MML-Based Approach for Finite Dirichlet Mixture Estimation and Selection

Nizar Bouguila and Djemel Ziou

Département d'Informatique, Faculté des Sciences
Université de Sherbrooke
Sherbrooke, Qc, Canada J1K 2R1.
{nizar.bouguila, djemel.ziou}@usherbrooke.ca

Abstract. This paper proposes an unsupervised algorithm for learning a finite Dirichlet mixture model. An important part of the unsupervised learning problem is determining the number of clusters which best describe the data. We consider here the application of the Minimum Message length (MML) principle to determine the number of clusters. The Model is compared with results obtained by other selection criteria (AIC, MDL, MMDL, PC and a Bayesian method). The proposed method is validated by synthetic data and summarization of texture image database.

1 Introduction

Statistical models are widely used in various fields such as image processing, pattern recognition, machine learning and remote sensing [1]. In these models, data is characterized in terms of its likely behavior, by means of a probability. The performance of the resulting algorithms depends heavily on the accuracy of the probabilistic models employed. Among the probability models, finite mixtures of densities are widely used [2]. Finite mixtures of distributions are a flexible and powerful modeling which has provided a mathematical based approach to the statistical modeling of a wide variety of random phenomena. This makes them an excellent choice in Bayesian learning. In statistical pattern recognition, finite mixtures permit a formal approach to unsupervised learning. The adoption of this model-based approach to clustering brings important advantages: for instance, the selection of the number of clusters or the assessment of the validity of 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 which best describes the data. For this purpose, many approaches have been suggested, such as the Minimum Message Length (MML) [3], Akaike's Information Criterion (AIC) [4], the Minimum Description Length (MDL) [5], the MMDL [6] and the partition coefficient (PC) [7]. Besides, many Bayesian model selection approaches was proposed such as the model of Roberts et al. [8]. In this paper, we consider MML and Dirichlet mixtures. MML has been used especially in the case of Gaussian, Poisson, Von Miss circular mixtures [9] and recently in the case of Gamma [10] mixtures. However, we have proven in a

previous work that the Dirichlet may provide a better fit [11] [12]. From an information-theory point of view, the minimum message length 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 [13].
Let us consider a set of data $\mathcal{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, ...\boldsymbol{X}_N)$ controlled by a mixture of distributions with parameters $\Theta = (\theta_1, \theta_2, ..., \theta_M)$, where M is the number of clusters, and θ_j is a vector which contains the parameters of the j^{th} distribution. According to information theory, the optimal number of clusters of the mixture is that which allows a minimum amount of information, measured in nats, needed to transmit $\mathcal{X}$ efficiently from a sender to a receiver. The message length is defined as :$MessLen = -log(P(\Theta|\mathcal{X}))$. The minimum message length principle has strong connections with Bayesian inference, and hence uses an explicit prior distribution over parameter values [9]. Baxter [9] gives us the formula for the message length for a mixture of distributions:

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

where $h(\Theta)$ is the prior probability, $p(\mathcal{X}|\Theta)$ is the likelihood, and $|F(\theta)|$ 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. The estimation of the number of clusters is carried out by finding the minimum with regards to Θ of the message length $MessLen$. In dimension dim, the Dirichlet pdf is defined by:

$$p(\boldsymbol{X}|\boldsymbol{\alpha}) = \frac{\Gamma(|\boldsymbol{\alpha}|)}{\prod_{i=1}^{dim+1}\Gamma(\alpha_i)} \prod_{i=1}^{dim+1} X_i^{\alpha_i - 1} \quad (2)$$

where $\sum_{i=1}^{dim} X_i < 1$, $|\boldsymbol{X}| = \sum_{i=1}^{dim} X_i$, $0 < X_i < 1 \quad \forall i = 1 \dots dim$, $X_{dim+1} = 1 - |\boldsymbol{X}|$, $|\boldsymbol{\alpha}| = \sum_{i=1}^{dim+1} \alpha_i$, $\alpha_i > 0 \quad \forall i = 1 \dots dim + 1$. This distribution is the multivariate extension of the 2-parameter Beta distribution. A Dirichlet mixture with M components is defined as :

$$p(\boldsymbol{X}|\Theta) = \sum_{j=1}^{M} p(\boldsymbol{X}|\boldsymbol{\alpha}_j)p(j) \quad (3)$$

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 $\Theta = (\alpha, \boldsymbol{P})$, where $\alpha = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_M)^T$ and $\boldsymbol{P} = (p(1), \cdots, p(M))^T$ is the mixing parameters vector. In the next two sections, we will calculate the Fisher information $F(\Theta)$ and the prior probability density function $h(\Theta)$. Section 4 is devoted to the experimental results.

2 Fisher Information for a Mixture of Dirichlet

Fisher information is the determinant of the Hessian matrix of the logarithm of minus the likelihood of the mixture. In our case, we have a $((M \times (dim + 2)) \times (M \times (dim + 2)))$ Hessian matrix defined by:

$$H_{l_1 l_2} = \frac{\partial^2}{\partial\theta_{l_1}\theta_{l_2}}(-logp(\mathcal{X}|\Theta)) \tag{4}$$

where $l_1 = 1 \ldots M \times (dim+2)$ and $l_2 = 1 \ldots M \times (dim+2)$. 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 α and the vector $\boldsymbol{P}$ are independent because any prior idea one might have about α would usually not be greatly influenced by one's idea about the value of the mixing parameters vector $\boldsymbol{P}$. Furthermore, we assume that the components of α are also independent. The Fisher information is then:

$$F(\theta) \simeq F(\boldsymbol{P}) \prod_{j=1}^{M} F(\boldsymbol{\alpha}_j) \tag{5}$$

where $F(\boldsymbol{P})$ is the Fisher information with regards to the mixing parameters of the mixture and $F(\boldsymbol{\alpha}_j)$ the Fisher information with regards to the vector $\boldsymbol{\alpha}_j$ of a single Dirichlet distribution. In what follows we will compute each of these separately. For $F(\boldsymbol{P})$, 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, ..., M^{th} cluster. The number of trials of the j^{th} cluster is a multinomial distribution of parameters $p(1), p(2), ..., p(M)$. In this case, the determinant of the Fisher information matrix is:

$$F(P) = \frac{N}{\prod_{j=1}^{M} p(j)} \tag{6}$$

where N is the number of data elements. For $F(\boldsymbol{\alpha}_j)$, let us consider the jth cluster $\mathcal{X}_j = (\boldsymbol{X}_l, \ldots, \boldsymbol{X}_{l+n_j-1})$ of the mixture, where $l \leq N$, with parameter $\boldsymbol{\alpha}_j$. The choice of the jth cluster allows us to simplify the notation without loss of generality. The Hessian matrix when we consider the vector $\boldsymbol{\alpha}_j$ is given by:

$$H(\boldsymbol{\alpha}_j)_{k_1 k_2} = \frac{\partial^2}{\partial\alpha_{jk_1}\partial\alpha_{jk_2}}(-logp(\mathcal{X}_j|\boldsymbol{\alpha}_j)) \tag{7}$$

where $k_1 = 1 \ldots dim + 1$ and $k_2 = 1 \ldots dim + 1$. We can write the negative of the log-likelihood function as follows:

$$-logp(\mathcal{X}_j|\boldsymbol{\alpha}_j) = -log(\prod_{i=l}^{l+n_j-1} p(\boldsymbol{X}_i|\boldsymbol{\alpha}_j)) = -\sum_{i=l}^{l+n_j-1} logp(\boldsymbol{X}_i|\boldsymbol{\alpha}_j) \tag{8}$$

We have:

$$-\frac{\partial logp(\mathcal{X}_j|\boldsymbol{\alpha}_j)}{\partial \alpha_{jk}} = n_j(-\Psi(|\boldsymbol{\alpha}_j|) + \Psi(\alpha_{jk})) - \sum_{i=l}^{l+n_j-1} log(X_{ik}) \tag{9}$$

Where Ψ is the digamma function. Then,

$$-\frac{\partial^2 logp(\mathcal{X}_j|\boldsymbol{\alpha}_j)}{\partial \alpha_{jk_1}\partial \alpha_{jk_2}} = -n_j\Psi'(|\boldsymbol{\alpha}_j|) \tag{10}$$

$$-\frac{\partial^2 logp(\mathcal{X}_j|\boldsymbol{\alpha}_j)}{\partial^2 \alpha_{jk}} = -n_j(\Psi'(|\boldsymbol{\alpha}_j|) - \Psi'(\alpha_{jk})) \tag{11}$$

Where Ψ' is the trigamma function. We remark that $H(\boldsymbol{\alpha}_j)_{k_1k_2}$ can be written as:

$$H(\boldsymbol{\alpha}_j)_{k_1k_2} = D + \gamma \boldsymbol{a}\boldsymbol{a}^T \tag{12}$$

where $D = diag[n_j\Psi'(\alpha_{j1}), \ldots, n_j\Psi'(\alpha_{jdim+1})]$, $\gamma = -n_j\Psi'(|\boldsymbol{\alpha}_j|)$, $\boldsymbol{a}^T = \mathbf{1}$ and $\gamma \neq (\sum_{k=1}^{dim+1} \frac{a_k^2}{D_{kk}})^{-1}$, then by the theorem (Theorem 8.4.3) given by Graybill [14], the determinant of the matrix $H(\boldsymbol{\alpha}_j)_{k_1k_2}$ is given by:

$$F(\boldsymbol{\alpha}_j) = (1 + \gamma \sum_{k=1}^{dim+1} \frac{a_k^2}{D_{kk}}) \prod_{k=1}^{dim+1} D_{kk} \tag{13}$$

then

$$F(\boldsymbol{\alpha}_j) = (1 - \Psi'(|\boldsymbol{\alpha}_j|) \sum_{k=1}^{dim+1} \frac{1}{\Psi'(\alpha_{jk})}) n_j^{dim+1} \prod_{k=1}^{dim+1} \Psi'(\alpha_{jk}) \tag{14}$$

Once we have the Fisher information for a single Dirichlet distribution, we can use it to calculate the Fisher information for a mixture of Dirichlet distributions. Eq. 5 is rewritten as:

$$F(\Theta) \simeq \frac{N}{\prod_{j=1}^{M} p(j)} \prod_{j=1}^{M} (1 - \Psi'(|\boldsymbol{\alpha}_j|) \sum_{k=1}^{dim+1} \frac{1}{\Psi'(\alpha_{jk})}) n_j^{dim+1} \prod_{k=1}^{dim+1} \Psi'(\alpha_{jk}) \tag{15}$$

3 Prior Distribution $h(\Theta)$

The performance of the MML criterion is dependent on the choice of the prior distribution $h(\Theta)$. Several criteria have been proposed for the selection of prior $h(\Theta)$. 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 inside each cell. Since α and the vector $\boldsymbol{P}$ are independent, we have:

$$h(\Theta) = h(\alpha)h(\boldsymbol{P}) \tag{16}$$

We will now define the two densities $h(\alpha)$ and $h(\boldsymbol{P})$. The $\boldsymbol{P}$ vector 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 $\boldsymbol{P}$ to [15]:

$$h(\boldsymbol{P}) = \frac{1}{(M-1)!} \tag{17}$$

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

$$h(\alpha) = \prod_{j=1}^{M} h(\boldsymbol{\alpha}_j) \tag{18}$$

We will now calculate $h(\boldsymbol{\alpha}_j)$. In fact, we assume that the components of $\boldsymbol{\alpha}_j$ are independent and in the absence of other knowledge about the $\alpha_{jk}, k = 1, \ldots, dim+1$, we use the principle of ignorance by assuming that $h(\alpha_{jk})$ is locally uniform over the range $[0, e^6 \frac{|\hat{\boldsymbol{\alpha}}_j|}{\hat{\alpha}_{jk}}]$ (in fact, we know experimentally that $\alpha_{jk} < e^6 \frac{|\hat{\boldsymbol{\alpha}}_j|}{\hat{\alpha}_{jk}}$), where $\hat{\boldsymbol{\alpha}}_j$ is the estimated vector. We choose the following uniform prior in accordance with Ockham's razor (a simple priors which give good results):

$$h(\alpha_{jk}) = \frac{e^{-6}\hat{\alpha}_{jk}}{|\hat{\boldsymbol{\alpha}}_j|} \tag{19}$$

By substituting Eq. 19 in Eq. 18, we obtain:

$$h(\boldsymbol{\alpha}_j) = \frac{e^{-6(dim+1)}}{|\hat{\boldsymbol{\alpha}}_j|^{dim+1}} \prod_{k=1}^{dim+1} \hat{\alpha}_{jk} \tag{20}$$

and

$$h(\alpha) = \prod_{j=1}^{M} h(\boldsymbol{\alpha}_j) = e^{-6M(dim+1)} \prod_{j=1}^{M} \frac{\prod_{k=1}^{dim+1} \hat{\alpha}_{jk}}{|\hat{\boldsymbol{\alpha}}_j|^{dim+1}} \tag{21}$$

So, substituting Eq. 21 and Eq. 17 in Eq. 16, we obtain:

$$log(h(\Theta)) = -\sum_{j=1}^{M-1} log(j) - 6M(dim+1) - (dim+1)\sum_{j=1}^{M} log(|\hat{\boldsymbol{\alpha}}_j|) + \sum_{j=1}^{M}\sum_{k=1}^{dim+1} log(\hat{\alpha}_{jk}) \tag{22}$$

The expression of MML for a finite mixture of Dirichlet distributions is obtained by substituting equations (22) and (15) in equation (1). The complete algorithm of estimation and selection is then as follows:

Algorithm

For each candidate value of M:

1. Estimate the parameters of the Dirichlet mixture using the algorithm in [11] [12].

2. Calculate the associated criterion MML(M) using Eq. 1.
3. Select the optimal model M^* such that:

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

4 Experimental Results

We compare the results from the MML approach with those obtained using the same model parameters (from the EM algorithm) using other model-order selection criteria/techniques. The methods we compare are the minimum description length (MDL) [5], The MMDL (Mixture MDL)[6], the Akaikes's information criterion (AIC) [4], the Partition coefficient (PC) [7] and a Bayesian criterion, which we call B, proposed by Roberts et al. [8].

4.1 Synthetic data

In the first application we investigate the properties of our model selection on three two-dimensional toy problems. We choose $dim = 2$ purely for ease of representation. In the first example, data were generated from five Dirichlet densities with different parameters. The parameters were: $\alpha_{11} = 10$, $\alpha_{12} = 16$, $\alpha_{13} = 40$, $\alpha_{21} = 23$, $\alpha_{22} = 50$, $\alpha_{23} = 32$, $\alpha_{31} = 15$, $\alpha_{32} = 19$, $\alpha_{33} = 6$, $\alpha_{41} = 29$, $\alpha_{42} = 8$, $\alpha_{43} = 55$, $\alpha_{51} = 60$, $\alpha_{52} = 40$, $\alpha_{53} = 16$. A total of 100 samples for each of densities were taken. The resultant mixture is presented in Fig. 1.a. From table 1, we can see that only the MML found the exact number of clusters. In the

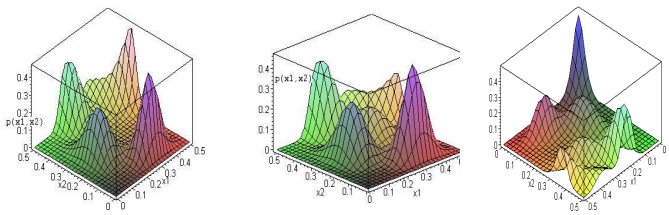
Table 1. values for the six criteria for the first two-dimensional generated data set.

Number of clusters	MML	MDL	AIC	PC	MMDL	B
1	-207.26	-206.16	-401.15	N/A	-206.16	270.41
2	-208.12	-207.02	-401.87	0.63	-207.93	274.45
3	-209.43	-207.89	-401.90	**0.76**	-209.45	278.84
4	-209.61	**-208.00**	**-403.44**	0.75	-210.40	**280.13**
5	**-210.36**	-207.54	-401.12	0.70	-210.33	272.02
6	-208.61	-207.01	-400.67	0.67	**-211.79**	272.98
7	-207.36	-204.43	-399.82	0.65	-209.59	273.17
8	-206.16	-200.12	-398.34	0.66	-207.33	273.91

second example, data were generated from six Dirichlet densities with different parameters. The parameters were: $\alpha_{11} = 10$, $\alpha_{12} = 16$, $\alpha_{13} = 40$, $\alpha_{21} = 23$, $\alpha_{22} = 50$, $\alpha_{23} = 32$, $\alpha_{31} = 15$, $\alpha_{32} = 19$, $\alpha_{33} = 6$, $\alpha_{41} = 29$, $\alpha_{42} = 8$, $\alpha_{43} = 55$ $\alpha_{51} = 60$, $\alpha_{52} = 40$, $\alpha_{53} = 16$, $\alpha_{61} = 30$, $\alpha_{62} = 30$, $\alpha_{63} = 30$. A total of 100 samples for each of the fourth first densities and a total of 50 for each of the two last densities were taken. The resultant mixture is presented in Fig. 1.b. From table 2, we can see that only the MML found the exact number of clusters.

Table 2. values for the six criteria for the second two-dimensional generated data set.

Number of clusters	MML	MDL	AIC	PC	MMDL	B
1	-287.65	-276.16	-476.52	N/A	-276.16	320.73
2	-288.23	-277.09	-477.09	0.71	-278.31	318.77
3	-288.93	-277.65	-477.54	0.76	-279.20	320.51
4	-289.33	**-278.92**	-477.78	**0.77**	-281.32	320.13
5	-289.79	-278.80	**-478.33**	0.72	**-282.29**	**320.84**
6	**-290.12**	-276.85	-476.97	0.70	-281.65	319.05
7	-287.54	-274.66	-476.80	0.69	-280.11	319.86
8	-287.11	-272.82	-476.66	0.68	-297.80	320.06



(a) (b) (c)

Fig. 1. Mixture densities for the generated data sets

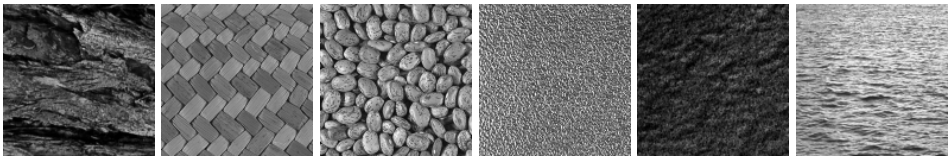
In the last example, data were generated from seven densities. The parameters were: $\alpha_{11} = 10$, $\alpha_{12} = 14$, $\alpha_{13} = 40$, $\alpha_{21} = 23$, $\alpha_{22} = 50$, $\alpha_{23} = 32$, $\alpha_{31} = 15$, $\alpha_{32} = 19$, $\alpha_{33} = 6$, $\alpha_{41} = 29$, $\alpha_{42} = 8$, $\alpha_{43} = 55$, $\alpha_{51} = 60$, $\alpha_{52} = 40$, $\alpha_{53} = 16$, $\alpha_{61} = 30$, $\alpha_{62} = 30$, $\alpha_{63} = 30$, $\alpha_{71} = 10$, $\alpha_{72} = 10$, $\alpha_{73} = 40$. A total of 100 samples for each of the three first densities and a total of 50 samples for each of the four last densities were taken. The resultant mixture is presented in Fig. 1.c. From table 3, we can see that only the MML found the exact number of clusters.

Table 3. values for the six criteria for the third two-dimensional generated data set.

Number of clusters	MML	MDL	AIC	PC	MMDL	B
1	-310.18	-300.54	-512.02	N/A	-300.54	378.22
2	-310.87	-300.89	-512.16	0.66	-301.49	380.14
3	-311.22	-301.15	-512.43	0.67	-302.71	379.64
4	-311.93	-301.87	-512.76	0.69	-304.27	379.06
5	-312.37	-302.12	**-513.86**	**0.76**	-305.62	378.83
6	-313.37	**-303.76**	-513.64	0.71	**-308.94**	**380.53**
7	**-313.55**	-301.09	-513.66	0.72	-308.18	379.03
8	-313.49	-300.87	-513.05	0.67	-308.09	379.76

4.2 Real data

The second 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* grey 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", "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. 2. In order to determine the vector of characteristics for each image, we used



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

Fig. 2. Sample images from each group. (a) Bark, (b) Fabric, (c) Food, (d) Metal, (e) Sand, (f) Water.

the cooccurrence matrix introduced by Haralick et al. [16]. For relevant representation of texture, many cooccurrences should be computed, each one considering a given neighborhood and direction. In our application, we have considered considering 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 calculate the corresponding cooccurrence matrix, then derive from it the following features: Mean, Energy, Contrast, and Homogeneity. Thus, each image was represented by an $16D$ feature vector. By

applying our algorithm to the texture database, only the MML criterion found six categories (see table 4). Then, 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

Table 4. Number of clusters found by the six criteria..

Number of clusters	MML	MDL	AIC	PC	MMDL	B
1	-12945.1	-12951.4	-25643.9	N/A	-12951.4	12543.11
2	-12951.12	-13001.52	-25780.12	0.72	-13002.17	12897.21
3	-12960.34	-13080.37	-25930.23	0.73	-13381.82	12799.54
4	-13000.76	-13206.73	-26000.57	**0.82**	-13209.81	12730.13
5	-13245.18	**-13574.98**	-26111.04	0.78	**-13578.60**	**13003.2**
6	**-13765.04**	-13570.09	-26312.64	0.77	-13576.34	13000.11
7	-13456.71	-13493.5	**-26401.50**	0.74	-13499.53	12761.23
8	-13398.16	-13387.56	-26207.92	0.69	-13393.69	12900.19
9	-13402.64	-13125.41	-26009.95	0.71	-13132.34	12980.32
10	-13100.82	-13001.8	-25999.23	0.80	-13007.81	12580.32

the texture image classification is given in Table 5. 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 5, 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 5. 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

5 Conclusion

We have presented a MML-based criterion to select the number of components in Dirichlet mixtures. The results presented indicate clearly that the MML model

selection method which is based upon information theory outperforms the other methods. The validation was based on synthetic data and an interesting applications which involves texture image database summarization.

Aknowledegment

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. A. K. Jain, R. P. W. Duin and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, January 2000.
2. G.J. McLachlan and D. Peel. *Finite Mixture Models*. New York: Wiley, 2000.
3. C.S Wallace and D.M. Boulton. An Information Measure for Classification. *Computer Journal*, 11(2):195–209, 1968.
4. H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control*, AC-19(6):716–723, 1974.
5. J. Rissanen. Modeling by Shortest Data Description. *Automatica*, 14:465–471, 1987.
6. M. A. T. Frigueiredo, J. M. N. Leitao and A. K. Jain. On Fitting Mixture Models. In E. Hancock and M. Pellilo, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 54–69, 1999.
7. J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.
8. S.J. Roberts, D. Husmeier, I. Rezek and W. Penny. Bayesian Approaches to Gaussian Mixture Modeling. *IEEE Transactions on PAMI*, 20(11):1133–1142, November 1998.
9. R.A Baxter. *Minimum Message Length Inference: Theory and Applications*. Ph.D. Thesis, Monash University, Clayton, Victoria, Australia, 1996.
10. D. Ziou and N. Bouguila. Unsupervised Learning of a Gamma Finite Mixture Using MML: Application to SAR Image Analysis . In *17th International Conference on Pattern Recognition, ICPR2004*, pages 280–283, 2004.
11. 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.
12. N. Bouguila, D. Ziou, and J. Vaillancourt. Novel Mixtures Based on the Dirichlet Distribution: Application to Data and Image Classification. In Petra Perner and Azriel Rosenfeld, editors, *Machine Learning and Data Mining in Pattern Recognition*, pages 172–181, 2003.
13. D.L. Dowe and G. Farr. An Introduction to MML Inference. Technical report, Department of Computer Science, Monash University, 1997.
14. F. A. Graybill. *Matrices with applications in Statistics*. Wadsworth, California, 1983.
15. R. A. Baxter and J. J. Olivier. Finding Overlapping Components with MML. *Statistics and Computing*, 10():5–16, 2000.
16. R. M. Haralick, K. Shanmugan and I. Dinstein. Texture features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 8:610–621, 1973.