal way. Indeed, an important part of the modeling problem concerns determining the number of consistent components that best describe the data. For this purpose, many approaches have been suggested. From a computational point of view, these approaches can be classified into three classes: deterministic, stochastic, and resampling methods. Stochastic approaches include Markov Chain Monte Carlo (MCMC) methods, which can be used in two different ways for mixture models. The first is the implementation of model selection criteria [5], [6]. The second is fully Bayesian and consists of resampling from the full posterior distribution, with the number of clusters considered to be unknown [7]. To select the number of clusters, resampling schemes [8] and cross-validation approaches [9] have also been used. The deterministic methods can themselves be classified in two main classes. In the first, we have approximate Bayesian criteria like the Schwarz's Bayesian information criterion (BIC) [10] and the Laplace empirical criterion (LEC) [2]. The second class contains approaches based on information/coding theory concepts such as the minimum message length (MML) [11], [12], Akaike's information criterion (AIC) [13], [14], the different versions of the minimum description length (MDL) criterion, which have been developed by Rissanen in a series of papers [15], [16], [17], and the mixture minimum description length (MMDL) [18]. Note that the first version of MDL [15] coincides formally (but not conceptually) with BIC. A more detailed survey of selection criteria approaches can be found in [2]. In this paper, we are interested in deterministic methods, specifically, in MML, since the other two approaches (stochastic and resampling schemes) are still far too computationally demanding in computer vision and pattern recognition applications. According to Baxter and Oliver [19], the MML criterion gave better results than the AIC and MDL criteria for artificial mixtures of Gaussians, but Roberts et al. found that MML and MDL are almost identical for Gaussian distributions [20]. Interesting comparisons between MML and MDL can be found in [21, Section 10.2] and [22, Section 11.4.3].

- N. Bouguila is with the Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, 1455 de Maisonneuve Blvd. West, EV-007-632, QC H3G 1T7, Canada. E-mail: bouguila@ciise.concordia.ca.
- D. Ziou is with the Département d'Informatique, Faculté des Sciences, Université de Sherbrooke, 2500 Boulevard de l'Université, Sherbrooke, QC J1K 2R1, Canada. E-mail: djemel.ziou@usherbrooke.ca.

Manuscript received 15 June 2005; revised 12 Feb. 2006; accepted 21 Dec. 2006; published online 18 Jan. 2007.
Recommended for acceptance by Z. Ghahramani.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0319-0605. Digital Object Identifier no. 10.1109/TPAMI.2007.1095.

0162-8828/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

In this paper, we consider MML and the generalized Dirichlet mixture. Wallace and Boulton first applied MML encoding and produced a practical program for unsupervised classification called SNOB [11], [23], [24], [25], [26], [27], [28]. MML has been used especially in the case of Gaussian, Poisson, and von Mises circular mixtures [28], in the case of spatially correlated classes of Gaussian distributions [29], and, recently, in the case of Gamma [30], [31] and Student-t [32] mixtures. At the same time, other models such as generalized Dirichlet mixtures have not received much attention.

The rest of the paper is organized as follows: In Section 2, we present the generalized Dirichlet distribution in detail. We determine the MML expression for a generalized Dirichlet mixture in Section 3. The complete estimation and selection algorithm is given in Section 4. Section 5 is devoted to the experimental results when the MML approach is compared to other selection criteria.

2 The Generalized Dirichlet Mixture

In dimension d, the generalized Dirichlet probability density function (pdf) is defined by [33]

$$p(X_1,\dots,X_d)=\prod_{i=1}^{d}\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}X_i^{\alpha_i-1}\left(1-\sum_{j=1}^{i}X_j\right)^{\gamma_i} \tag{1}$$

for $\sum_{i=1}^{d}X_i<1$ and $0<X_i<1$ for $i=1\dots d$, where $\alpha_i>0$, $\beta_i>0$, $\gamma_i=\beta_i-\alpha_{i+1}-\beta_{i+1}$ for $i=1\dots d-1$, and $\gamma_d=\beta_d-1$. Note that the generalized Dirichlet distribution is reduced to a Dirichlet distribution when $\beta_i=\alpha_{i+1}+\beta_{i+1}$

$$p(X_1,\dots,X_d)=\frac{\Gamma(\alpha_1+\alpha_2+\dots+\alpha_d+\alpha_{d+1})}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_d)\Gamma(\alpha_{d+1})}\left(1-\sum_{i=1}^{d}X_i\right)^{\alpha_{d+1}-1}\prod_{i=1}^{d}X_i^{\alpha_d-1}, \tag{2}$$

where $\alpha_{d+1}=\beta_d$. The mean and variance of the Dirichlet distribution satisfy the following conditions:

$$E(X_i)=\frac{\alpha_i}{\sum_{l=1}^{d+1}\alpha_l}, \tag{3}$$

$$Var(X_i)=\frac{\alpha_i(\sum_{i=1}^{d+1}\alpha_i-\alpha_i)}{(\sum_{i=1}^{d+1}\alpha_i)^2(\sum_{i=1}^{d+1}\alpha_i+1)}, \tag{4}$$

and the covariance between X_i and X_j is

$$Cov(X_i,X_j)=-\frac{\alpha_i\alpha_j}{(\sum_{i=1}^{d+1}\alpha_i)^2(\sum_{i=1}^{d+1}\alpha_i+1)}. \tag{5}$$

Thus, any two random variables in $\vec{X}=(X_1,\dots,X_d)$ are negatively correlated, which is not always the case. Wong [33] studied the generalized Dirichlet distribution and showed that the general moment function is

$$E(X_1^{r_1},X_2^{r_2},\dots,X_d^{r_d})=\prod_{i=1}^{d}\frac{\Gamma(\alpha_i+\beta_i)\Gamma(\alpha_i+r_i)\Gamma(\beta_i+\delta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)\Gamma(\alpha_i+\beta_i+r_i+\delta_i)}, \tag{6}$$

where $\delta_i=r_{i+1}+r_{i+2}+\dots+r_d$ for $i=1,2,\dots,d-1$, and $\delta_d=0$. Then, we can show that the mean and the variance of the generalized Dirichlet distribution satisfy the following conditions [33]:

$$E(X_i)=\frac{\alpha_i}{\alpha_i+\beta_i\prod_{k=1}^{i-1}}\frac{\beta_k+1}{\alpha_k+\beta_k}, \tag{7}$$

$$Var(X_i)=E(X_i)\left(\frac{\alpha_i+1}{\alpha_i+\beta_i+1}\prod_{k=1}^{i-1}\frac{\beta_k+1}{\alpha_k+\beta_k+1}-E(X_i)\right), \tag{8}$$

and the covariance between X_i and X_j is

$$Cov(X_i,X_j)=E(X_j)\left(\frac{\alpha_i}{\alpha_i+\beta_i+1}\prod_{k=1}^{i-1}\frac{\beta_k+1}{\alpha_k+\beta_k+1}-E(X_i)\right). \tag{9}$$

Note that the generalized Dirichlet distribution has a more general covariance structure than the Dirichlet distribution [34]. In addition to these properties, it has been shown that $(X_1,\dots,X_l)$, for any $l<d$, follows an l-variate generalized Dirichlet distribution and that the generalized Dirichlet is conjugate to the multinomial distribution [33]. Compared to the Gaussian distribution, the generalized Dirichlet has a smaller number of parameters that makes the estimation and the selection more accurate as we will show in the experimental results. We note that the generalized Dirichlet distribution is defined in the compact support [0, 1] in contrast of the Gaussian, for example, which is defined in $\mathbb{R}$. However, we can generalize it easily to be defined in a compact support of the form $[A,B]$, where $(A,B)\in\mathbb{R}^2$ (see, for example, [35] in the case of Beta distribution). Having a compact support is an interesting property for a given density because of the nature of data in general. Generally, we model data that are compactly supported, such as data originating from videos, images, or text. Besides, as a generalization of the Dirichlet, this distribution offers high flexibility and ease of use for the approximation of both symmetric and asymmetric distributions and can be used in many applications such as image processing [34], biology [36], and text modeling [37]. Numerous other properties of this distribution are given in [38], [33]. A generalized Dirichlet mixture with M components is defined as

$$p(\vec{X}|\Theta)=\sum_{j=1}^{M}p\left(\vec{X}|\vec{\alpha}_j\right)p(j), \tag{10}$$

where $0<p(j)\leq 1$, and $\sum_{j=1}^{M}p(j)=1$. In this case, the parameters of a mixture for M clusters are denoted by $\Theta=(\alpha,\vec{P})$, where $\alpha=(\vec{\alpha}_1,\cdots,\vec{\alpha}_M)^T$, $\vec{\alpha}_j=(\alpha_{j1},\beta_{j1},\cdots,\alpha_{jd},\beta_{jd})$, $j=1,\cdots,M$, and $\vec{P}=(p(1),\cdots,p(M))^T$ is the mixing parameter vector.

3 The MML Criterion for a Generalized Dirichlet Mixture

From an information-theory point of view, the MML approach is based on evaluating statistical models according to their ability to compress a message containing the data.

High compression is obtained by forming good models of the data to be coded. For each model in the model space, the message includes two parts. The first part encodes the model, using only prior information about the model and no information about the data. The second part encodes only the data in a way that makes use of the model encoded in the first part [39]. Let us consider a set of data $\mathcal{X} = (\vec{X}_1, \vec{X}_2, \dots \vec{X}_N)$ controlled by a mixture of distributions with parameters Θ. The optimal number of clusters of the mixture is that which minimimizes the amount of information (measured in bits, if base-2 logarithm is used, or in nits, if natural logarithm is adopted [21]) needed to transmit $\mathcal{X}$ efficiently from a sender to a receiver. The message length is defined as minus the logarithm of the posterior probability (interesting discussion about the difference between maximizing a density and maximizing a probability can be found in [28]). The MML principle has strong connections with Bayesian inference and, hence, uses an explicit prior distribution over parameter values [40]. Wallace and Dowe [28] and Baxter and Oliver [19] give us the formula for the message length for a mixture of distributions

$$MessLen \simeq -\log(h(\Theta)) - \log\bigg(p(\mathcal{X}|\Theta)) + \frac{1}{2}\log(|F(\Theta)|\bigg) + \frac{N_p}{2}(1 + \log(\kappa_{N_p})), \tag{11}$$

where $h(\Theta)$ is the prior probability, $p(\mathcal{X}|\Theta)$ is the likelihood, $F(\Theta)$ is the expected Fisher information matrix, and $|F(\Theta)|$ is its determinant. N_p is the number of parameters to be estimated and is equal to $(2d+1)M$ in our case. κ_{N_p} is the optimal quantization lattice constant for $\mathbb{R}^{N_p}$ [41] and we have $\kappa_1 = 1/12 \simeq 0.083$ for $N_p = 1$. As N_p grows, κ_{N_p} tends to the asymptotic value given by $\frac{1}{2\pi e} \simeq 0.05855$. We note that κ_{N_p} does not vary much; thus, we can approximate it by $\frac{1}{12}$. More details and discussions about the MML principle can be found in [21], [42]. The estimation of the number of clusters is carried out by finding the minimum with regards to Θ of the message length $MessLen$. Note that the MML criterion is very similar, but conceptually different, to another one called the LEC [2]

$$\log(p(\mathcal{X}|\Theta)) - \log(h(\Theta)) - \frac{1}{2}N_p\log(2\pi) + \frac{1}{2}\log(|F(\Theta)|). \tag{12}$$

Apart from the lattice constant, the MML has the same form as the LEC [2]. In the following sections, we will calculate the determinant of the Fisher information matrix $|F(\Theta)|$ and the prior pdf $h(\Theta)$ for a mixture of generalized Dirichlet distributions.

3.1 Fisher Information for a Mixture of Generalized Dirichlet Distributions

The Fisher information matrix is the expected value of the Hessian minus the logarithm of the likelihood. It is difficult, in general, to obtain the expected Fisher information matrix of a mixture analytically [2], [43]. Then, we use the complete-data Fisher information matrix as proposed by Figueiredo and Jain in [12], that is, the Fisher information matrix is computed after the vectors in the data set are assigned to the different clusters [25], [26], [28]. The complete-data Fisher information matrix has a block-diagonal structure and its determinant is equal to the product of the determinant of the information matrix for each component times the determinant of the information matrix of $\vec{P}$

$$|F(\Theta)| \simeq |F(\vec{P})| \prod_{j=1}^{M} |F(\vec{\alpha}_j)|, \tag{13}$$

where $|F(\vec{P})|$ is the Fisher information with regards to the mixing parameters vector, and $|F(\vec{\alpha}_j)|$ is the Fisher information with regards to the vector $\vec{\alpha}_j$ of a single generalized Dirichlet distribution. In what follows, we will compute each of these separately. For $|F(\vec{P})|$, it should be noted that the mixing parameters satisfy the requirement $\sum_{j=1}^{M} p(j) = 1$. Consequently, it is possible to consider the generalized Bernoulli process with a series of trials, each of which has M possible outcomes labeled first cluster, second cluster, ..., Mth cluster. The number of trials of the jth cluster is a multinomial distribution of parameters $p(1), p(2), \dots, p(M)$. In this case, the determinant of the Fisher information matrix is [31]

$$|F(\vec{P})| = \frac{N^{M-1}}{\prod_{j=1}^{M} p(j)}, \tag{14}$$

where N is the number of data elements. For $F(\vec{\alpha}_j)$, let us consider the jth cluster $\mathcal{X}_j = (\vec{X}_t, \dots, \vec{X}_{t+n_j-1})$ of the mixture, where $t \leq N$, with parameter $\vec{\alpha}_j$. The choice of the jth cluster allows us to simplify the notation without loss of generality. The problem now is how to find the determinant of the Fisher information matrix with regards to the vector $\vec{\alpha}_j$. Indeed, we have a $(2 \times d) \times (2 \times d)$ matrix that is not easy to compute, especially for high-dimensional data. Here, we try to find an alternative method to overcome this difficulty by using an interesting property of the generalized Dirichlet distribution. If a vector $\vec{X}_i = (X_{i1}, \dots, X_{id})$ has a generalized Dirichlet distribution, then we can construct a vector $\vec{W}_i = (W_{i1}, \dots, W_{id})$ using the following geometric transformation T defined by

$$W_{il} = T(X_{il}) = \begin{cases} X_{il} & \text{if } l = 1 \\ X_{il}/(1 - X_{i1} - \dots - X_{il-1}) & \text{for } l = 2, 3, \dots, d. \end{cases} \tag{15}$$

In this vector $\vec{W}_i$, each $W_{il}, l = 1, \dots, d$ has a Beta distribution with parameters α_{jl} and β_{jl}, and the parameters $\{\alpha_{jl}, \beta_{jl}, l = 1, \dots, d\}$ define the generalized Dirichlet distribution that $\vec{X}_i$ follows [33]. Thus, the Fisher information with regards to the vector $\vec{\alpha}_j$ of a single generalized Dirichlet distribution is approximated by

$$|F(\vec{\alpha}_j)| \simeq \prod_{l=1}^{d} |F(\alpha_{jl}, \beta_{jl})|, \tag{16}$$

where $|F(\alpha_{jl}, \beta_{jl})|$ is the Fisher information of a single Beta distribution with parameters $(\alpha_{jl}, \beta_{jl})$. The Hessian matrix in the case of a Beta distribution with parameters $(\alpha_{jl}, \beta_{jl})$ is given by

$$H(\alpha_{jl},\beta_{jl}) = \begin{pmatrix} -\frac{\partial^2}{\partial^2\alpha_{jl}}\log(p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})) & -\frac{\partial^2}{\partial\alpha_{jl}\partial\beta_{jl}}\log(p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})) \\ -\frac{\partial^2}{\partial\beta_{jl}\partial\alpha_{jl}}\log(p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})) & -\frac{\partial^2}{\partial^2\beta_{jl}}\log(p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})) \end{pmatrix}, \tag{17}$$

where $\mathcal{W}_{jl} = (W_{tl}, \ldots, W_{(t+n_j-1)l})$, and p_{beta} is the Beta distribution, given by

$$p_{Beta}(W_{il}|\alpha_{jl},\beta_{jl}) = \frac{\Gamma(\alpha_{jl}+\alpha_{jl})}{\Gamma(\alpha_{jl})\Gamma(\alpha_{jl})} W_{il}^{\alpha_{jl}-1}(1-W_{il})^{\beta_{jl}-1}. \tag{18}$$

Thus (see the Appendix),

$$|F(\alpha_{jl},\beta_{jl})| = n_j^2\Big(\Psi'(\alpha_{jl})\Psi'(\beta_{jl}) - \Psi'(\alpha_{jl}+\beta_{jl})\Big(\Psi'(\alpha_{jl})+\Psi'(\beta_{jl})\Big)\Big). \tag{19}$$

By substituting (19) in (16), we obtain

$$|F(\vec{\alpha}_j)| \simeq n_j^{2d}\prod_{l=1}^{d}\Big(\Psi'(\alpha_{jl})\Psi'(\beta_{jl}) - \Psi'(\alpha_{jl}+\beta_{jl})\Big(\Psi'(\alpha_{jl})+\Psi'(\beta_{jl})\Big)\Big). \tag{20}$$

Once we have the Fisher information for a single generalized Dirichlet distribution, we can use it to calculate the Fisher information for a mixture of generalized Dirichlet distributions. By substituting (20) and (14) in (13), we obtain

$$|F(\Theta)| \simeq \frac{N^{M-1}}{\prod_{j=1}^{M}p(j)}\prod_{j=1}^{M}n_j^{2d}\Big[\prod_{l=1}^{d}\Big(\Psi'(\alpha_{jl})\Psi'(\beta_{jl}) - \Psi'(\alpha_{jl}+\beta_{jl})\Big(\Psi'(\alpha_{jl})+\Psi'(\beta_{jl})\Big)\Big)\Big], \tag{21}$$

$$\log(|F(\Theta)|) \simeq (M-1)\log(N) - \sum_{j=1}^{M}\log(p(j)) + 2d\sum_{j=1}^{M}\log(n_j) + \sum_{j=1}^{M}\sum_{l=1}^{d}\log\Big(\Big|\Psi'(\alpha_{jl})\Psi'(\beta_{jl}) - \Psi'(\alpha_{jl}+\beta_{jl})\Big(\Psi'(\alpha_{jl})+\Psi'(\beta_{jl})\Big)\Big|\Big). \tag{22}$$

3.2 Prior Distribution $h(\Theta)$

The performance of the MML criterion is dependent on the choice of the prior distribution $h(\Theta)$. In the absence of other knowledge about the mixture parameters, we model the parameters of the different components as a priori independent from the mixing probabilities, that is,

$$h(\Theta) = h(\vec{P})h(\alpha). \tag{23}$$

We know that the vector $\vec{P}$ is defined on the simplex $\{(p(1),\ldots,p(M)) : \sum_{j=1}^{M}p(j) = 1\}$; thus, a natural choice, as a prior, for this vector is the Dirichlet distribution

$$h(\vec{P}) = \frac{\Gamma(\sum_{j=1}^{M}\eta_j)}{\prod_{j=1}^{M}\Gamma(\eta_j)}\prod_{j=1}^{M}p(j)^{\eta_j-1}, \tag{24}$$

where $\vec{\eta} = (\eta_1,\ldots,\eta_M)$ is the parameter vector of the Dirichlet distribution. The choice of $\eta_1 = 1,\ldots,\eta_M = 1$ gives a uniform prior over the space $p(1)+\ldots+p(M) = 1$. This prior is given by [19], [28]

$$h(\vec{P}) = (M-1)!. \tag{25}$$

For $h(\alpha)$, since $\vec{\alpha}_j$, $j = 1\ldots M$ are assumed to be independent, we obtain

$$h(\alpha) = \prod_{j=1}^{M}h(\vec{\alpha}_j). \tag{26}$$

For $h(\vec{\alpha}_j)$, we know experimentally that $\sum_{l=1}^{d}(\alpha_{jl}+\beta_{jl}) < 2de^5$, so the vector $\vec{\alpha_j}$ can be defined on the simplex $\{(\alpha_{j1},\beta_{j1},\ldots,\alpha_{jd},\beta_{jd}) : \sum_{l=1}^{d}(\alpha_{jl}+\beta_{jl}) < 2de^5\}$. Then, we can consider as a prior a Dirichlet distribution with parameters $\vec{\eta} = (\eta_1,\ldots,\eta_{2d+1})$

$$h(\vec{\alpha}_j) = \frac{\Gamma(\sum_{l=1}^{2d+1}\eta_l)}{(2de^5)^{\sum_{l=1}^{2d+1}\eta_l-1}\prod_{l=1}^{2d+1}\Gamma(\eta_l)}\Big(2de^5 - \sum_{l=1}^{d}(\alpha_{jl}+\beta_{jl})\Big)^{\eta_{2d+1}-1}\prod_{l=0}^{d-1}\alpha_{jl+1}^{\eta_{2l+1}-1}\beta_{jl+1}^{\eta_{2l+2}-1}. \tag{27}$$

The choice of $\eta_1 = 1,\ldots,\eta_{2d+1} = 1$ gives a uniform prior

$$h(\vec{\alpha}_j) = \frac{(2d)!}{(2de^5)^{2d}} \tag{28}$$

and

$$h(\alpha) = \prod_{j=1}^{M}h(\vec{\alpha}_j) = (2de^5)^{-2Md}((2d)!)^M. \tag{29}$$

Substituting (29) and (25) in (23), we obtain

$$\log(h(\Theta)) = \sum_{j=1}^{M-1}\log(j) - 10Md - 2Md\log(2d) + M\sum_{j=1}^{2d}\log(j). \tag{30}$$

The expression of MML for a finite mixture of generalized Dirichlet distributions is obtained using (30), (22), and (11)

$$\begin{aligned} MessLen = MML(M) &= -\sum_{j=1}^{M-1}\log(j) + 10Md + 2Md\log(2d) - M\sum_{j=1}^{2d}\log(j) \\ &+ \frac{(M-1)\log(N)}{2} - \frac{1}{2}\sum_{j=1}^{M}\log(p(j)) + d\sum_{j=1}^{M}\log(n_j) \\ &- \log(p(\mathcal{X}|\Theta)) - \frac{N_p}{2}\log(12) + \frac{N_p}{2} \\ &+ \frac{1}{2}\sum_{j=1}^{M}\sum_{l=1}^{d}\log\Big(\Big|\Psi'(\alpha_{jl})\Psi'(\beta_{jl}) - \Psi'(\alpha_{jl}+\beta_{jl})\Big(\Psi'(\alpha_{jl})+\Psi'(\beta_{jl})\Big)\Big|\Big). \end{aligned} \tag{31}$$

TABLE 1
Parameters of the Different Generated Data Sets
(n_j Represents the Number of the Elements in Cluster j)

	j	α_{j1}	β_{j1}	α_{j2}	β_{j2}	n_j
Data set 1	1	12	50	35	20	100
	2	32	60	13	20	100
Data set 2	1	12	50	35	20	120
	2	32	60	13	20	120
	3	20	60	20	60	160
Data set 3	1	3	43	32	100	120
	2	70	100	5	55	120
	3	40	80	26	20	80
	4	15	90	50	50	80

TABLE 2
Parameters of the Different Generated Data Sets
(n_j Represents the Number of the Elements in Cluster j)

	j	α_{j1}	β_{j1}	α_{j2}	β_{j2}	n_j
Data set 4	1	3	43	32	100	100
	2	70	100	5	55	100
	3	40	80	26	20	100
	4	15	90	50	50	100
	5	20	60	20	60	100
Data set 5	1	3	43	32	100	200
	2	70	100	5	55	200
	3	40	80	26	20	200
	4	15	90	50	50	200
	5	20	60	20	60	100
	6	31	141	295	430	100
Data set 6	1	3	43	32	100	200
	2	70	100	5	55	200
	3	40	80	26	20	200
	4	15	90	50	50	100
	5	20	60	20	60	100
	6	31	141	295	430	100
	7	118	275	41	63	100

4 Estimation and Selection Algorithm

In this section, we summarize the algorithm for estimating the number of clusters for a mixture of generalized Dirichlet distributions. The input to this algorithm consists of a data set of vectors. Its output is the number of components and the estimated parameters. Normally, the estimation of the parameters is based on the minimization of the message length. However, as the MML estimates are very similar to the maximum likelihood (ML) estimates, we used the ML approach for the estimation of the mixture parameters [19]. The maximization defining the ML estimates is under the constraints $0 < p(j) \leq 1$ and $\sum_{j=1}^{M} p(j) = 1$. Obtaining ML estimates of the mixture parameters is possible through expectation-maximization (EM) and related techniques [43]. The EM algorithm [44] is a general approach to ML in the presence of incomplete data. In EM, the "complete" data are considered to be $Y_i = \{\vec{X}_i, \vec{Z}_i\}$, where $\vec{Z}_i = (Z_{i1}, \dots, Z_{iM})$ with

$$Z_{ij} = \begin{cases} 1 & \text{if } \vec{X}_i \text{ belongs to class } j \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

constituting the "missing" data. When we maximize the likelihood function, we do not obtain a closed-form solution for the α parameters. In [45], we have used the Fisher scoring method for the estimation of these parameters. This method involves the inverse of the $(2 \times d) \times (2 \times d)$ Fisher information matrix, which is not easy to compute, especially for high-dimensional data. Here, we try to find an alternative method to overcome this difficulty. As Scott and Thompson have observed, *the problem of density estimation in higher dimensions involves first of all finding where the action is* [46]. We therefore begin by identifying the important classes by an efficient initialization algorithm and use the interesting properties of the generalized Dirichlet distribution to refine the estimates. In order to estimate the α parameters, we have used the transformation given by (15). In the vector $\vec{W}_i$ obtained by this geometric transformation, each W_{il}, $l = 1, \dots, d$ has a Beta distribution with parameters α_{il} and β_{il}. The parameters $\{\alpha_{il}, \beta_{il}, l = 1, \dots, d\}$ define the generalized Dirichlet distribution of $\vec{X}_i$ [33]. The problem of estimating the parameters of a generalized Dirichlet mixture can thus be reduced to the estimation of the parameters of d Beta mixtures. This entails maximizing the following equation for every dimension l:

$$\Phi_W(\theta_l, \mathcal{W}) = \sum_{i=1}^{N} \log\Big(\sum_{j=1}^{M} p_{beta}(W_{il}|\theta_{jl})p(j)\Big), \tag{33}$$

where $\mathcal{W} = (W_{1l}, \dots, W_{Nl})$, $0 < l \leq d$, $\theta_l = (\alpha_{1l}, \beta_{1l}, \dots, \alpha_{Ml}, \beta_{Ml})$, $\theta_{jl} = (\alpha_{jl}, \beta_{jl})$, and $p(j)$ are the mixing parameters. To maximize (33), we resolve the following equations:

$$\frac{\partial}{\partial \alpha_{jl}} \Phi_W(\theta_l, \mathcal{W}) = 0 \quad \forall \quad 0 < l \leq d, \tag{34}$$

$$\frac{\partial}{\partial \beta_{jl}} \Phi_W(\theta_l, \mathcal{W}) = 0 \quad \forall \quad 0 < l \leq d. \tag{35}$$

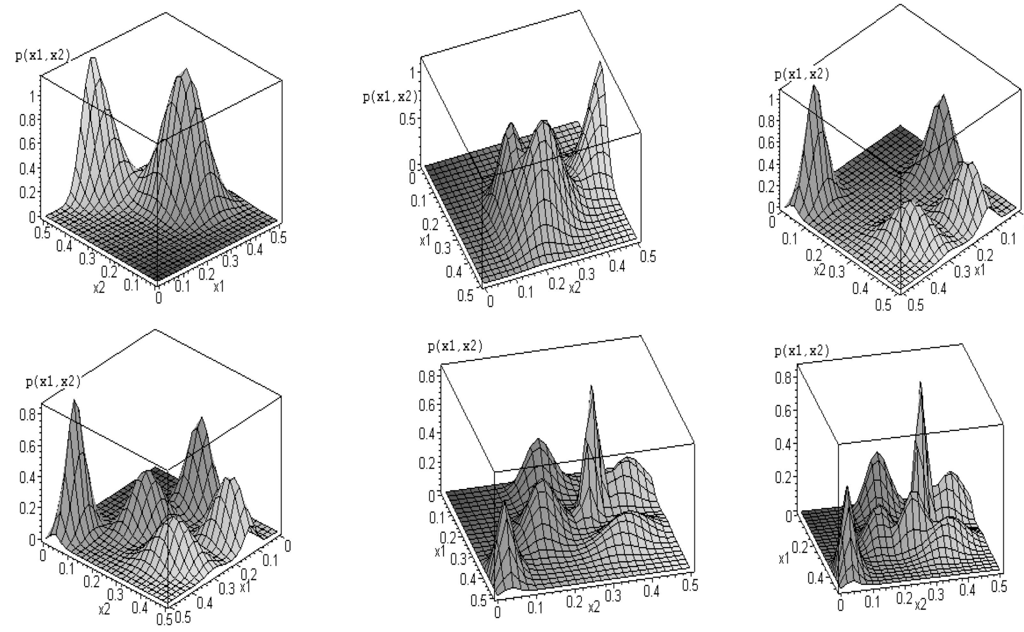


Fig. 1. Mixture densities for the generated data sets.

In order to estimate the θ_{jl} parameters, we will use Fisher's scoring method [47]. The scoring method is based on the first, second, and mixed derivatives of the function $\Phi_W(\theta_{jl}, \mathcal{W})$. We therefore compute these derivatives. Given a set of initial estimates, Fisher's scoring method can now be used. The iterative scheme of the Fisher method is given by the following equation:

$$\begin{pmatrix}\hat{\alpha'}_{jl}\\ \hat{\beta'}_{jl}\end{pmatrix}^{(t)} = \begin{pmatrix}\hat{\alpha'}_{jl}\\ \hat{\beta'}_{jl}\end{pmatrix}^{(t-1)} + V^{(t-1)} \times \begin{pmatrix}\frac{\partial\Phi_W}{\partial\hat{\alpha'}_{jl}}\\ \frac{\partial\Phi_W}{\partial\hat{\beta'}_{jl}}\end{pmatrix}^{(t-1)}, \qquad (36)$$

where j is the class number, $1 \leq j \leq M$, l is the current dimension, $1 \leq l \leq d$, and α'_{jl} and β'_{jl} are unconstrained real numbers. Indeed, we require that the α_{jl} and β_{jl} be strictly positive and we want the parameters upon which we will derive to be unconstrained so we reparametrize, setting $\alpha_{jl} = e^{\alpha'_{jl}}$ and $\beta_{jl} = e^{\beta'_{jl}}$.

The matrix V is obtained as the inverse of the Fisher information matrix $\mathbf{I}$. The information matrix $\mathbf{I}$ is

$$\mathbf{I} = \begin{pmatrix} -E[\frac{\partial^2}{\partial^2\alpha'_{jl}}\Phi_W(\theta_l, \mathcal{W})] & -E[\frac{\partial^2}{\partial\alpha'_{jl}\partial\beta'_{jl}}\Phi_W(\theta_l, \mathcal{W})] \\ -E[\frac{\partial^2}{\partial\beta'_{jl}\partial\alpha'_{jl}}\Phi_W(\theta_l, \mathcal{W})] & -E[\frac{\partial^2}{\partial^2\beta'_{jl}}\Phi_W(\theta_l, \mathcal{W})] \end{pmatrix}. \qquad (37)$$

Given sufficiently accurate starting values, the convergence of a sequence of iterates, produced by the Fisher scoring method, to a solution $\hat{\theta}_{jl}$ is locally quadratic. That is, given a norm $\|.\|$ on the parameter space, there is a constant h such that

$$\|\theta_{jl}^{(t)} - \hat{\theta}_{jl}\| \leq h\|\theta_{jl}^{(t-1)} - \hat{\theta}_{jl}\|^2 \qquad (38)$$

holds for $t = 1, 2, \ldots$. Quadratic convergence is very fast: This is regarded as the most important advantage of the Fisher scoring method allowing it to overcome the slow convergence of the EM algorithm. This rapid convergence can be improved by introducing a stochastic step in the EM algorithm [48], [49] that prevents the sequence of estimates Θ^t from staying near an unstable stationary point of the likelihood function [49]. In this step, each vector is assigned to a component j with probability $\hat{Z}_{ij}$. Then, we are using a partial assignment that is different of the total assignment used in [11], [50]. Interesting comparisons between partial and total assignments can be found in [23], [24], [25], [26], [28], [32]. In order to make our algorithm less sensitive to local maxima, we have used some initialization schemes including the Fuzzy C-Means and the method of moments (MM) [34]. Our initialization method can be resumed as follows:

INITIALIZATION Algorithm

1. INPUT: d-dimensional data $\vec{X}_i$, $i = 1, \ldots, N$ and number of clusters M.
2. Apply the Fuzzy C-Means to obtain the elements, covariance matrix, and mean of each component.
3. Compute the $\vec{W}_i = (W_{i1}, \ldots, W_{id})$ from the $\vec{X}_i$. $W_{i1} = X_{i1}$ and $W_{il} = X_{il}/V_{il-1}$ for $l = 2, 3 \ldots, d$, where $V_{il} = 1 - X_{i1} - X_{i2} - \ldots - X_{il}$.
4. Apply the MM for each component j and for each dimension l to obtain the vector of parameters $\vec{\theta}_{jl}$.
5. Assign the data to clusters, assuming that the current model is correct.
6. If the current model and the new model are sufficiently close to each other, terminate, else go to 4.

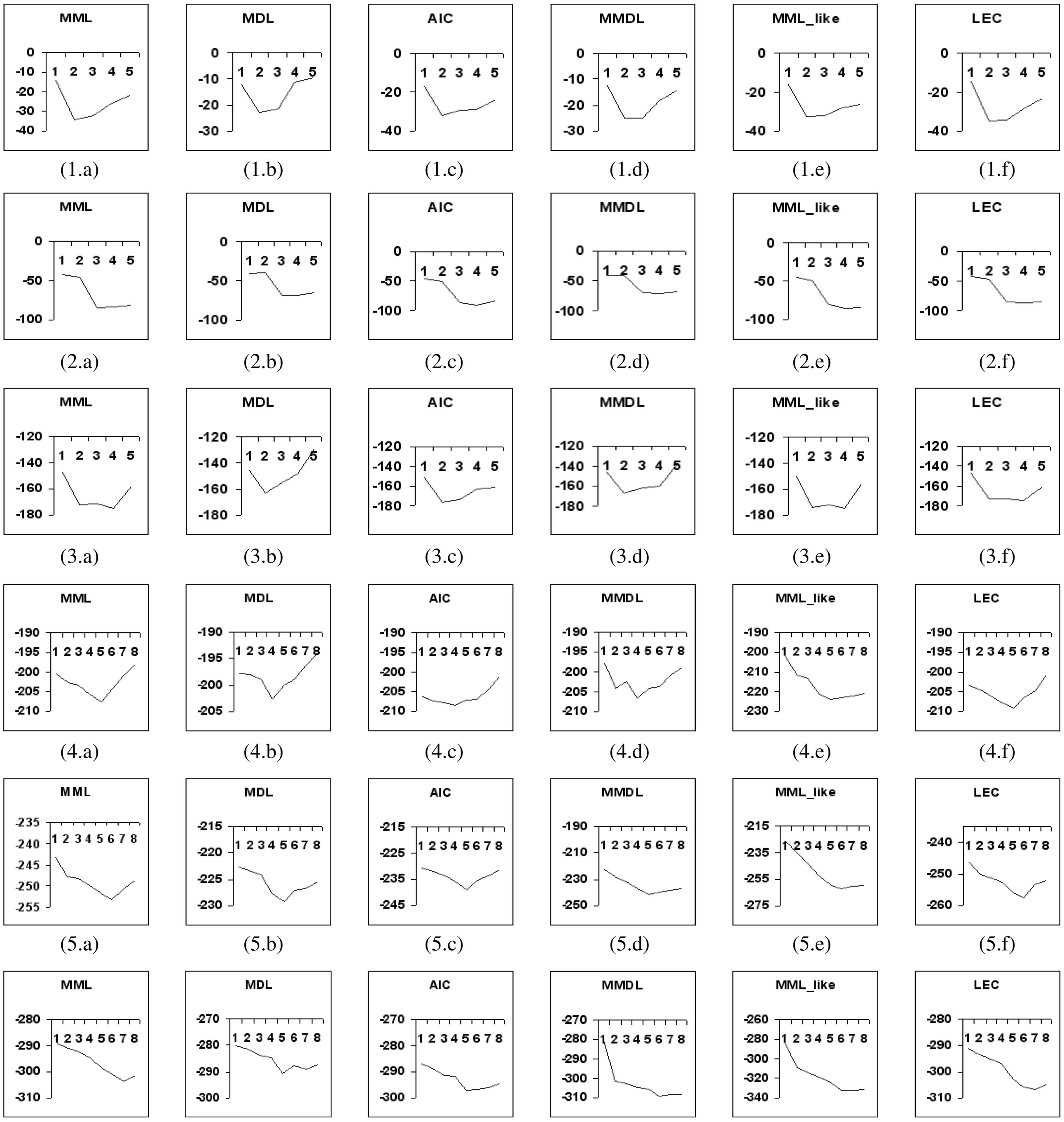
With this initialization method at hand, the complete estimation and selection algorithm is given as follows:

Algorithm

For each candidate value of M:

1. Apply the INITIALIZATION Algorithm.
2. E-Step: Compute the *posterior* probabilities:

$$\hat{Z}_{ij} = \frac{p(\vec{X}_i|\vec{\Theta}_j)p(j)}{\sum_{j=1}^{M} p(\vec{X}_i|\vec{\Theta}_j)p(j)}.$$



(1.a) (1.b) (1.c) (1.d) (1.e) (1.f)

(2.a) (2.b) (2.c) (2.d) (2.e) (2.f)

(3.a) (3.b) (3.c) (3.d) (3.e) (3.f)

(4.a) (4.b) (4.c) (4.d) (4.e) (4.f)

(5.a) (5.b) (5.c) (5.d) (5.e) (5.f)

(6.a) (6.b) (6.c) (6.d) (6.e) (6.f)

Fig. 2. Number of clusters found by the different criteria for the different generated data sets.

3. S-Step: For each sample value $\vec{X}_i$, draw $\vec{Z}_i$ from the multinomial distribution of order one with M categories having probabilities specified by the $\hat{Z}_{ij}$.
4. M-Step:
   - Update the $\vec{\theta}_{jl}$ using (36), $j = 1, \dots, M$ and $l = 1, \dots, d$.
   - Update the $p(j) = \frac{1}{N}\sum_{i=1}^{N} \hat{Z}_{ij}$, $j = 1, \dots, M$.
5. Calculate the associated criterion MML(M) using (31).
6. Select the optimal model M^* such that

$$M^* = arg\ min_M MML(M).$$

5 Experimental Results

5.1 Comparison with Other Criteria

Here, we will compare the results from the MML approach with those obtained using other deterministic model-order selection criteria/techniques. The methods we compare with are the MDL proposed by Rissanen in [15], the MMDL [18], the AIC [13], the MML-like criterion, which we call MML_{like}, proposed by Figueiredo and Jain in [12], and the LEC [2]. In general, the deterministic criteria can be expressed as

$$C(\hat{\Theta}(M), M) = -\log\Big[p\Big(\mathcal{X}|\hat{\Theta}(M)\Big)\Big] + f(M), \tag{39}$$

TABLE 3
Results for the First Generated Data Set
with 400 Simulations Draws

Number of clusters	MML	AIC	MDL	MMDL	MML$_{like}$	LEC
1	13	34	41	26	22	13
2 true	387	362	359	374	378	387
3	0	4	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0

where $f(M)$ is an increasing function that penalizes higher values of M. The optimal number of clusters is selected according to

$$\hat{M} = arg\,min\{C(\hat{\Theta}(M), M), M = M_{min}, \dots, M_{max}\}. \quad (40)$$

In spite of this common point, these criteria can be conceptually different. These criteria are given by the following equations:

$$MDL(M) = -\log(p(\mathcal{X}|\Theta)) + \frac{N_p}{2}\log(N), \quad (41)$$

where N_p is the number of parameters estimated, equal to $(2d+1)M$ in our case.

$$AIC(M) = -\log(p(\mathcal{X}|\Theta)) + \frac{N_p}{2}, \quad (42)$$

TABLE 4
Results for the Fourth Generated Data Set
with 400 Simulations Draws

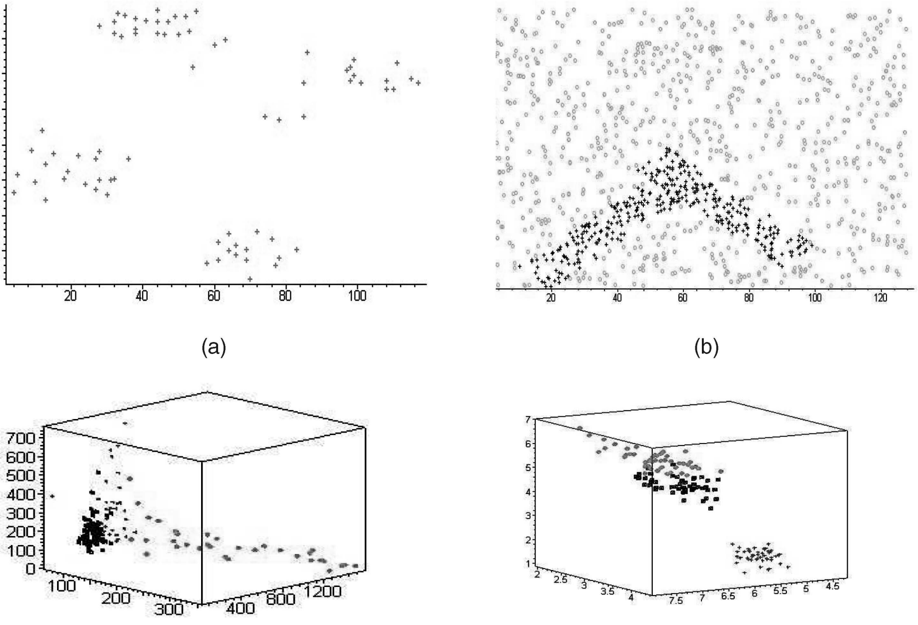
Number of clusters	MML	AIC	MDL	MMDL	MML$_{like}$	LEC
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	8	6	6	2	0
5	2	153	168	53	30	2
6	28	112	101	192	181	29
7 true	369	121	124	148	183	368
8	1	6	1	1	4	1

$$MMDL(M) = -\log(p(\mathcal{X}|\Theta)) + \frac{1}{2}N_p\log(N) + \frac{c}{2}\sum_{j=1}^{M}\log(p(j)), \quad (43)$$

where c is the number of parameters describing each component, equal to $2d+1$ in our case

$$MML_{like}(M) = -\log(p(\mathcal{X}|\Theta)) + \frac{M}{2}\log\left(\frac{N}{12}\right) + \frac{c}{2}\sum_{j=1}^{M}\log\left(N\frac{p(j)}{12}\right) + \frac{N_p}{2}.$$

To obtain the LEC selection criterion for a generalized Dirichlet mixture, we substitute (29) and (25) in (12)



(a)

(b)

(c)

(d)

Fig. 3. Examples of the data sets used. (a) The Ruspini data set. (b) The Chevron data set. (c) The Diabetes data set. (d) The Iris data set when we take the first three dimensions.

TABLE 5
Characteristics of the Data Used in the Second Application and the Results Given by Generalized Dirichlet and Gaussian Mixtures

Data set	Dimension	Samples	Classes	Accuracy	
				Generalized Dirichlet	Gaussian
Ruspini	2	75	4	98.67	97.34
chevron	2	350	2	91.95	88.72
Diabetes	3	145	3	94.22	91.88
Iris	4	150	3	98.66	97.33
Breast Cancer	9	683	2	98.97	97.65
Heart Disease	13	270	2	96.00	89.70

$$\begin{aligned} LEC(M) = & -\log(p(\mathcal{X}|\Theta)) - \sum_{j=1}^{M-1}\log(j) \\ & + 10Md + 2Md\log(2d) - M\sum_{j=1}^{2d}\log(j) \\ & + \frac{1}{2}\sum_{j=1}^{M}\sum_{l=1}^{d}\log(|\Psi'(\alpha_{jl})\Psi'(\beta_{jl}) \\ & - \Psi'(\alpha_{jd}+\beta_{jd})(\Psi'(\alpha_{jl})+\Psi'(\beta_{jl}))|) \\ & + d\sum_{j=1}^{M}\log(n_j) - \frac{1}{2}N_p\log(2\pi) \\ & + \frac{(M-1)\log(N)}{2} - \frac{1}{2}\sum_{j=1}^{M}\log(p(j)). \end{aligned} \tag{44}$$

5.2 Synthetic Data

In the first application, we investigate the properties of these model selection methods on six 2D synthetic data sets. We choose $d = 2$ purely for ease of representation. In fact, we tested the effectiveness of the methods for selecting the number of clusters by generating data sets using different parameters. We then attempted to estimate the parameters and the number of clusters of the mixtures representing these data sets. The parameters of these generated data sets are given in Tables 1 and 2. In Fig. 1, which represents the resultant mixtures, we see that we obtain different shapes (symmetric and asymmetric modes). Fig. 4 gives the number of clusters calculated for the generated data sets. In this figure, the values of the different criteria were averaged over 400 simulations draws. Tables 3 and 4 show the number of clusters calculated for the first and the sixth generated data sets for 400 simulation draws. The results presented for the generated data sets indicate clearly that the MML and LEC outperform the other criteria. This can be explained by the fact that these two criteria contain prior terms that the other criteria do not have. Note that the MML criterion is very similar to the LEC [2]. Indeed, the LEC criterion is reduced to the MML by taking uniform priors over the parameters and by choosing the asymptotic value $\kappa_{N_p} = \frac{1}{2\pi e}$ in (11). The MDL, for example, can be viewed as an approximation of the MML criterion. In fact, we have $F(\Theta) = NF^{(1)}(\Theta)$, where $F(\Theta)$ is the Fisher matrix of the entire population and $F^{(1)}(\Theta)$ is the Fisher matrix for a single observation. Therefore,

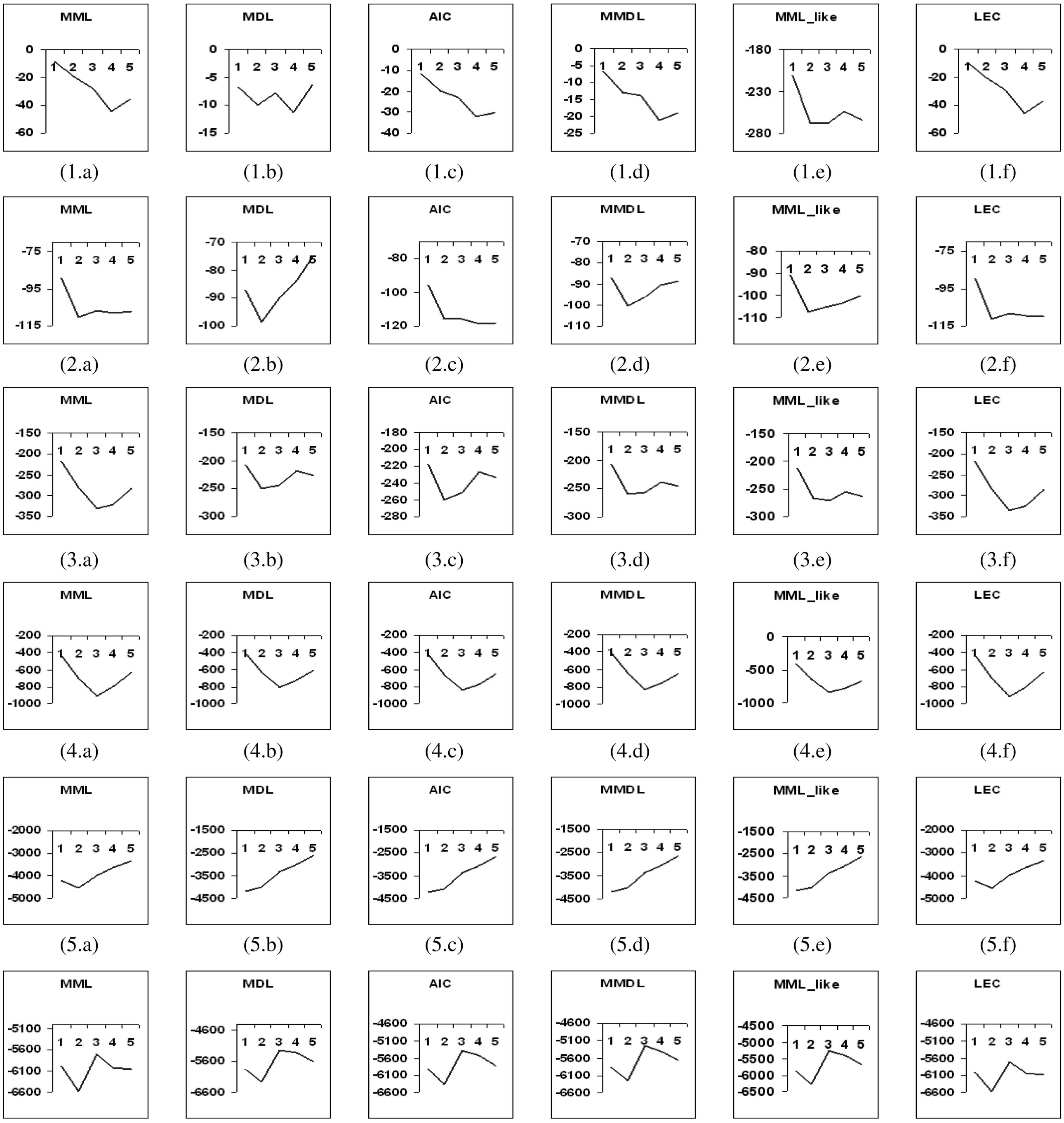
$$\begin{aligned} \log(|F(\Theta)|) &= \log\left(N^{N_p}|F^{(1)}(\Theta)|\right) \\ &= N_p\log(N) + \log\left(|F^{(1)}(\Theta)|\right), \end{aligned}$$

where $|F^{(1)}(\Theta)|$ is the Fisher information for a single observation. For a large N, we can remove the terms $\log(|F^{(1)}(\Theta)|)$ and $\frac{N_p}{2}(1-\log(12))$ from (11). Then, by assuming a flat prior $h(\Theta)$ and dropping it from (11), we obtain the well-known MDL selection criterion. The most important problem in using the MDL criterion is that all data points have equal importance in estimating each component of the parameter vector. This is not the case in mixtures, where each data point has its own weight in estimating different parameters. This point becomes apparent if we compute the Fisher matrix for the single $j\text{th}$ cluster of the mixture that leads to $F(\Theta_j) = Np(j)F^{(1)}(\Theta_j)$, where $F^{(1)}(\Theta_j)$ denotes the Fisher matrix associated with a single observation. As we have $|F(\Theta)| = \prod_{j=1}^{M}|F(\Theta_j)|$, we obtain

$$\begin{aligned} \log(|F(\Theta)|) &= \log\left(\prod_{j=1}^{M}(Np(j))^{c}\log\left(|F^{(1)}(\Theta_j)\right)|\right) \\ &= \sum_{j=1}^{M}c\log(Np(j)) + \sum_{j=1}^{M}\log\left(|F^{(1)}(\Theta_j)|\right), \end{aligned}$$

where c is the number of parameters defining each component. For a large N, we can drop the terms $\log(|F^{(1)}(\Theta_j)|)$ and $\frac{N_p}{2}(1-\log(12))$ from (11). We obtain

$$\begin{aligned} \log(|F(\Theta)|) &= \sum_{j=1}^{M}d(\log(N) + \log p(j)) \\ &= N_p\log(N) + c\sum_{j=1}^{M}\log p(j). \end{aligned}$$



(1.a) (1.b) (1.c) (1.d) (1.e) (1.f)

(2.a) (2.b) (2.c) (2.d) (2.e) (2.f)

(3.a) (3.b) (3.c) (3.d) (3.e) (3.f)

(4.a) (4.b) (4.c) (4.d) (4.e) (4.f)

(5.a) (5.b) (5.c) (5.d) (5.e) (5.f)

(6.a) (6.b) (6.c) (6.d) (6.e) (6.f)

Fig. 4. Number of clusters found by the different criteria for the different real data sets.

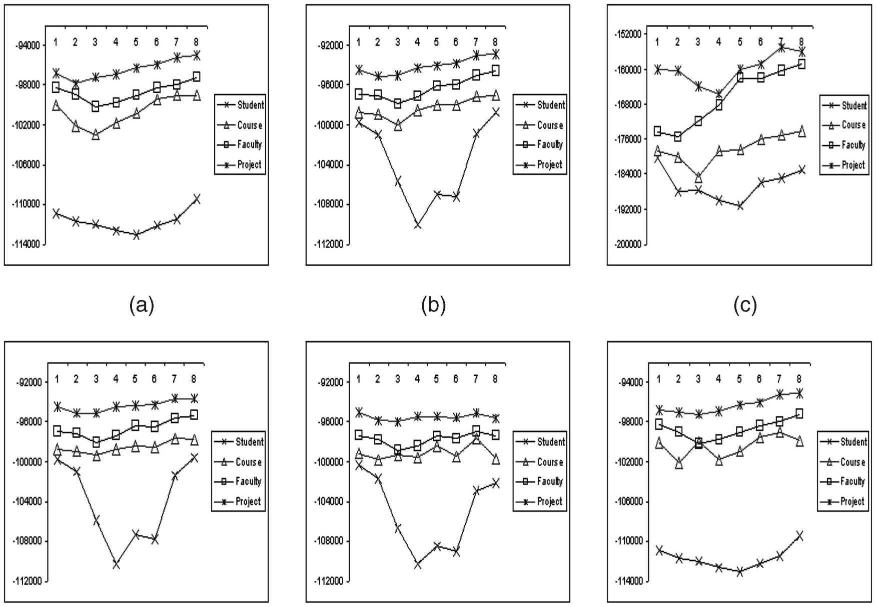
The result is the MMDL criterion. This is can explain the fact that the MMDL outperforms MDL. The MML_{like} criterion is also derived from the MML criterion by taking $F(\Theta_j) = Np(j)F^{(1)}(\Theta_j)$, $F(\vec{P}) = NF(Mul)$, where $F(Mul)$ is the Fisher matrix of a multinomial distribution, and assuming the following noninformative Jeffreys priors: $h(\Theta_j) \propto \sqrt{|F(\Theta_j)|}$ and

$$h(\vec{P}) \propto \sqrt{|F(Mul)|} = \frac{1}{\prod_{j=1}^{M} p(j)}$$

[12]. Note that we can show easily that $\mathrm{MML}_{like}(\mathrm{M}) = \mathrm{MMDL}(\mathrm{M}) + \frac{N_p}{2}(1 + \log(\frac{1}{12}))$. Then, the MML_{like} criterion can be obtained using the same approach used for MMDL but by keeping the order-1 term $\frac{N_p}{2}(1 + \log(\frac{1}{12}))$. This can explain the fact that the MML_{like} performance is comparable to, but slightly better than, MMDL. The AIC and MDL criteria perform comparably. However, AIC overfits (chooses the number of clusters greater than the true number) more often than MDL and the other criteria. This is explained by noting that AIC regularizes only with $0.5N_p$.

TABLE 6
Number of Clusters Determined by the Different Criteria Using Both Generalized Dirichlet and Gaussian Mixtures

		Ruspini	chevron	Diabetes	Iris	Breast Cancer	Heart Disease
Generalized Dirichlet	MML	**4**	**2**	**3**	**3**	**2**	**2**
	AIC	2	3	2	**3**	**2**	1
	MDL	2	**2**	2	**3**	**2**	1
	MMDL	2	**2**	2	**3**	**2**	1
	MML$_{like}$	3	**2**	**3**	**3**	**2**	1
	LEC	**4**	**2**	**3**	**3**	**2**	**2**
Gaussian	MML	**4**	**2**	4	**3**	**2**	1
	AIC	2	3	5	4	3	1
	MDL	2	**2**	**3**	**3**	1	1
	MMDL	2	**2**	**3**	**3**	1	1
	MML$_{like}$	2	**2**	**3**	**3**	1	1
	LEC	**4**	**2**	4	**3**	**2**	1



(a) (b) (c)

(d) (e) (f)

Fig. 5. Number of clusters determined to represent each of the four classes (Course, Faculty, Project, and Student) when we consider a training set of 2,000 Web pages. (a) MML. (b) MDL. (c) AIC. (d) MMDL. (e) MML$_{like}$. (f) LEC.

5.3 Real Data Clustering

In the second application, we validate our model using six standard multidimensional data sets (Ruspini, Chevron, Diabetes, Iris, Breast Cancer, and Heart Disease) that differ in dimension, size, and complexity. These data sets were obtained from the machine learning repository at the University of California, Irvine [51]. The Ruspini [52] data set contains 2D data in four groups (see Fig. 3a). Chevron is another 2D data set [53]. The data in Chevron arise from the processing of a series of images taken by a reconnaissance aircraft in which a large number of points are identified as representing possible mines, but many of these are in fact noise (see Fig. 3b). Diabetes is a 3D data set involving 145 observations used for diabetes diagnosis [54]. The data set is composed of three clusters that are overlapping and are far from spherical in shape (see Fig. 3c). Breast Cancer [55] is a nine-dimensional data set that contains two classes and 683 samples. Iris [56] comprises 50 samples for each of the three classes presented in the data, *Iris Versicolor*, *Iris Verginica*, and *Iris Setosa*; each datum is four-dimensional and consists of measures of the plants' morphology (see Fig. 3d). Heart disease is a 13-dimensional data set that contains two classes and 270 samples. Table 5 gives the characteristics of these data sets and the accuracy of classification (we give the exact number of clusters) when we use both the generalized Dirichlet and the Gaussian mixtures. We have considered a diagonal covariance matrix in the case of the Gaussian to avoid numerical problems [19]. Fig. 4 shows the number of clusters found by our algorithm when we use the six criteria for the different data sets. In these tables, we can see clearly that only the MML and LEC criteria found the correct number of clusters each time. The superiority of MML is also reported by Agusta and Dowe in [57], where they show that MML beats both AIC and BIC, in bit costing, with multivariate Gaussian mixtures, for the Iris and Diabetes data sets. Table 6 shows the number of clusters obtained when we use both the generalized Dirichlet and Gaussian mixtures.

5.4 Web Mining: Classification of Web Pages

The goal of this application is to understand the textual content of a Web page based on statistical features by considering single-word statistics, that is, the frequency of word occurrence [58]. The goal of this application is to prove the modeling capabilities of our algorithm. We begin by presenting the generalized Dirichlet mixture classifier. If the feature vectors $\vec{X}$ are annotated by providing class labels, we are able to perform supervised learning using the generalized Dirichlet mixture. Consider a data set $\mathcal{X}_l = \{(\vec{X}_i, C_i) | i = 1, \dots, N)\}$, where $C_i \in \{1, 2, \dots, M\}$, and M is the number of classes. The joint density of feature vectors $\vec{X}$ and class labels C is $p(\vec{X}, C) = p(\vec{X}|C)p(C)$, where $p(\vec{X}|C)$ is the class density, and $p(C)$ is the mixing probability. The classifier is designed by adapting the generalized Dirichlet mixtures to each class separately using the training data. Thus, the density of each class is itself a generalized Dirichlet mixture and can be written as

$$p(\vec{X}|C) = \sum_{k=1}^{M_c} p(\vec{X}|k, C)p(k|C), \qquad (45)$$

where M_c is the number of clusters calculated for class C using MML, $p(k|C)$ represent the mixing parameters, and $p(\vec{X}|k, C)$ is the generalized Dirichlet density. Labels are assigned to the test data using the Bayesian rule by selecting the maximum posterior probability given by the following equation:

$$p(C|\vec{X}) \propto p(\vec{X}|C)p(C). \qquad (46)$$

For the experiments we used the WebKB[1] data set, which contains Web pages gathered from university computer science departments. There are about 8,280 documents, and they are divided into seven categories: Student, Faculty, Staff, Course, Project, Department, and Other. Among these seven categories, student, faculty, staff, course, and project are the four most populous entity categories. The associated subset is typically called WebKB4 and contains 4,199 Web pages. In this paper, we perform experiments on the four-category data set: Course, Faculty, Project, and Student. In our experiments, we first select the top 200 words. The feature selection is done with the Rainbow package [59]. Suitable selection of the data is required for good performance. This concerns removing stop words and words that have a little influence (less than 50 occurrences in our experiments). Moreover, we keep only word stems and define the term vector as a complete set of words occurring in all the Web pages. A Web page histogram is the vector containing the frequency of occurrence of each word from the term vector and defines the content of the Web page. Normalizing all histogram vectors, each Web page i will be represented by a vector $\vec{X}_i = (X_{i1}, \dots, X_{i200})$, where X_{ij}, $j = 1, \dots, 200$ represents the probability of term j in document i. The data are then randomly split 10 times into a test set of ($N_{test} = 2, 199$) and training sets of increasing sizes, ($N_{train} = 1, 100 \dots 2, 000$). Fig. 5 shows the number of clusters determined by the different criteria used to represent each of the four classes (Course, Faculty, Project, and Student) when we consider a training set of 2,000 Web pages. Fig. 6 shows the learning curves for the different criteria. In this figure, we observe the classification error as a function of the number of documents in the training set. The proposed generalized Dirichlet mixture classifier achieves the best classification rate when the MML criterion is used to learn the training sets.

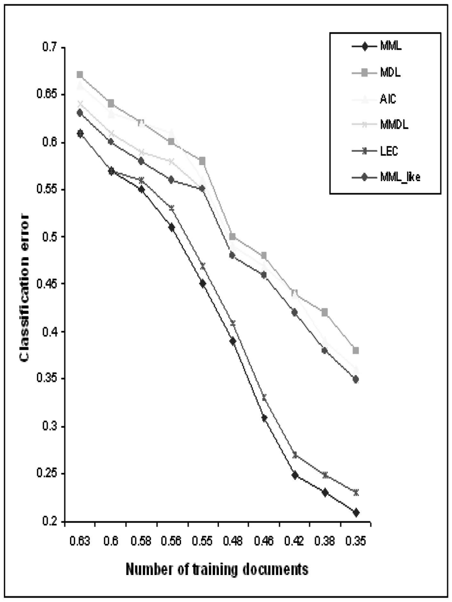


Fig. 6. Learning curves for the different selection criteria.

5.5 Texture Image Database Summarization for Efficient Retrieval

The fourth application concerns the summarization of texture image databases. Interactions between users and multimedia databases can involve queries like "Retrieve images that are similar to this image." A number of techniques have been

[1] This data set is available on the Internet. See http://www.cs.cmu.edu/~textlearning.

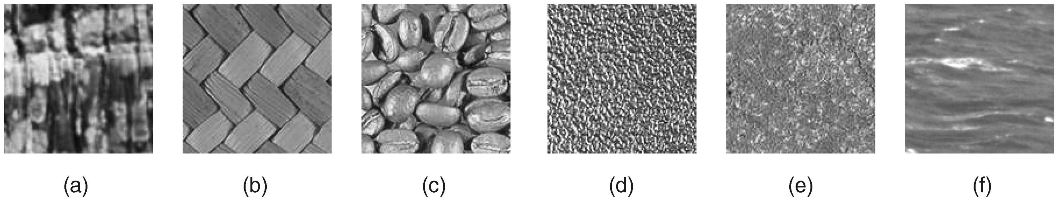


Fig. 7. Sample images from each group. (a) Bark. (b) Fabric. (c) Food. (d) Metal. (e) Sand. (f) Water.

developed to handle pictorial queries, for example, QBIC [60], Photobook [61], Blobworld [62], VisualSeek [63], and Atlas [64]. Summarizing the database is very important because it simplifies the task of retrieval by restricting the search for similar images to a smaller domain of the database [65]. Summarization is also very efficient for browsing [66]. Knowing the categories of images in a given database allows the user to find the images he or she is looking for more quickly. Using mixture decomposition, we can find natural groupings of images and represent each group by the most representative image in the group. In other words, after appropriate features are extracted from the images, the feature space can be partitioned into regions that are relatively homogeneous with respect to the chosen set of features. By identifying the homogeneous regions in the feature space, the task of summarization is accomplished. For the experiment described in this paper, we used the *Vistex* color texture database obtained from the Massachusetts Institute of Technology (MIT) Media Lab. In our experimental framework, each of the 512×512 images from the *Vistex* database was divided into 64×64 images. Since each 512×512 "mother image" contributes 64 images to our database, ideally, all of the 64 images should be classified in the same class. In the experiment, six homogeneous texture groups "Bark," "Fabric," "Food," "Metal," "Water," and "Sand" were used to create a new database. A database with 1,920 images was obtained. Four images from each of the Bark, Fabric, and Metal texture groups were used to obtain 256 images for each of these categories, and six images from Water, Food, and Sand were used to obtain 384 images for these categories. Examples of images from each of the categories are shown in Fig. 7. In order to determine the vector of characteristics for each image, we have computed a set of features derived from the correlogram [67]. It has been noted that to obtain good results, many correlograms should be computed, each one considering a given neighborhood and direction. Some studies show that considering the following neighborhoods is sufficient for co-occurrence matrices, in the case of gray-level images, to obtain good results in general: $(1;0)$, $(1;\frac{\pi}{4})$, $(1;\frac{\pi}{2})$, and $(1;\frac{3\pi}{4})$ [68]. For each of these neighborhoods, we calculated the corresponding correlogram and then derived from it the following features that have been proposed for co-occurrence matrices: Mean, Variance, Energy, Correlation, Entropy, Contrast, Homogeneity, and Cluster Prominence [69]. Thus, each image was characterized by a 36-dimensional vector. Applying our algorithm to the texture database, only the MML and LEC criteria found six categories (see Fig. 8). However, all the criteria failed to find the exact number of clusters when we use a Gaussian mixture with diagonal covariance matrices (four clusters in the case of MML and LEC, three clusters using MDL, MMDL, and MML_{like}, and seven clusters by the AIC). The classification was performed using the Bayesian decision rule after the class-conditional densities were estimated. The confusion matrix for the texture image classification application is given in Table 7. In this confusion matrix, the cell (*classi*, *classj*) represents the number of images

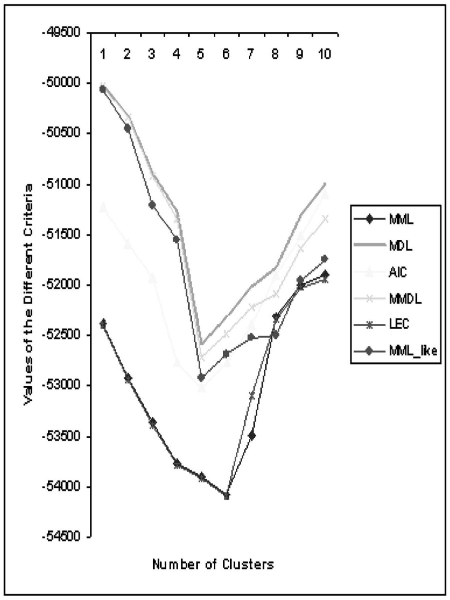


Fig. 8. Number of clusters found by each of the different criteria for the texture image database summarization.

TABLE 7
Confusion Matrix for Image Classification by a Generalized Dirichlet Mixture

	Bark	Fabric	Food	Metal	Sand	Water
Bark	254	0	0	0	2	0
Fabric	0	251	5	0	0	0
Food	0	8	376	0	0	0
Metal	0	0	0	250	0	6
Sand	3	0	0	0	381	0
Water	3	0	0	6	2	373

TABLE 8
Confusion Matrix for Image Classification by a Gaussian Mixture

	Bark	Fabric	Food	Metal	Sand	Water
Bark	240	0	0	3	8	5
Fabric	0	236	12	0	4	4
Food	0	12	365	4	0	3
Metal	0	2	2	242	4	6
Sand	8	2	0	0	370	4
Water	5	1	0	10	5	363

from *classi* that are classified as *classj*. The number of images misclassified was small: 35 in all, which represents an accuracy of 98.18 percent. Table 8 shows the confusion matrix for the Gaussian mixture when we suppose that we obtain the correct number of clusters (an accuracy of 94.59).

After the database was summarized, we conducted another experiment designed to retrieve images similar to a query. First, we defined a measure to determine the closest component to the query vector. Next, another distance measure was used to determine the similarity between the query vector and the feature vectors in the closest component. The *posterior* probabilities were used to choose the component nearest to the query. After selecting the closest component, the 2-norm was applied to find the images most similar to the query. To measure the retrieval rates, each image was used as a query, and the number of relevant images among those that were retrieved was noted. Precision and recall, which are the measures most commonly used by the information retrieval community, were then computed using (47) and (48). These measures were then averaged over all the queries and are defined as follows:

$$\text{precision} = \frac{\text{number of relevant retrieved images}}{\text{total number of retrieved images}}, \tag{47}$$

$$\text{recall} = \frac{\text{number of relevant retrieved images}}{\text{total number of relevant images}}. \tag{48}$$

As each 512×512 image from *Vistex* contributes 64 images to our database, given a query image, ideally, all 64 images should be retrieved and are considered to be relevant. Table 9 presents the retrieval rates obtained in terms of precision and recall. The results are shown when 16, 48, 64, 80, and 96 images were retrieved from the database in response to a query.

6 Conclusion

In this paper, we have focused on high-dimensional data clustering. We have presented an MML-based criterion to select the number of components in generalized Dirichlet mixtures. The algorithm proposed is motivated by the great number of pattern recognition and image processing applications that involve such types of data. In contrast with other methods that use dimensionality reduction, our algorithm uses the full dimensionality of the data. In fact, it is based on the statistical properties of the data through the use of generalized Dirichlet finite mixture models. The data is transformed in such a way that density estimation in the transformed space is simpler. The generalized Dirichlet distribution has the advantage that, by varying its parameters, it permits multiple modes and asymmetry and can thus approximate a wide variety of shapes. Besides, it has a more general covariance structure than the Dirichlet. Generalized Dirichlet mixtures allow more modeling flexibility than mixtures of Gaussians, without the explosion in the number of parameters. We estimated the parameters of this mixture using the ML and Fisher scoring methods and by introducing a stochastic step. The results presented indicate clearly that the MML and LEC model selection methods outperform the other methods. This can be explained by the fact that these two criteria contain prior terms that the others do not have. From the experimental results that involve generated data, real data set clustering, Web page classification, and texture image database summarization for efficient retrieval, we can say that the generalized Dirichlet distribution and the MML approach offer strong modeling capabilities for both low and high-dimensional data.

TABLE 9
Recall and Precision Obtained for the Texture Database

	No. of retrieved images				
Measure	16	48	64	80	96
Recall	0.24	0.74	0.93	0.95	0.98
Precision	0.97	0.98	0.93	0.84	0.73

Appendix

Proof of Equation (19)

We can write the negative of the log-likelihood function of a Beta distribution as follows:

$$\begin{aligned} -\log p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl}) &= -\log\left(\prod_{i=t}^{t+n_j-1} p_{Beta}(W_{il}|\alpha_{jl},\beta_{jl})\right) \\ &= -\sum_{i=t}^{t+n_j-1} \log p_{Beta}(W_{il}|\alpha_{jl},\beta_{jl}). \end{aligned} \tag{49}$$

By substituting (18) into (49), we obtain

$$\begin{aligned} &-\log p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl}) = \\ &\quad n_j\Big(-\log(\Gamma(\alpha_{jl}+\beta_{jl})) + \log(\Gamma(\alpha_{jl})) + \log(\Gamma(\beta_{jl}))\Big) \\ &\quad -\sum_{i=t}^{t+n_j-1}\Big((\alpha_{jl}-1)\log(W_{il}) + (\beta_{jl}-1)\log(1-W_{il})\Big) \end{aligned} \tag{50}$$

and we have

$$-\frac{\partial \log p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})}{\partial \alpha_{jl}} = n_j(-\Psi(\alpha_{jl}+\beta_{jl})+\Psi(\alpha_{jl})) - \sum_{i=t}^{t+n_j-1} \log(W_{il}), \tag{51}$$

$$-\frac{\partial \log p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})}{\partial \beta_{jl}} = n_j(-\Psi(\alpha_{jl}+\beta_{jl})+\Psi(\beta_{jl})) - \sum_{i=t}^{t+n_j-1} \log(1-W_{il}), \tag{52}$$

where Ψ is the digamma function. Then,

$$-\frac{\partial^2 \log p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})}{\partial \alpha_{jl}\partial \beta_{jl}} = -\frac{\partial^2 \log p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})}{\partial \beta_{jl}\partial \alpha_{jl}} = -n_j\Psi'(\alpha_{jl}+\beta_{jl}), \tag{53}$$

$$-\frac{\partial^2 \log p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})}{\partial^2 \alpha_{jl}} = -n_j(\Psi'(\alpha_{jl}+\beta_{jl})- \Psi'(\alpha_{jl})), \tag{54}$$

$$-\frac{\partial^2 \log p_{beta}(\mathcal{W}_{jl}|\alpha_{jl},\beta_{jl})}{\partial^2 \beta_{jl}} = -n_j(\Psi'(\alpha_{jl}+\beta_{jl}) - \Psi'(\beta_{jl})), \tag{55}$$

where Ψ' is the trigamma function. Thus,

$$|F(\alpha_{jl},\beta_{jl})| = n_j^2\Bigg(\Psi'(\alpha_{jl})\Psi'(\beta_{jl}) - \Psi'(\alpha_{jl}+\beta_{jl})\Bigg(\Psi'(\alpha_{jl})+\Psi'(\beta_{jl})\Bigg)\Bigg). \tag{56}$$

□

Acknowledgments

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council (NSERC) of Canada, Bell Canada's support through its Bell University Laboratories R&D program, and a start-up grant from Concordia University. The authors would like to thank the anonymous referees and the associate editor for their helpful comments. An earlier, shorter version of this work has been accepted for publication in [1].

References

[1] N. Bouguila and D. Ziou, "MML-Based Approach for High-Dimensional Learning Using the Generalized Dirichlet Mixture," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition—Workshops,* p. 53, 2005.

[2] G.J. McLachlan and D. Peel, *Finite Mixture Models.* John Wiley & Sons, 2000.

[3] B.S. Everitt and D.J. Hand, *Finite Mixture Distributions.* Chapman and Hall, 1981.

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

[5] H. Bensmail, G. Celeux, A. Raftery, and C. Robert, "Inference in Model-Based Cluster Analysis," *Statistics and Computing,* vol. 7, pp. 1-10, 1997.

[6] K. Roeder and L. Wasserman, "Practical Bayesian Density Estimation Using Mixture of Normals," *J. Am. Statistical Assoc.,* vol. 92, pp. 894-902, 1997.

[7] S. Richardson and P. Green, "On Bayesian Analysis of Mixtures with Unknown Number of Components," *J. Royal Statistic Soc. B,* vol. 59, pp. 731-792, 1997.

[8] G. McLachlan, "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture," *J. Royal Statistic Soc. C,* vol. 36, pp. 318-324, 1987.

[9] P. Smyth, "Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood," *Statistics and Computing,* vol. 10, no. 1, pp. 63-72, 2000.

[10] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics,* vol. 6, no. 2, pp. 461-464, 1978.

[11] C.S. Wallace and D.M. Boulton, "An Information Measure for Classification," *The Computer J.,* vol. 11, no. 2, pp. 195-209, 1968.

[12] M.A.T. Figueiredo and A.K. Jain, "Unsupervised Learning of Finite Mixture Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 3, pp. 4-37, Mar. 2002.

[13] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. Automatic Control,* vol. 9, no. 6, pp. 716-723, 1974.

[14] H. Bozdogan, "Determining the Number of Component Clusters in the Standard Multivariate Normal Mixture Model Using Model Selection Criteria," Technical Report A83-1, Quantitative Methods Dept., Univ. of Illinois, 1983.

[15] J. Rissanen, "Modeling by Shortest Data Description," *Automatica,* vol. 14, pp. 465-471, 1978.

[16] J. Rissanen, "Universal Coding, Information, Prediction and Estimation," *IEEE Trans. Information Theory,* vol. 30, no. 4, pp. 629-636, 1984.

[17] A.R. Barron, J. Rissanen, and B. Yu, "The Minimum Description Length Principle in Coding and Modeling," *IEEE Trans. Information Theory,* vol. 44, no. 6, pp. 2743-2760, 1998.

[18] M.A.T. Figueiredo, J.M.N. Leitao, and A.K. Jain, "On Fitting Mixture Models," *Energy Minimization Methods in Computer Vision and Pattern Recognition,* E. Hancock and M. Pellilo, eds. Springer, pp. 54-69, 1999.

[19] R.A. Baxter and J.J. Oliver, "Finding Overlapping Components with MML," *Statistics and Computing,* vol. 10, no. 1, pp. 5-16, 2000.

[20] S.J. Roberts, D. Husmeier, I. Rezek, and W. Penny, "Bayesian Approaches to Gaussian Mixture Modeling," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1133-1142, 1998.

[21] C.S. Wallace, *Statistical and Inductive Inference by Minimum Message Length.* Springer, 2005.

[22] J.W. Comley and D.L. Dowe, "Minimum Message Length, MDL and Generalised Bayesian Networks with Asymmetric Languages," *Advances in Minimum Description Length: Theory and Applications,* pp. 265-294, 2005.

[23] C.S. Wallace, "An Improved Program for Classification," *Proc. Ninth Australian Computer Science Conf.,* pp. 357-366, 1986.

[24] C.S. Wallace, "Classification by Minimum-Message-Length Inference," *Proc. Advances in Computing and Information,* S.G. Akl et al., eds., pp. 72-81, 1990.

[25] C.S. Wallace and D.L. Dowe, "Intrinsic Classification by MML—The Snob Program," *Proc. Seventh Australian Joint Conf. Artificial Intelligence,* pp. 37-44, 1994.

[26] C.S. Wallace and D.L. Dowe, "MML Mixture Modelling of Multi-State, Poisson, von Mises Circular and Gaussian Distributions," *Proc. Sixth Int'l Workshop Artificial Intelligence and Statistics,* pp. 529-536, 1997.

[27] C.S. Wallace and D.L. Dowe, "MML Mixture Modelling of Multi-State, Poisson, von Mises Circular and Gaussian Distributions," *Proc. 28th Symp. Interface, Computing Science and Statistics,* pp. 608-613, 1997.

[28] C.S. Wallace and D.L. Dowe, "MML Clustering of Multi-State, Poisson, von Mises Circular and Gaussian Distributions," *Statistics and Computing,* vol. 10, no. 1, pp. 73-83, 2000.

[29] C.S. Wallace, "Intrinsic Classification of Spatially Correlated Data," *The Computer J.,* vol. 41, no. 8, pp. 602-611, 1998.

[30] D. Ziou and N. Bouguila, "Unsupervised Learning of a Gamma Finite Mixture Using MML: Application to SAR Image Analysis," *Proc. 17th Int'l Conf. Pattern Recognition,* pp. 280-283, 2004.

[31] Y. Agusta and D.L. Dowe, "Unsupervised Learning of Gamma Mixture Models Using Minimum Message Length," *Proc. Third IASTED Conf. Artificial Intelligence and Applications,* M.H. Hamza, ed., pp. 457-462, 2003.

[32] Y. Agusta and D.L. Dowe, "MML Clustering of Continuous-Valued Data Using Gaussian and t Distributions," *Proc. 15th Australian Joint Conf. Artificial Intelligence,* B. McKay and J. Slaney, eds., pp. 143-154, 2002.

[33] T. Wong, "Generalized Dirichlet Distribution in Bayesian Analysis," *Applied Math. and Computation,* vol. 97, pp. 165-181, 1998.

[34] N. Bouguila, D. Ziou, and J. Vaillancourt, "Unsupervised Learning of a Finite Mixture Model Based on the Dirichlet Distribution and Its Application," *IEEE Trans. Image Processing,* vol. 13, no. 11, pp. 1533-1543, 2004.

[35] R.J. Beckman and G.L. Tietjen, "Maximum Likelihood Estimation for the Beta Distribution," *J. Statistics and Computational Simulation,* vol. 7, pp. 253-258, 1978.

[36] K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, and D. Haussler, "Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology," *Computer Applications in the Biosciences,* vol. 12, no. 4, pp. 327-345, 1996.

[37] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[38] R.J. Connor and J.E. Mosimann, "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution," *J. Am. Statistical Assoc.,* vol. 64, pp. 194-206, 1969.

[39] C.S. Wallace and D.L. Dowe, "Minimum Message Length and Kolmogorov Complexity," *The Computer J.,* vol. 42, no. 4, pp. 270-283, 1999.

[40] J.J. Oliver and R.A. Baxter, "MML and Bayesianism: Similarities and Differences," Technical Report 205, Dept. Computer Science, Monash Univ., July 1994.

[41] J. Conway and N. Sloane, *Sphere Packings, Lattice, and Groups.* Springer, 1993.

[42] C.S. Wallace and P.R. Freeman, "Estimation and Inference by Compact Coding," *J. Royal Statistical Soc. B,* vol. 49, pp. 240-252, 1987.

[43] G.J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions.* Wiley-Interscience, 1997.

[44] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc., B,* vol. 39, no. 1, pp. 1-38, 1977.

[45] N. Bouguila and D. Ziou, "A Powerful Finite Mixture Model Based on the Generalized Dirichlet Distribution: Unsupervised Learning and Applications," *Proc. 17th Int'l Conf. Pattern Recognition,* pp. 68-71, 2004.

[46] D.W. Scott and J.R. Thompson, "Probability Density Estimation in Higher Dimensions," *Computer Science and Statistics,* pp. 173-179, 1983.

[47] C.R. Rao, *Advanced Statistical Methods in Biomedical Research.* John Wiley & Sons, 1952.

[48] G. Celeux and J. Diebolt, "The SEM Algorithm: A Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem," *Computational Statistics Quarterly,* vol. 2, no. 1, pp. 73-82, 1985.

[49] G. Celeux and J. Diebolt, "A Stochastic Approximation Type EM Algorithm for the Mixture Problem," *Stochastics and Stochastics Reports,* vol. 41, pp. 119-134, 1992.

[50] R.T. Edwards and D.L. Dowe, "Single Factor Analysis in MML Mixture Modelling," *Proc. Second Pacific-Asia Conf. Knowledge Discovery and Data Mining,* pp. 96-109, 1998.

[51] C.L. Blake and C.J. Merz, *Repository of Machine Learning Databases.* Dept. Information and Computer Sciences, Univ. of California, Irvine, 1998, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[52] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data.* John Wiley & Sons, 1990.

[53] C. Fraley and A.E. Raftery, "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer J.,* vol. 41, no. 8, 1998.

[54] G.M. Reaven and R.G. Miller, "An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis," *Diabetologia,* vol. 16, pp. 17-24, 1979.

[55] R. Kothari and D. Pitts, "On Finding the Number of Clusters," *Pattern Recognition Letters,* vol. 20, pp. 405-416, 1999.

[56] E. Anderson, "The Irises of the Gaspe Peninsula," *Bull. Am. Iris Soc.,* vol. 59, pp. 2-5, 1935.

[57] Y. Agusta and D.L. Dowe, "Unsupervised Learning of Correlated Multivariate Gaussian Mixture Models Using MML," *Proc. 16th Australian Joint Conf. Artificial Intelligence,* T.D. Gedeon and L.C. Fung, eds., pp. 477-489, 2003.

[58] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, 1989.

[59] A.K. McCallum, "Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering," http://www.cs.cmu.edu/mccallum/bow, 1996.

[60] W. Niblack, R. Barber, W. Equitz, M. Flickner, E.H. Glasman, D. Yanker, P. Faloutsos, and G. Taubin, "The QBIC Project: Querying Images by Content Using Color, Texture and Shape," Technical Report RJ 9203, IBM, 1993.

[61] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-Based Manipulation of Image Databases," *Int'l J. Computer Vision,* vol. 18, no. 3, pp. 233-254, 1996.

[62] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 8, pp. 1026-1038, Aug. 2002.

[63] J.R. Smith and S.F. Chang, "VisualSEEK: A Fully Automated Content-Based Image Query System," *Proc. Fourth ACM Int'l Conf. Multimedia,* pp. 87-98, 1996.

[64] M.L. Kherfi, D. Ziou, and A. Bernardi, "Combining Positive and Negative Examples in Relevance Feedback for Content-Based Image Retrieval," *J. Visual Comm. and Image Representation,* vol. 14, pp. 428-457, 2003.

[65] A. Vailaya, M.A.T. Figueiredo, A.K. Jain, and H. Zhang, "Image Classification for Content-Based Indexing," *IEEE Trans. Image Processing,* vol. 10, no. 1, pp. 117-130, 2001.

[66] S. Newsman, B. Sumengen, and B.S. Manjunath, "Category-Based Image Retrieval," *Proc. Seventh IEEE Int'l Conf. Image Processing, Special Session on Multimedia Indexing, Browsing, and Retrieval,* 2001.

[67] J. Huang, S.R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image Indexing Using Color Correlograms," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* p. 762, 1997.

[68] T. Randen and J.H. Husoy, "Filtering for Texture Classification: A Comparative Study," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 4, pp. 291-310, Apr. 1999.

[69] M. Unser, "Sum and Difference Histograms for Texture Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 8, no. 1, pp. 118-125, 1986.

Nizar Bouguila received the BEng from the University of Tunis in 2000 and the MSc and PhD degrees from Sherbrooke University in 2002 and 2006, respectively, all in computer science. He is currently an assistant professor at the Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, Quebec. In 2007, Dr. Bouguila received the best PhD thesis award in engineering and natural sciences from Sherbrooke University and was a runner-up for the prestigious NSERC doctoral prize. His research interests include image processing, machine learning, 3D graphics, computer vision, and pattern recognition. He is a member of the IEEE.

Djemel Ziou received the BEng degree in computer science from the University of Annaba, Algeria, in 1984 and the PhD degree in computer science from the Institut National Polytechnique de Lorraine (INPL), France, in 1991. From 1987 to 1993, he served as a lecturer in several universities in France. During the same period, he was a researcher at the Centre de Recherche en Informatique de Nancy (CRIN) and the Institut National de Recherche en Informatique et Automatique (INRIA) in France. Presently, he is a full professor in the Department of Computer Science, Université de Sherbrooke, Quebec, Canada. He is the holder of the Natural Sciences and Engineering Research Council (NSERC)/Bell Canada Research chair in personal imaging. He has served on numerous conference committees as a member or chair. He heads the laboratory MOIVRE and the consortium CoRIMedia, which he founded. His research interests include image processing, information retrieval, computer vision, and pattern recognition.