On Fitting Finite Dirichlet Mixture Using ECM and MML

Nizar Bouguila and Djemel Ziou

Université de Sherbrooke,
Sherbrooke, Qc, Canada J1K 2R1
{nizar.bouguila, djemel.ziou}@usherbrooke.ca

Abstract. Gaussian mixture models are being increasingly used in pattern recognition applications. However, for a set of data other distributions can give better results. In this paper, we consider Dirichlet mixtures which offer many advantages [1]. The use of the ECM algorithm and the minimum message length (MML) approach to fit this mixture model is described. Experimental results involve the summarization of texture image databases.

1 Introduction

Finite mixture models have continued to receive increasing attention over the years [2]. These models are used in various fields such as image processing, pattern recognition, machine learning and remote sensing. For multivariate data attention has focused on the use of Gaussian components. However, for many applications the Gaussian can fail when the partitions are clearly non-Gaussian. In [1], we have demonstrated that the Dirichlet can be a good choice to overcome the problems of the Gaussian. In dimension dim the Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{dim+1})$ is given by:

$$p(\boldsymbol{X}|\boldsymbol{\alpha}) = \frac{\Gamma(|\boldsymbol{\alpha}|)}{\prod_{i=1}^{dim+1}\Gamma(\alpha_i)} \prod_{i=1}^{dim+1} X_i^{\alpha_i - 1} \tag{1}$$

where $\sum_{i=1}^{dim} X_i < 1$, $|\boldsymbol{X}| = \sum_{i=1}^{dim} X_i$, $0 < X_i < 1 \ \forall i = 1 \ldots dim$, $X_{dim+1} = 1 - |\boldsymbol{X}|$, and $|\boldsymbol{\alpha}| = \sum_{i=1}^{dim+1} \alpha_i$, $\alpha_i > 0 \quad \forall i = 1 \ldots dim + 1$. This distribution is the multivariate extension of the 2-parameter Beta distribution. The mean of the Dirichlet distribution is given by:

$$\mu_i = E(X_i) = \frac{\alpha_i}{|\boldsymbol{\alpha}|} \tag{2}$$

A mixture with M components is defined as : $p(\boldsymbol{X}|\Theta) = \sum_{j=1}^{M} p(\boldsymbol{X}|\boldsymbol{\alpha}_j)p(j)$ where $p(j)$ $(0 < p(j) < 1$ and $\sum_{j=1}^{M} p(j) = 1)$ are the mixing parameters and $p(\boldsymbol{X}|\boldsymbol{\alpha}_j)$ is the Dirichlet distribution. The symbol Θ refers to the entire set of parameters to be estimated: $\Theta = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_M, p(1), \ldots, p(M))$, where $\boldsymbol{\alpha}_j$ is the

S. Singh et al. (Eds.): ICAPR 2005, LNCS 3686, pp. 172–182, 2005.

parameters vector of the j^{th} component. The EM algorithm is a popular method for iterative maximum likelihood (ML) estimation of finite mixture distributions. This algorithm, however, is unattractive when the M-Step is complicate [2]. This is the case of the Dirichlet mixture. Indeed, the M-Step involves the inverse of the $(dim + 1) \times (dim + 1)$ Fisher information matrix which is not easy to compute especially for high-dimensional data. In this paper, we introduce another approach based on the ECM algorithm which replace a complicated M-step of the EM algorithm with several computationally simpler CM-Steps [3]. The determination of the number of components is based on the MML approach. The rest of the paper is organized as follows. Section II, discusses the basic concepts of the EM algorithm and proposes the ECM algorithm as a method to overcome the problems of the EM in the case of Dirichlet mixtures. In Section III, we present the MML approach for the selection of the number of clusters. Section IV is devoted to experimental results, and Section V ends the paper with some concluding remarks.

2 ML Estimation of a Dirichlet Mixture Using ECM

We consider now ML estimation for a M-component mixture of Dirichlet distributions. Given the set of independent vectors $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$, the log-likelihood corresponding to an M-component mixture is:

$$L(\Theta, \mathcal{X}) = log \prod_{i=1}^{N} p(\boldsymbol{X_i}|\Theta) = \sum_{i=1}^{N} log \sum_{j=1}^{M} p(\boldsymbol{X}_i|\boldsymbol{\alpha}_j)p(j) \tag{3}$$

It's well-known that the ML estimate: $\hat{\Theta}_{ML} = argmax_{\Theta}\{L(\Theta, \mathcal{X})\}$ which can not be found analytically. The maximization defining the ML estimates is subject to 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 EM and related techniques [2]. The EM algorithm is a general approach to maximum likelihood in the presence of incomplete data. In EM, the "complete" data are considered to be $Y_i = \{\boldsymbol{X}_i, \boldsymbol{Z}_i\}$, where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ with $Z_{ij} = 1$ if $\boldsymbol{X_i}$ belongs to class j and $Z_{ij} = 0$ otherwise. The relevant assumption is that the density of an observation $\boldsymbol{X}_i$ given $\boldsymbol{Z}_i$ is given by $\prod_{j=1}^{M} p(\boldsymbol{X_i}|\boldsymbol{\alpha}_j)^{Z_{ij}}$. The resulting *complete-data log-likelihood* is:

$$L(\Theta, \mathcal{Z}, \mathcal{X}) = \sum_{i=1}^{N} \sum_{j=1}^{M} Z_{ij} log(p(\boldsymbol{X_i}|\boldsymbol{\alpha}_j)p(j)) \tag{4}$$

The EM algorithm produces a sequence of estimates $\{\Theta^t, t = 0, 1, 2 \ldots\}$ by applying two steps in alternation (until some convergence criterion is satisfied):

1. **E-step:** Compute $\hat{Z}_{ij}$ given the parameter estimates from the initialization:

$$\hat{Z}_{ij} = \frac{p(\boldsymbol{X_i}|\boldsymbol{\alpha}_j)p(j)}{\sum_{j=1}^{M} p(\boldsymbol{X_i}|\boldsymbol{\alpha}_j)p(j)} \tag{5}$$

2. **M-step:** Update the parameter estimates according to:

$$\hat{\Theta} = argmax_{\Theta} L(\Theta, \mathcal{Z}, \mathcal{X}) \tag{6}$$

The quantity $\hat{Z}_{ij}$ is the conditional expectation of Z_{ij} given the observation $\boldsymbol{X_i}$ and parameter vector Θ. The value Z^*_{ij} of $\hat{Z}_{ij}$ at a maximum of Eq. 4 is the conditional probability that observation i belongs to class j (the *posterior* probability); the classification of an observation X_i is taken to be $\{k/Z^*_{ik} = max_j Z^*_{ij}\}$, which is the Bayes rule. The EM algorithm has been shown to monotonically increase the log-likelihood function. When we maximize Eq. 6, we obtain:

$$p(j)^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \hat{Z}_{ij}^{(t-1)} \tag{7}$$

However, we do not obtain a closed-form solution for the $\boldsymbol{\alpha}_j$ parameters. We therefore use the Fisher scoring method to estimate these parameters [1]. The inconvenient of this approach is that it involves the inverse of the $(dim + 1) \times (dim + 1)$ Fisher information matrix which is not easy to compute especially for high-dimensional data. One of reasons of the popularity of the EM algorithm is that the M-step involves only complete-data ML estimation. But, if the M-Step is complicated as in the case of the Dirichlet mixture, the EM algorithm becomes less attractive. In many cases, however, the ML estimation is simpler if maximization is undertaken conditional on some functions of the parameters. For this goal, Meng and Rubin [3] introduced an algorithm called ECM which replaces a complicated M-step of the EM algorithm with several computationally simpler CM-Steps. As a consequence the ECM converges more slowly than the EM in terms of number of iterations, but can be faster in total computer time. Another important advantage of the ECM is the preservation of the convergence properties of the EM, such as its monotone convergence. Now, we focus on the use of this algorithm for the estimation of Dirichlet mixture.
By substituting Eq. 2 in Eq. 1, the Dirichlet distribution can be written as the following:

$$p(\boldsymbol{X}||\boldsymbol{\alpha}|, \boldsymbol{\mu}) = \frac{\Gamma(|\boldsymbol{\alpha}|)}{\prod_{i=1}^{dim+1} \Gamma(\mu_i|\boldsymbol{\alpha}|)} \prod_{i=1}^{dim+1} X_i^{\mu_i|\boldsymbol{\alpha}|-1} \tag{8}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{dim+1})$. By this reparameterization, the parameters of the Dirichlet mixture to estimate will be $\xi = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M, |\boldsymbol{\alpha}_1|, \ldots, |\boldsymbol{\alpha}_M|, p(1), \ldots, p(M))$. This set of parameters can be divided into three subsets $\xi_1 = (|\boldsymbol{\alpha}_1|, \ldots, |\boldsymbol{\alpha}_M|)$, $\xi_2 = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M)$, and $\xi_3 = (p(1), \ldots, p(M))$. Then, the different parameters ξ_1, ξ_2 and ξ_3 can be calculated independently. The likelihood for ξ_1 alone is:

$$p(\mathcal{X}|\xi_1) \propto \prod_{i=1}^{N} \left[\sum_{j=1}^{M} p(j) \frac{\Gamma(|\boldsymbol{\alpha}_j|)}{\prod_{l=1}^{dim+1} \Gamma(\mu_{jl}|\boldsymbol{\alpha}_j|)} \prod_{l=1}^{dim+1} X_{il}^{\mu_{jl}|\boldsymbol{\alpha}_j|-1} \right] \tag{9}$$

For the estimation of $|\boldsymbol{\alpha_j}|$, we use a Newton-Raphson method:

$$|\boldsymbol{\alpha_j}|^{(t)} = |\boldsymbol{\alpha_j}|^{(t-1)} - \left(\frac{\partial^2 logp(\mathcal{X}|\xi_1^{(t-1)})}{\partial^2 |\boldsymbol{\alpha_j}|}\right)^{-1} \frac{\partial logp(\mathcal{X}|\xi_1^{(t-1)})}{\partial |\boldsymbol{\alpha_j}|} \tag{10}$$

The likelihood for ξ_2 alone is:

$$p(\mathcal{X}|\xi_2) \propto \prod_{i=1}^{N} \left[\sum_{j=1}^{M} p(j) \prod_{l=1}^{dim+1} \frac{X_{il}^{\mu_{jl}|\boldsymbol{\alpha}_j|-1}}{\Gamma(\mu_{jl}|\boldsymbol{\alpha}_j|)} \right] \tag{11}$$

By maximizing $p(\mathcal{X}|\xi_2)$ taking into account the constraint $\sum_{l=1}^{dim+1} \mu_{jl} = 1$, we obtain:

$$\mu_{jl}^{(t)} = \frac{\mu_{jl}^{(t-1)} \sum_{i=1}^{N} p(\boldsymbol{\mu_j}^{(t-1)}|\boldsymbol{X_i}) \Big(log(X_{il}) - \Psi(\mu_{jl}^{(t-1)}|\boldsymbol{\alpha_j}|^{(t)}) \Big)}{\sum_{l=1}^{dim+1} \left[\mu_{jl}^{(t-1)} \sum_{i=1}^{N} p(\boldsymbol{\mu_j}^{(t-1)}|\boldsymbol{X_i}) \Big(log(X_{il}) - \Psi(\mu_{jl}^{(t-1)}|\boldsymbol{\alpha_j}|^{(t)}) \Big) \right]} \tag{12}$$

Then, on the iteration t of the ECM algorithm, the E-Step is the same as given above for the EM algorithm, but the M-Step is replaced by three CM-Steps, as follows:

- **CM-Step1:** Calculate $\xi_1^{(t)}$ using Eq. 10 with ξ_2 fixed at $\xi_2^{(t-1)}$ and ξ_3 fixed at $\xi_3^{(t-1)}$.
- **CM-Step2:** Calculate $\xi_2^{(t)}$ using Eq. 12 with ξ_1 fixed at $\xi_1^{(t)}$ and ξ_3 fixed at $\xi_3^{(t-1)}$.
- **CM-Step2:** Calculate $\xi_3^{(t)}$ using Eq. 7 with ξ_1 fixed at $\xi_1^{(t)}$ and ξ_2 fixed at $\xi_2^{(t)}$.

3 MML Approach for the Determination of the Number of Clusters

3.1 MML Principle

Let us consider a set of data $\mathcal{X}$ controlled by a mixture of distributions with vector of parameters ξ. According to information theory [4], the optimal number of clusters of the mixture is that which requires a minimum amount of information, measured in nats, to transmit $\mathcal{X}$ efficiently from a sender to a receiver. The message length is defined as minus the logarithm of the posterior probability.

$$MessLen = -log(P(\xi|\mathcal{X})) \tag{13}$$

The MML principle has strong connections with Bayesian inference, and hence uses an explicit prior distribution over parameter values. Wallace [5] and Baxter [6] give us the formula for the message length for a mixture of distributions:

$$MessLen \simeq -log(h(\xi)) - log(p(\mathcal{X}|\xi)) + \frac{1}{2}log(|F(\xi)|) - \frac{N_p}{2}log(12) + \frac{N_p}{2} \quad (14)$$

where $h(\xi)$ is the prior probability, $p(\mathcal{X}|\xi)$ is the likelihood, and $|F(\xi)|$ is the Fisher information, defined as the determinant of the Hessian matrix of minus the log-likelihood of the mixture. N_p is the number of parameters to be estimated and is equal to $M(dim+3)$ in our case. The estimation of the number of clusters is carried out by finding the minimum with regards to ξ of the message length $MessLen$. We will determine the expression of MML for a Dirichlet mixture.

3.2 Fisher Information for a Mixture of Dirichlet Distributions

Fisher information is the determinant of the Hessian matrix of the logarithm of minus the likelihood of the mixture. The Hessian matrix of a mixture leads to a complicated analytical form of MML which cannot be easily reproduced. We will approximate this matrix by formulating two assumptions, as follows. First, it should be recalled that (ξ_1, ξ_2) and ξ_3 are independent because any prior idea one might have about (ξ_1, ξ_2) would usually not be greatly influenced by one's idea about the value of the mixing parameter vector ξ_3. Furthermore, we assume that ξ_1 and ξ_2 are also independent. The Fisher information is then [6]:

$$F(\xi) \simeq F(\xi_1)F(\xi_2)F(\xi_3) \quad (15)$$

where $F(\xi_3)$ is the Fisher information with regards to the probability of the mixture. $F(\xi_1)$ and $F(\xi_2)$ are the Fisher information with regards to the vectors ξ_1 and ξ_2. In what follows we will compute each of these separately. For $F(\xi_3)$, 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, ..., M^{th} cluster. The number of trials of the j^{th} cluster is a multinomial distribution of parameters $p(1), p(2), \ldots, p(M)$. In this case, the determinant of the Fisher information matrix is [6]:

$$F(\xi_3) = \frac{N}{\prod_{j=1}^{M} p(j)} \quad (16)$$

For $F(\xi_1)$ and $F(\xi_3)$, we assume that the components of ξ_1 and ξ_2 are independent, then:

$$F(\xi_1) = \prod_{j=1}^{M} F(|\boldsymbol{\alpha}_j|) \quad (17)$$

$$F(\xi_2) = \prod_{j=1}^{M} F(\boldsymbol{\mu}_j) \quad (18)$$

let us consider the jth cluster $\mathcal{X}_j = (\boldsymbol{X}_l, \ldots, \boldsymbol{X}_{l+nj-1})$ of the mixture, where $l \leq N$, with parameters $|\boldsymbol{\alpha}_j|$ and $\boldsymbol{\mu}_j$. The choice of the jth cluster allows us to

simplify the notation without loss of generality. The Hessian matrix when we consider the vector $\boldsymbol{\mu}_j$ is given by:

$$H(\boldsymbol{\mu}_j) = \frac{\partial^2}{\partial \mu_{jk_1} \partial \mu_{jk_2}} (-logp(\mathcal{X}_j|\boldsymbol{\mu}_j)) \tag{19}$$

where $k_1 = 1 \ldots dim + 1$ and $k_2 = 1 \ldots dim + 1$. Straight forward manipulations give us the determinant of the matrix $H(\boldsymbol{\mu}_j)$:

$$F(\boldsymbol{\mu}_j) = n_j^{dim+1} |\boldsymbol{\alpha}_j|^{2(dim+1)} \prod_{k=1}^{dim+1} \Psi^{'}(\mu_{jk}|\boldsymbol{\alpha}_j|) \tag{20}$$

By substituting Eq. 20 in Eq. 18 we obtain:

$$F(\xi_2) = \prod_{j=1}^{M} \left(n_j^{dim+1} |\boldsymbol{\alpha}_j|^{2(dim+1)} \prod_{k=1}^{dim+1} \Psi^{'}(\mu_{jk}|\boldsymbol{\alpha}_j|) \right) \tag{21}$$

Now we determine the Fisher information when we consider $|\boldsymbol{\alpha}_j|$. The second derivative is given by:

$$-\frac{\partial^2 logp(\mathcal{X}_j||\boldsymbol{\alpha}_j|)}{\partial^2 |\boldsymbol{\alpha}_j|} = n_j \left(-\Psi^{'}(|\boldsymbol{\alpha}_j|) + \sum_{k=1}^{dim+1} \mu_{jk}^2 \Psi^{'}(\mu_{jk}|\boldsymbol{\alpha}_j|) \right) \tag{22}$$

and represent the Fisher information. By substituting Eq. 22 in Eq. 17, we obtain:

$$F(\xi_1) = \prod_{j=1}^{M} n_j \left(-\Psi^{'}(|\boldsymbol{\alpha}_j|) + \sum_{k=1}^{dim+1} \mu_{jk}^2 \Psi^{'}(\mu_{jk}|\boldsymbol{\alpha}_j|) \right) \tag{23}$$

Finally the complete Fisher information for the mixture is found by substituting Eq. 16, Eq. 21 and Eq. 23 in Eq. 15.

3.3 Prior Distribution $h(\xi)$

The performance of the MML criterion is dependent on the choice of the prior distribution $h(\xi)$. Several criteria have been proposed for the selection of prior $h(\xi)$. Following Bayesian inference theory, the prior density of a parameter is either constant on the whole range of its values or the value range is split into cells and the prior density is assumed to be constant within each cell. Since ξ_1, ξ_2 and ξ_3 are independent, we have:

$$h(\xi) = h(\xi_1)h(\xi_2)h(\xi_3) \tag{24}$$

We will now define the three densities $h(\xi_1)$, $h(\xi_2)$, and $h(\xi_3)$. The vector ξ_3 has M dependent components; i.e. the sum of the mixing parameters is one. Thus, we omit one of these components, say $p(M)$. The new vector has $(M - 1)$ independent components. We treat the $p(j)$, $j = 1 \ldots M - 1$ as being the

parameters of a multinomial distribution. With the $(M-1)$ remaining mixing parameters, $(M-1)!$ possible vectors can be formed. Thus, we set the uniform prior density of ξ_3 to [6]:

$$h(\xi_3) = \frac{1}{(M-1)!} \tag{25}$$

For $h(\xi_2)$, since $\boldsymbol{\mu}_j$, $j = 1 \ldots M$ are assumed to be independent:

$$h(\xi_2) = \prod_{j=1}^{M} h(\boldsymbol{\mu}_j) \tag{26}$$

Using the same approach as for the vector ξ_3, we set the uniform prior density of $\boldsymbol{\mu}_j$ to:

$$h(\boldsymbol{\mu}_j) = \frac{1}{dim!} \tag{27}$$

Indeed, $\sum_{k=1}^{dim+1} \mu_{jk} = 1$. By substituting Eq. 27 in Eq. 26, we obtain:

$$h(\xi_2) = \frac{1}{dim!^M} \tag{28}$$

For $h(\xi_1)$, since $|\boldsymbol{\alpha}_j|$, $j = 1 \ldots M$ are assumed to be independent:

$$h(\xi_1) = \prod_{j=1}^{M} h(|\boldsymbol{\alpha}_j|) \tag{29}$$

We will now calculate $h(|\boldsymbol{\alpha}_j|)$. In the absence of other knowledge about the $|\boldsymbol{\alpha}_j|$, we use the principle of ignorance by assuming that $h(|\boldsymbol{\alpha}_j|)$ is locally uniform over the ranges $[0, e^3|\hat{\boldsymbol{\alpha}}_{pop}|]$ (in fact, we know experimentally that $|\boldsymbol{\alpha}_j| < e^3|\hat{\boldsymbol{\alpha}}_{pop}|$, where $|\hat{\boldsymbol{\alpha}}_{pop}|$ is the estimated parameter when we consider the entire population. We choose the following uniform priors in accordance with Ockham's razor (a simple priors which give good results) [7]:

$$h(|\boldsymbol{\alpha}_j|) = \frac{e^{-3}}{|\hat{\boldsymbol{\alpha}}_{pop}|} \tag{30}$$

By substituting Eq. 30 in Eq. 29, we obtain

$$h(\xi_1) = \prod_{j=1}^{M} \frac{e^{-3}}{|\hat{\boldsymbol{\alpha}}_{pop}|} = \frac{e^{-3M}}{|\hat{\boldsymbol{\alpha}}_{pop}|^M} \tag{31}$$

By substituting Eq. 31, Eq. 28 and Eq. 25 in Eq. 24, we obtain:

$$h(\xi) = \frac{e^{-3M}}{|\hat{\boldsymbol{\alpha}}_{pop}|^M (M-1)! dim!^M} \tag{32}$$

The expression of MML for a finite mixture of Dirichlet distributions is obtained by substituting Eq. 32 and Eq. 15 in Eq. 14.

3.4 Estimation and Selection Algorithm

The algorithm of selection and estimation is thus as follows:

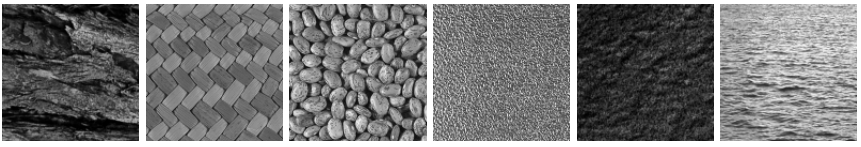
Algorithm
For each candidate value of M:

1. Initialization
2. E-Step: Compute the *posterior* probabilities:
   $\hat{Z}_{ij} = \frac{p(\boldsymbol{X_i}|\boldsymbol{\alpha}_j)p(j)}{\sum_{j=1}^{M} p(\boldsymbol{X_i}|,\boldsymbol{\alpha}_j)p(j)}$
3. CM-Steps:
   (a) **CM-Step1:** Calculate $\xi_1^{(t)}$ using Eq. 10 with ξ_2 fixed at $\xi_2^{(t-1)}$ and ξ_3 fixed at $\xi_3^{(t-1)}$.
   (b) **CM-Step2:** Calculate $\xi_2^{(t)}$ using Eq. 12 with ξ_1 fixed at $\xi_1^{(t)}$ and ξ_3 fixed at $\xi_3^{(t-1)}$.
   (c) **CM-Step2:** Calculate $\xi_3^{(t)}$ using Eq. 7 with ξ_1 fixed at $\xi_1^{(t)}$ and ξ_2 fixed at $\xi_2^{(t)}$.
4. If the convergence test is passed, terminate, else go to 2.
5. Calculate the associated criterion MML(M) using Eq. 14.
6. Select the optimal model M^* such that: $M^* = \arg\min_M MML(M)$

details about the initialization algorithm can be found in [1]. The convergence test can involve the stabilization of the parameters or the likelihood function.

4 Experimental Results

The application concerns the summarization of 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 developed to handle pictorial queries. 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. Summarization is also very efficient for browsing. 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, it allows us to partition the feature space 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, we used the *Vistex* gray-level texture database obtained from the 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",



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

Fig. 1. Sample images from each group. (a) Bark, (b) Fabric, (c) Food, (d) Metal, (e) Sand, (f) Water.

Table 1. Number of clusters found by three criteria (MML, MDL and AIC)

Number of clusters	MML	MDL	AIC
1	-12945.10	-12951.40	-12974.90
2	-12951.12	-13001.52	-13019.12
3	-12960.34	-13080.37	-13094.23
4	-13000.76	-13206.73	-13225.57
5	-13245.18	**-13574.98**	**-13591.04**
6	**-13765.04**	-13570.09	-13587.64
7	-13456.71	-13493.50	-13519.50
8	-13398.16	-13387.56	-13405.92
9	-13402.64	-13125.41	-13141.95
10	-13100.82	-13001.80	-13020.23

Table 2. Confusion matrix for image classification by a Dirichlet mixture

	Bark	Fabric	Food	Metal	Sand	Water
Bark	250	0	0	0	6	0
Fabric	0	248	8	0	0	0
Food	0	9	375	0	0	0
Metal	0	0	0	250	0	6
Sand	4	0	0	0	380	0
Water	3	0	0	7	2	372

"Metal", "Water" and "Sand" were used to create a new database. A database with 1920 images of size 64 × 64 pixels was obtained. Four images from each of the Bark, Fabric and Metal texture groups and 6 images from Water, Food and Sand were used. Examples of images from each of the categories are shown in Fig. 1. In order to determine the vector of characteristics for each image, we used the cooccurrence matrix introduced by Haralick et al. [8]. For relevant representation of texture, many cooccurrences should be computed, each one considering a given neighborhood and direction. In our application, we have considered the following four neighborhoods: $(1; 0)$, $(1; \frac{\pi}{4})$, $(1; \frac{\pi}{2})$, and $(1; \frac{3\pi}{4})$. For each of these neighborhoods, we calculated the corresponding cooccurrence matrix, then derived from it the following features: Mean, Energy, Contrast, and Homogeneity [9]. Thus, each image was represented by a $16D$ feature vector. By

Table 3. 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

applying our algorithm to the texture database using MML and other different selection selection criteria such that MDL and AIC [2], only the MML criterion found six categories (see Table 1). In what follows we use the selection found by the MML. The classification was performed using the Bayesian decision rule after the class-conditional densities were estimated. The confusion matrix for the texture image classification is given in Table 2. In this confusion matrix, the cell $(classi, classj)$ represents the number of images from $classi$ which are classified as $classj$. The number of images misclassified was small: 45 in all, which represents an accuracy of 97.65 percent. From Table 2, we can see clearly that the errors are due essentially to the presence of macrotexture, i.e., the texture at large scale, (between Fabric and Food for example) or because of microtexture, i.e., the texture at pixel level (between Metal and Water for example). Table 3 shows the confusion matrix for the Gaussian mixture.

5 Conclusion

In this paper, we have proposed a new method based on the ECM algorithm to estimate the parameters of a Dirichlet mixture. The ECM algorithm replaces a complicated M-step of the EM algorithm with several computationally simpler CM-Steps. The number of clusters is determined using an MML-based approach. From the experimental results, we can say that the Dirichlet distribution offers strong modeling capabilities.

Acknowledgement

The completion of this research was made possible thanks to the the Natural Sciences and Engineering Research Council of Canada, Heritage Canada and Bell Canada's support through its Bell University Laboratories R&D program.

References

1. N. Bouguila, D. Ziou and J. Vaillancourt. Unsupervised Learning of a Finite Mixture Model Based on the Dirichlet Distribution and its Application. *IEEE Transactions on Image Processing*, 13(11):1533–1543, November 2004.
2. G.J. McLachlan and D. Peel. *Finite Mixture Models.* New York: Wiley, 2000.

3. X. L. Meng and D. B. Rubin. Maximum Likelihood Estimation via the ECM Algorithm: a General Framework. *Biometrika*, 80(2):267–278, 1993.
4. C.E. Shannon. A Mathematical Theory of Communication. *Bell System Tech.*, 27:379–423, 1948.
5. C. S. Wallace and D. L. Dowe. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing* , 10(1):73–83, 2000.
6. R. A. Baxter and J. J. Olivier. Finding Overlapping Components with MML. *Statistics and Computing*, 10(1):5–16, 2000.
7. W. Jeffreys and J. Berger. Ockham's Razor and Bayesian Analysis. *American Scientist*, 80:64–72, 1992.
8. R. M. Haralick, K. Shanmugan and I. Dinstein. Texture features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 8:610–621, 1973.
9. M. Unser. Sum and Difference Histograms for Texture Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(1):118–125, 1986.