A Model Based Mixture Supervised Classification Approach in Hyperspectral Data Analysis

M. Murat Dundar and David Landgrebe, *Life Fellow, IEEE*
School of Electrical and Computer Engineering Purdue University

A Model Based Mixture Supervised Classification Approach in Hyperspectral Data Analysis

M. Murat Dundar and David Landgrebe, *Life Fellow, IEEE*
School of Electrical and Computer Engineering Purdue University

***Abstract*—It is well-known that there is a strong relation between class definition precision and classification accuracy in pattern classification applications. In hyperspectral data analysis, usually classes of interest contain one or more components and may not be well represented by a single Gaussian density function. In this paper, a model based mixture classifier, which uses mixture models to characterize class densities, is discussed. However, a key outstanding problem of this approach is how to choose the number of components and determine their parameters for such models in practice, and to do so in the face of limited training sets where estimation error becomes a significant factor. The proposed classifier estimates the number of subclasses and class statistics simultaneously by choosing the best model. The structure of class covariances is also addressed through a model-based covariance estimation technique introduced in this paper.**

Index Terms–**Gaussian mixtures, Expectation-maximization, Covariance estimator**

I. INTRODUCTION

In hyperspectral analysis, materials of practical interest, such as agricultural crops, forest plantations, natural vegetation, minerals, and fields of interest in urban areas exist in a number of states and are observed in a number of conditions of illumination. It is thus necessary to characterize them not with a single average or typical spectral response, but with a family of responses.

One of the most powerful and versatile means for representing such a family of responses quantitatively is to model each as a multivariate probability density function, as this makes possible classification by assigning the class to a sample based on likelihood. Maximum likelihood methods have long been used in the field of digital communication systems and have made possible communication in very noisy and complex environments. To use these methods for multispectral data classification, one must determine the probabilistic density function that correctly models each class of a data set. Most commonly this is done by estimating at least the first two orders of statistics, the mean vector and the covariance matrix, for each class.

Quantifying higher order statistics of each class rather than just mean and covariance to better fit the data into a parametric class density function might seem desirable at first sight. Indeed, it is well established that, in theory a complete description of an arbitrary distribution can be made by the use of statistics of all orders, as in an infinite series. The reason it is customary to use no more than two orders of statistics, the mean vector and the covariance matrix, arises from the practical problem of estimating these two statistics from the available set of labeled data. If one were to estimate higher order statistics one would definitely need more labeled samples to arrive at an adequately precise estimate. Usually the number of these samples is very limited since the labeling of such samples is one of the most onerous and time-consuming aspects of designing a classifier. Indeed, there often is not much information available about the scene to use in labeling the samples. Thus it is not practical to use statistics beyond the second order [1].

Another solution might be to use nonparametric class density estimations. A nonparametric classifier does not rely on any assumption about the structure of the underlying density function. The classifier may become the *Bayes* classifier and the resulting error, *Bayes* error, the smallest achievable error, if the density estimates converge to the true densities, which is only achieved when an infinite number of labeled samples are used. When the densities are estimated non-parametrically with limited number of samples, the estimate is far less reliable with larger bias and variance than the parametric counterpart [2].

The normal mixture model, which is the sum of one or more weighted Gaussian components, combine much of the flexibility of nonparametric methods with certain of the analytic advantages of parametric methods. Under fairly weak conditions and given enough components a mixture model can approximate a given density arbitrarily closely allowing great flexibility.

Although, mixture models are widely used in applied statistics and recently in knowledge discovery and data mining, they have not attracted much attention in the remote sensing community. In [3], Hoffbeck and Landgrebe have introduced a method for using mixture models to characterize the class densities. Their approach is to partition the data into a designated number of clusters, and compare the mixture models corresponding to each of these partitions by means of a selection criterion. They used a nearest means clustering algorithm to partition the data and leave-one-out likelihood criterion (LOOL) to select the best fit. This method has two shortcomings. First, statistics estimation, which is done through maximum likelihood, does not address the ill-conditioned cluster covariance case. This is almost unavoidable when clustering the

data of limited size to several subclusters in hyperdimensional space. Second, nearest mean clustering alone does not provide an efficient model fit.

In this work, we will address these issues by proposing a method based on model-based expectation-maximization (EM). When EM is used, data is better fit into the model and hence more reliable measures of fit are obtained. EM requires an initialization, which is done by initial partitioning of the data through a nearest mean clustering algorithm. The model based approach, which utilizes the common and sample covariance matrices as well as their trace and diagonal forms not only helps us to identify the underlying class structure providing better characterization but also produces robust estimates of component covariance matrices when the number of labeled samples in each cluster is less than the dimensionality.

The measures of fit are computed through an approximate Bayes factor, which is known as the Bayesian Information Criterion (BIC) [8].

The design of the proposed mixture classifier involves three core stages. These are:

1) Initialization and clustering
2) Expectation-maximization
3) Model selection

This paper is organized as follows. In the first section, we give the necessary background in expectation-maximization (EM) for mixture models. In the second section, a model based approach for identifying mixtures is introduced. In the third section experiments with real aerial data are performed to test the proposed classifier.

II. EM FOR FINITE MIXTURE MODELS

The mixture model approach assumes that the data $X=\{\mathbf{x}_1,...,\mathbf{x}_n\}$ in R^d (a d-dimensional feature space) arises from a linear combination of component density functions resulting in a mixture probability density function of the form:

$$f(X)=\sum_{k=1}^{K}\tau_k f_k(X\mid\mu_k,\Sigma_k) \tag{1}$$

where K is the number of mixture components, the τ_k's are the mixing proportions with $\sum_{k=1}^{K}\tau_k=1$, $0\le\tau_k\le 1,\ 1\le k\le K$ and $f_k(X\mid\mu_k,\Sigma_k)$ denotes the conditional density of class k given mean vector μ_k and covariance matrix Σ_k.

There are two commonly used approaches in mixture analysis, the mixture approach and the classification approach. In short, the mixture approach aims to maximize the likelihood over the mixture parameters, whereas the classification approach aims to maximize the likelihood over the mixture parameters and the identifying labels of the mixture components for each sample.

In the mixture approach there is no direct interest in the discrete labeling of the samples. More specifically, the parameter set is chosen to maximize the log-likelihood.

$$L(\Theta\mid X)=\sum_{i=1}^{n}\ln[\sum_{k=1}^{K}\tau_k f_k(\mathbf{x}_i\mid\mu_k,\Sigma_k)] \tag{2}$$

In the classification approach, the indicator vectors, $\mathbf{z}_i=(z_{ik},k=1,...,K)$ with $z_{ik}=1$ or 0 identifies the mixture component, according as $\mathbf{x}_i$ $(1\le i\le n)$ has been drawn from the k^{th} component or from another one and they are treated as unknown parameters.

Our main concern is to identify the origin of each sample within a class; therefore we will adopt the classification log-likelihood approach in this paper. In what follows EM equations for the classification log-likelihood will be derived.

In EM for mixture models, the complete data are considered to be $X=(Y,Z)$ where $Y=\{\mathbf{y}_1,...,\mathbf{y}_n\}$ is observable (i.e., labeled samples) and $Z=\{\mathbf{z}_1,...,\mathbf{z}_n\}$ constitutes the "missing data" (i.e., unlabeled samples). In other words there exists a finite set of k states, and that each $\mathbf{y}_i$ is associated with an indicator vector $\mathbf{z}_i$ of length k whose components are all zero except for one indicating the unobserved state associated with $\mathbf{y}_i$.

Let $\mathbf{z}_i=\{z_{i1},...,z_{iK}\}$ with $z_{ik}=1$ if $\mathbf{y}_i$ belongs to component k and $z_{ik}=0$ otherwise and each $\mathbf{z}_i$ be independent and identically distributed according to a multinomial distribution of one draw on K categories with probabilities $\tau_1,...,\tau_K$ and $\mathbf{y}_i$ given $\mathbf{z}_i$ be independent and identically distributed. Then assuming $\mathbf{z}_i$ is associated with $\mathbf{y}_i$ and $\mathbf{y}_i$ belongs to component j, the probability of $\mathbf{z}_i$ becomes

$$f(\mathbf{z}_i)=\frac{1!}{0!...0!1!0!...0!}\tau_1^{0}...\tau_{j-1}^{0}\tau_j^{1}\tau_{j+1}^{0}...\tau_K^{0}=\tau_j \tag{3}$$

and the probability of $\mathbf{y}_i$ given $\mathbf{z}_i$ becomes

$$f(\mathbf{y}_i\mid\mathbf{z}_i)=f_j(\mathbf{y}_i\mid\mu_j,\Sigma_j) \tag{4}$$

Combining (3) and (4) to obtain $f(\mathbf{x}_i)$ yields

$$f(\mathbf{x}_i)=f_j(\mathbf{y}_i\mid\mu_j,\Sigma_j)\tau_j \tag{5}$$

Suppose j is unknown, since $z_{ik}=1$ if $\mathbf{y}_i$ belongs to component k and $z_{ik}=0$ otherwise (5) can be generalized as,

$$f(\mathbf{x}_i)=\prod_{k=1}^{K}[f_k(\mathbf{y}_i\mid\mu_k,\Sigma_k)\tau_k]^{z_{ik}} \tag{6}$$

For all i=1, ,n (6) can be written as,

The initial partitioning of the data has a crucial importance on the output of the EM stage, as EM will converge to a local maximum, which will be in the neighborhood of the starting point.

The discrete partitioning algorithms by which the initialization is done can be grouped into two types: The hierarchical and nonhierarchical approaches. The fact that initialization is not required makes the hierarchical methods very appealing at first sight. Some hierarchical methods even guarantee global optimality. The *Branch and Bound* algorithm [4] is of this type. Both the global optimality and a good initialization is what we are looking for. However, there is a major drawback of hierarchical methods, which makes them very impractical to use in some cases. That is, the time required for the algorithm to converge increases rapidly with the number of samples and the dimensionality of the data.

The scattering of the data is another negative factor on the performance of some of the hierarchical algorithms. Although some efficient implementation techniques have recently been developed, hierarchical algorithms are still far from accommodating a few thousand samples in hyperdimensional space. Hyperdimensionality is an inherent characteristic of hyperspectral data analysis, and an efficient identification of subclasses usually requires a considerable number of training samples. For this reason, a hierarchical approach is not adopted in this work.

$$f(X) = \prod_{k=1}^{K} \prod_{i=1}^{n} [f_k(y_i \mid \mu_k, \Sigma_k)\tau_k]^{z_{ik}} \tag{7}$$

Finally, the resulting complete-data log likelihood is

$$L(\Theta, Z \mid X) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} [\log \tau_k f_k(\mathbf{x}_i \mid \mu_k, \Sigma_k)] \tag{8}$$

The quantity z_{ik} (i=1, ,n and k=1, ,K) for equation (8) can be estimated as the conditional expectation of z_{ik} given the observation $\mathbf{x}_i$ and the parameter set Θ.

The EM algorithm iterates between an E-step in which values of z_{ik} are estimated from the data with the current parameter estimates as $\hat{z}_{ik}$, and an M-step in which the complete-data log likelihood (8), with z_{ik} replaced by its current conditional expectation $\hat{z}_{ik}$, is maximized with respect to the parameters. The outline of the algorithm is given in Fig. 1. Under fairly weak regularity conditions [11], the method can be shown to converge to a local maximum of the classification likelihood.

We use nearest mean clustering to initialize the EM algorithm. Compared to hierarchical methods, nearest-mean clustering is very efficient in computational time but does not guarantee a convergence to a global optimum point or a convergence at all and requires an initialization for itself. The initialization method we use is data dependent, which consists of initializing the centers of each cluster using principal components. For well-separated data, a shift in the initial cluster centers will not affect the result of the algorithm. However as the number of samples increases and the data becomes more scattered the algorithm is likely to produce a different output when started with a different set of cluster centers.

III. MODEL BASED MIXTURE IDENTIFICATION

As stated earlier, in hyperspectral analysis the performance of the classifier is highly restricted by the number of training samples available and due to the high dimensional nature of the data. Not surprisingly, this restriction is more severe in the mixture classifier than in any simple quadratic classifier. By partitioning the already small set of training data into multiple clusters and then estimating the cluster statistics, one ends up with a mixture model whose component statistics and thus overall statistics are ill-conditioned unless each cluster has at least d+1 samples, where d is the number of dimensions. Assuming each class of data is to be partitioned into a designated number of clusters K, we should have at least K(d+1) training samples for each class and these samples should be partitioned evenly among clusters when a clustering algorithm is run. This is practically hard to achieve, let alone having K(d+1) training samples for each class. Therefore the EM equations derived in the previous section based on the *unconstrained* covariance model is by itself not efficient.

One way around this problem is to use regularized covariance estimators based on leave-one-out likelihood (LOOL). LOOC introduced in [5], BLOOC introduced in [6], and MIXED_LOOC [7] are of this type. For a quadratic classifier, these estimators examine the sample covariance and the common covariance estimates, as well as their diagonal (LOOC) or trace forms (BLOOC), to determine which would be most appropriate. The same idea can be used in mixture classifiers by applying these estimators to each individual class training data. However in mixture identification, the main emphasis is on model fit. That is to say, statistics estimation is not the only concern as in simple quadratic classifiers. Therefore the success of LOOL based estimators would be very limited in mixture classifiers unless statistics estimation is accompanied with EM.

initialize $\hat{z}_{ik}$ (This is done through the output of nearest-mean clustering)
repeat

M-step: maximize (8) given $\hat{z}_{ik}$ ($f(\mathbf{x} \mid \mu, \Sigma)$ is a Gaussian density)

$$\hat{n}_k = \sum_{i=1}^{n} \hat{z}_{ik} \tag{9}$$

$$\hat{\tau}_k = \frac{\hat{n}_k}{n} \tag{10}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik} \mathbf{x}_i}{\hat{n}_k} \tag{11}$$

$$\hat{\Sigma}_k = \frac{\sum_{i-1}^{n} \hat{z}_{ik} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T}{\hat{n}_k} \tag{12}$$

E-step: compute z_{ik} given the parameter estimates from the M-step.

$$\hat{z}_{ik} = \frac{\hat{\tau}_k f_k(\mathbf{x}_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^{K} \hat{\tau}_j f_j(\mathbf{x}_i \mid \hat{\mu}_j, \hat{\Sigma}_j)} \tag{13}$$

until convergence criterion is satisfied

Figure 1. The EM algorithm for Gaussian mixture models. The indicator variables are initialized through the output of nearest-mean clustering and the algorithm is terminated when the relative difference between successive values of classification log likelihood falls below a threshold or the designated maximum number of iterations is exceeded.

On the other hand, when working with EM, the structure of the covariance matrix is needed as *a priori* information to avoid possible and unexpected breakdowns of the algorithm. Unfortunately LOOL based estimators do not provide this information, as the covariance matrix obtained this way need not necessarily be unconstrained. When a diagonal, a trace, or a common covariance form is favored alone by these estimators, and equation (12) is used to update the covariance matrices, the structure of these matrices cannot be preserved. As a result, the statistics estimation may change substantially during the EM stage.

In this paper we propose a mixture classifier whose core component is the model based mixture identification. The new approach will eliminate the above problems to a greater extent. In this framework the covariance matrix of each cluster is assumed to be a weighted mixture of constrained and unconstrained covariance matrices. The proposed covariance estimator has the following form:

$$C_{jk}(\alpha_{jk}) = (1 - \alpha_{jk})\Sigma_{jk} + \alpha_{jk}\Psi_j \tag{14}$$

where $0 \le \alpha_{jk} < 1,\ 1 \le j \le N,\ 1 \le k \le m_j$, Σ_{jk} is the unconstrained covariance of cluster k in class j, α_{jk} is the corresponding mixing parameter, m_j is the number of clusters in class j, and N is the total number of classes. Note that, m_j can vary from one up to a designated number K. In the above formulation, Ψ_j is the unknown covariance structure of class j, which is to be chosen among six possible covariance models through a process involving clustering, EM, and model selection.

In what follows, EM equations for the six covariance models considered in this paper will be derived. Note that a change in the covariance model only affects the update equation for the covariance matrix. The rest of the equations in Figure 1 remain the same. Before passing into the derivations, two additional matrices will be defined. These are the within class scatter matrix W_j estimated by,

$$\hat{W}_j = \sum_{k=1}^{m_j} \sum_{i=1}^{n_j} \hat{z}_{ik} (\mathbf{x}_i - \hat{\mu}_{jk})(\mathbf{x}_i - \hat{\mu}_{jk})^T \tag{15}$$

and the within cluster scatter matrix W_{jk} estimated by,

$$\hat{W}_{jk} = \sum_{i=1}^{n_j} \hat{z}_{ik}(\mathbf{x}_i - \hat{\mu}_{jk})(\mathbf{x}_i - \hat{\mu}_{jk})^T \quad (16)$$

where n_j is the number of samples, m_j is the number of clusters in class j, and $\hat{z}_{ik}$, $\hat{\mu}_{jk}$ are estimated through update formulas in Fig. 1. For a multivariate quadratic case, equation (8) can be expanded as follows:

$$L_j(\Sigma_{j1},\mathrm{K},\Sigma_{jm_j} \mid X) = \sum_{k=1}^{m_j}\sum_{i=1}^{n_j} -\hat{z}_{ik}(\log|\Sigma_{jk}| + tr[(\mathbf{x}_i - \hat{\mu}_{jk})(\mathbf{x}_i - \hat{\mu}_{jk})^T \Sigma_{jk}^{-1}] + c) \quad (17)$$

where c is a constant with respect to Σ_k and tr(.) is the trace operator. After substituting $\hat{W}_{jk}$ and $\sum_{i=1}^{n_j} \hat{z}_{ik} = \hat{n}_{jk}$ into (17), we end up with the following equation.

$$L_j(\Sigma_{j1},\mathrm{K},\Sigma_{jm_j} \mid X_j) = \sum_{k=1}^{m_j} -(\hat{n}_{jk}\log|\Sigma_{jk}| + tr[\hat{W}_{jk}\Sigma_{jk}^{-1}] + c) \quad (18)$$

Covariance models:

1) Unconstrained case: No restriction is placed on the covariance matrices Σ_{jk}. The update equation for this case is already given in (12). Maximizing (18) leads to the minimization of

$$L_j(\Sigma_{j1},\mathrm{K},\Sigma_{jm_j} \mid X_j) = \sum_{k=1}^{m_j} (\hat{n}_{jk}\log|\Sigma_{jk}| + tr[\hat{W}_{jk}\Sigma_{jk}^{-1}] + c) \quad (19)$$

and the covariance matrices Σ_{jk} are estimated by

$$\hat{\Sigma}_{jk} = \frac{1}{\hat{n}_{jk}}\hat{\mathbf{W}}_{jk} \quad (20)$$

2) $\Sigma_{jk} = \sigma_j \mathrm{I}$: All the clusters have the same spherical covariance matrix. In this situation, maximizing (18) leads to the minimization of

$$L_j(\sigma_j \mid X_j) = n_j d\log(\sigma_j) + \frac{1}{\sigma_j}\mathrm{tr}[\hat{W}_j] + c \quad (21)$$

and we get

$$\sigma_j = \frac{\mathrm{tr}[\hat{W}_j]}{n_j d} \quad (22)$$

3) $\Sigma_{jk} = \sigma_{jk}\mathrm{I}$: Clusters are spherical with different volumes. In this situation maximizing (18) leads to the minimization of

$$L_j(\sigma_{j1},\mathrm{K},\sigma_{jm_j} \mid X_j) = d\sum_{k=1}^{m_j} \hat{n}_{jk}\log(\sigma_{jk}) + \sum_{k=1}^{m_j}\frac{1}{\sigma_{jk}}\mathrm{tr}[\hat{W}_{jk}] + c \quad (23)$$

and we get

$$\sigma_{jk} = \frac{\mathrm{tr}[\hat{W}_{jk}]}{n_{jk} d} \quad (24)$$

4) $\Sigma_{jk} = D_j$: All the clusters have the same diagonal covariance matrix. In this situation maximizing (18) leads to the minimization of

$$L_j(D_j \mid X_j) = n_j\log(|D_j|) + \mathrm{tr}[\hat{W}_j D_j^{-1}] + c \quad (25)$$

and we get

$$D_j = \frac{\mathrm{diag}(\hat{W}_j)}{n_j} \quad (26)$$

5) $\Sigma_{jk} = D_{jk}$: All the clusters have different diagonal covariance matrices. In this situation maximizing (18) leads to the minimization of

$$L_j(D_{j1},\mathrm{K},D_{jm_j} \mid X_j) = \sum_{k=1}^{m_j}(\hat{n}_{jk}\log(|D_{jk}|) + \mathrm{tr}[\hat{W}_{jk}D_{jk}^{-1}] + c) \quad (27)$$

and we get

$$D_{jk} = \frac{\mathrm{diag}(\hat{W}_{jk})}{n_{jk}} \quad (28)$$

6) $\Sigma_{jk} = \Sigma_j$: All the clusters have the same covariance matrix. In this situation maximizing (18) leads to the minimization of

$$L_j(\Sigma_j \mid X_j) = \hat{n}_j\log(|\Sigma_j|) + \mathrm{tr}[\hat{W}_j\Sigma_j^{-1}] + c \quad (29)$$

and we get

$$\Sigma_j = \frac{\hat{W}_j}{n_j} \quad (30)$$

For a fixed number of clusters k, the training data for each class is fit into each of the six models through clustering and EM. Then, corresponding to each model, a measure of fit is computed and the model with the highest measure of fit is chosen as Ψ_j. Once the structure of the mixture is determined, the next step is to estimate the mixing proportions, α_{jk} which can be done by maximizing the LOOL for each cluster. This process is repeated by incrementing the number of clusters, k by one up to a designated number K, and a measure of fit is computed at each stage. Finally the mixture model with the highest measure of fit is chosen as the best fit.

Computing a measure of fit for a few thousand samples in a hyperdimensional space is computationally very challenging if not impossible. For the practicality it provides in this study, we adopt the Bayesian Information Criterion (BIC) [8] to compute the measures of fit. Although regularity conditions for the BIC do not hold for mixture models, there is considerable theoretical and practical support for its use in this context [14], [15]. The closed form expression for BIC is given below.

$$2L_{M_k}(X,\hat{\Theta}_k) - d_{M_k}\log(n) \quad (31)$$

where $L_{M_k}(X,\hat{\Theta}_k)$ is the logarithm of the maximized mixture likelihood for the model and d_{M_k} is the number of independent parameters to be estimated in the model. The term on the right in (31) is known as the *penalization term*. It penalizes the complexity of the model and this allows us to compare models with differing parameterizations, differing number of components or both. This is not possible by the log-likelihood alone, which increases as more terms are

added to the model. A standard convention for interpreting BIC differences is given in Table 1 and the number of independent parameters for the six models considered in the previous section is given in Table 2.

Table 1. Indication of evidence for different BIC values [8].

$2\log_e(B_{10}) \approx$ BIC difference	Evidence
0-2	Weak
2-6	Positive
6-10	Strong
>10	Decisive

Table 2. Number of independent parameters (A=kd+k-1, B=d(d+1)/2).

Model	Symbol	Number of parameters
Σ_{jk}	U	A+kB
$\Sigma_{jk} = \sigma_j I$	TE	A+1
$\Sigma_{jk} = \sigma_{jk} I$	TV	A+k
$\Sigma_{jk} = D_j$	DE	A+d
$\Sigma_{jk} = D_{jk}$	DV	A+kd
$\Sigma_{jk} = \Sigma_j$	E	A+B

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, several experiments are conducted to test the performance of the proposed mixture classifier (MC) and compare it with that of a simple quadratic classifier (SQC). To make a consistent comparison between these two classifiers both are designed similarly. In fact, the only difference is the maximum number of clusters (K) considered. Unlike the mixture classifier, K is always one for the simple quadratic classifier.

A. Experiment 1: Purdue Flightline

This data set is a flightline over the Purdue University West Lafayette Campus. The hyperspectral data used was collected on September 30, 1999 with the airborne HYMAP system [12], providing image data in 126 spectral bands in the visible and IR regions (0.4 μm — 2.4 μm). The system was flown at an altitude such that the pixel size is about 5 meters. The list of classes and number of labeled samples for each class is given in Table 3.

This data set is very challenging to analyze in many respects. First it is difficult to define a set of efficient informative classes. There are three parking garages and many parking lots throughout the campus, and they all show similar spectral behavior in the image data due to the presence of large numbers of cars. It is very hard to distinguish between these two unless another class is assigned for cars. Second, some of the parking lots are graveled, and they differ from others, which are made of asphalt. Third, rooftops are made of a great variety of materials including glass (garden frame). Fourth, paths are spectrally very similar to some rooftops and streets. Fifth, basketball fields, tennis courts, tracking fields, all of which may show different spectral behavior, are all grouped into the same class.

After defining the classes desired, the next step is to perform a feature extraction and choose the best set of features. At this point we avoid using parametric feature extraction techniques, as feature information relevant to mixture analysis may be lost when the class densities are assumed as Gaussian and a feature extraction is performed based on this assumption. We use the Nonparametric Weighted Feature Extraction (NWFE) technique recently introduced by Kuo and Landgrebe [9]. We randomly choose a portion (r) of the labeled samples for each class as training samples and perform a feature extraction using NWFE. Then the best 10, 20 and 30 features are selected respectively, and for each case a classification is performed. The classifiers are then tested using the remaining portion of the labeled samples. For r=0.2, d=30 the mixture classifier takes around 20 minutes to complete the entire job on a Pentium based computer with a clock frequency of 1.2 gigahertz. Results of these classifications are shown in Fig. 2, and the number of subclasses identified in each class is given in Table 3. A Purdue campus map is provided in Fig. 3a to designate some landmark structures. Note that this map is older than the aerial scene, it is slightly miss-scaled, and only the university facilities are shown. Based on the ground truth data a test field map is provided in Fig. 3b and thematic maps for the simple quadratic and mixture classifications are given in Fig. 3c and 3d for r=0.2.

The results reveal an interesting fact. For the simple quadratic classifier increasing the number of training samples doesn t lead to a better classifier performance especially when d is small. This is observed in Fig 2. for d=10 and d=20. For d=30 this effect is somewhat mitigated. In the mixture classifier case, we have observed that more training samples lead to better classifier performance. This difference in performance can more clearly be seen by comparing Fig. 3c and Fig. 3d for r=0.2 (see parts of the images taken inside the rectangles)

Table 3. List of classes and number of labeled samples in each class for experiment 2.

Class Name	Number of Labeled Samples
Roof	10916
Streets	1779
Shadow	521
Grass	1459
Trees	956
Others	512
Cars	1092
Path	457
Total number of samples	17692

Table 4. Number of subclasses identified in each class for experiment 2.

d	**10**			**20**			**30**		
r	**0.01**	**0.1**	**0.2**	**0.01**	**0.1**	**0.2**	**0.01**	**0.1**	**0.2**
K	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
Roof	3	5	3	1	3	5	1	3	4
Streets	1	2	3	1	1	2	1	1	1
Shadows	1	2	3	1	1	1	1	1	1
Grass	1	2	2	1	2	2	1	1	2
Trees	1	2	2	1	1	1	1	1	1
Others	2	2	2	1	2	2	1	1	1
Cars	1	1	1	1	1	1	1	1	1
Path	2	1	1	1	1	1	1	1	1

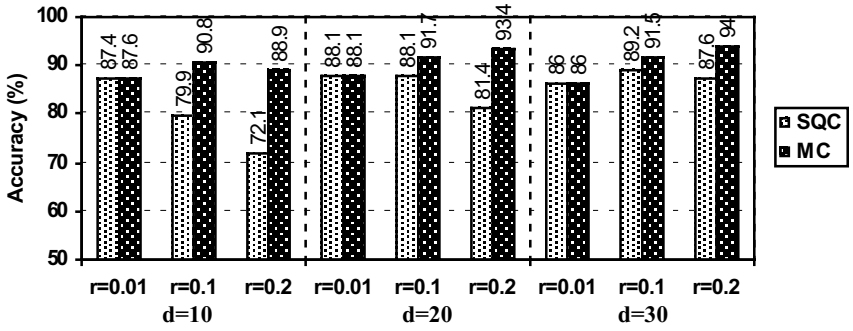

Figure 2. Comparison of classifier performances for experiment 2.

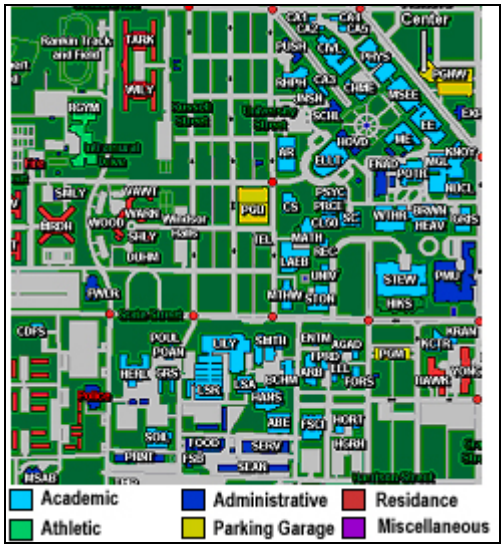

Figure 3a. False color image for ground truth data for Purdue Main Campus [13] (Only university facilities are shown).Original in color, see [16] for a color version.

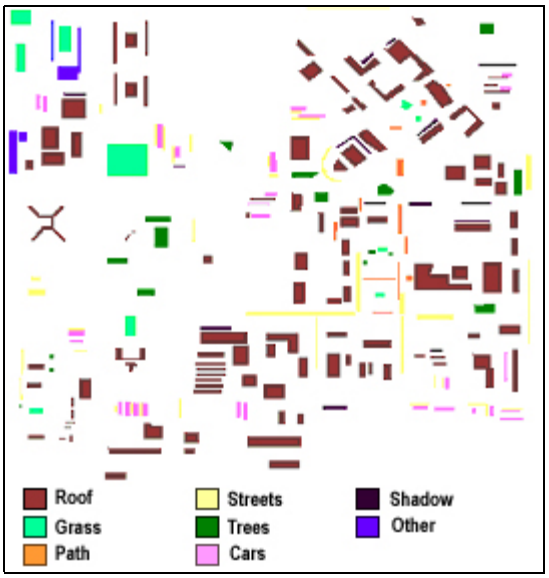

Figure 3b. False color image for test fields used in experiment 1. Original in color, see [16] for a color version.

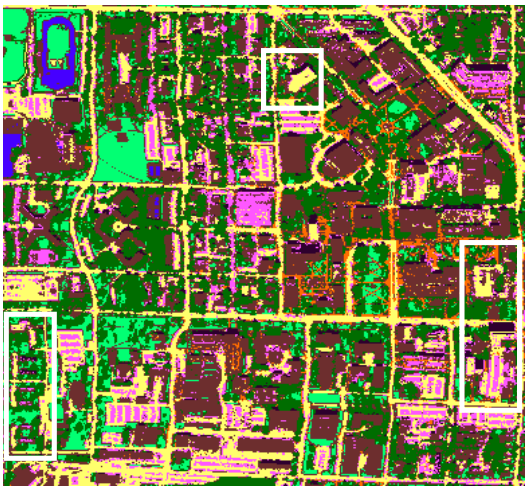

Figure 3c. False color image of the classification map obtained by the Simple Quadratic Classifier (d=30,r=0.2, Acc: 87.6 %). Original in color, see [16] for a color version.

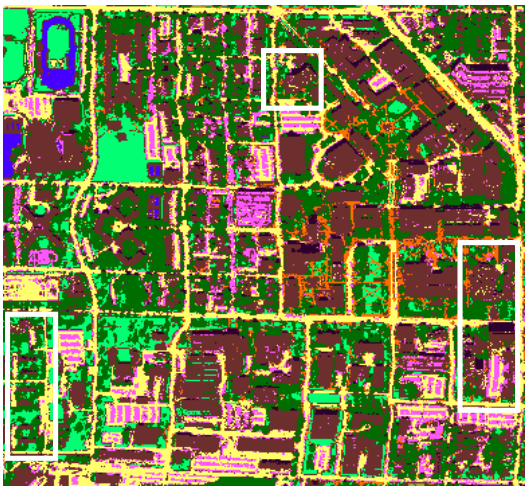

Figure 3d. False color image of the classification map obtained by the Mixture Classifier (d=30,r=0.2, Acc: 94.0 %). Original in color, see [16] for a color version.

The following experiments are performed using AVIRIS [10] data taken in 1992 over two sites: Cuprite, Nevada and Indian Pine.

B. Experiment 2: Cuprite Site

The Cuprite site covers an interesting geological feature called a hydrothermal alteration zone, which is exposed due to sparse vegetation. A total of 2744 labeled samples and 191 bands (0.40-1.34, 1.43-1.80, 1.96-2.46 m) are used in the experiment. The classifications are performed using the 10 features extracted by NWFE. Half of the labeled samples are used as training samples and the other half is used as testing samples. The numbers of subclasses identified in each class are shown in Table 5, and the classification results are shown in Fig 4.

Table 5. Number of subclasses identified in each class for experiment 2 (K=5).

Classes	# of subclasses
Alunite	2
Buddingtonite	1
Kaolinite	2
Quartz	1
Alluvium	4
Playa	1
Tuff	1
Argillized	1

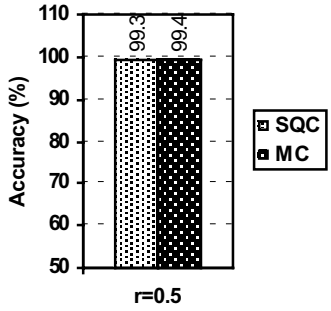

Figure 4. Comparison of classifier performances for experiment 2.

C. Experiment 3: Indian Pine Site

In this experiment, the data taken over the Indian Pine test site is used. This is a mixed forest/agricultural area in Indiana. A total of 2521 labeled samples and 191 bands (0.40-1.34, 1.43-1.80, 1.96-2.46 m) are used in the experiment. The classifications are performed using the 10 features extracted by NWFE. Half of the labeled samples are used as training samples and the other half is used as testing samples. The numbers of subclasses identified in each class are shown in Table 6 and the classification results are shown in Fig 5.

Table 6. Number of subclasses identified in each class for experiment 3 (K=5).

Classes	# of subclasses
Beans/Corn Residue	2
Corn/No Residue	1
Corn/Bean Residue	2
Beans/No Residue	2
Corn/Wheat Residue	2
Wheat/No Residue	2

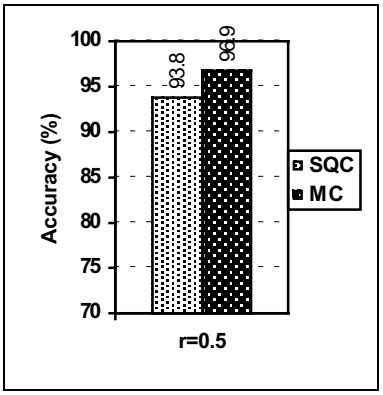

Figure 5. Comparison of classifier performances for experiment 3.

V. CONCLUSION

Experimental results favoring the proposed mixture classifier over the simple quadratic classifier in the majority of the cases show that characterizing class densities by a mixture density is useful in obtaining more precise class definitions. However as with most other classifiers, the success of the mixture classifier is also limited by the size of the training data set available. The main difference between a simple quadratic classifier and a mixture classifier is that the former does not guarantee a better performance when a larger training data set is used during the design process but the latter usually does, given enough components and separability among classes. In fact the larger the training data set, the more accurately the densities are estimated and hence the better the performance of the mixture classifier.

Apart from the limited training size problem, which is one of the most basic problems of hyperspectral data analysis, there are some other stages in the design of the mixture classifier that requires further attention.

One of these is the model selection stage. BIC doesn t provide reliable information on covariance

structure when the number of samples is very small. As a matter of fact, it will overestimate the number of components and introduce redundant parameters that will make estimations less reliable. A possible solution might be to use a common covariance for all classes and update EM equations accordingly.

Initial partitioning of the data is another stage that needs attention, as EM will converge to a local optimum that is closest to the initial starting point. We use nearest mean clustering to initialize EM but more efficient initializations can be found by trying different clustering algorithms.

REFERENCES

[1] D. A. Landgrebe, "On the Relationship Between Class Definition Precision and Classification Accuracy in Hyperspectral Analysis", *Proceedings of the International Geoscience and Remote Sensing Symposium*, Honolulu, Hawaii, 2000.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, San Diego: Academic Press, 1990, pp. 51-100.

[3] J. P. Hoffbeck and D. A. Landgrebe, "A Method for Estimating the Number of Components in a Normal Mixture Density Function", *Proceedings of the International Geoscience and Remote Sensing Symposium*, Honolulu, Hawaii, 2000.

[4] Warren L. G. Koontz, Patrenahalli M. Narendra, Keinosuke Fukunaga, "A Branch and Bound Clustering Algorithm", *IEEE Trans. Computers*, V. 24, September 1975.

[5] J.P. Hoffbeck and D.A. Landgrebe, "Covariance Matrix Estimation and Classification with Limited Training Data" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 18, NO. 7, pp. 763-767, July 1996.

[6] Saldju Tadjudin and David Landgrebe, "Covariance Estimation With Limited Training Samples," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 37, No. 4, pp. 2113-2118, July 1999.

[7] Bor-Chen Kuo and David A. Landgrebe, "A Covariance Estimator for Small Sample Size Classification Problems and Its Application to Feature Extraction**,"***IEEE Transactions on Geoscience and Remote Sensing,* Vol. 40, No. 4, pp 814-819, April 2002.

[8] R. E. Kass and A. Raftery, "Bayes Factors", *Journal of the American Statistical Association,* 90:773-795, ? 1995.

[9] Bor-Chen Kuo and David A. Landgrebe "Hyperspectral Data Classification Using Nonparametric Weighted Feature Extraction," International Geoscience and Remote Sensing Symposium, Toronto, Canada, 2002.

[10] Airborne Visible/Infrared Imaging Spectrometer, NASA Jet Propulsion Lab (http://makalu.jpl.nasa.gov)

[11] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extension*, Wiley, 1997.

[12] Kruse, F. A., Boardman, J. W., Lefkoff, A. B., Young, J. M., Kierein-Young, K. S., Cocks, T. D., Jenssen, R., and Cocks, P. A., HyMap: An Australian Hyperspectral Sensor Solving Global Problems - Results from USA HyMap Data Acquisitions: *Proceedings of the 10th Australasian Remote Sensing and Photogrammetry Conference*, Adelaide, Australia, August 21-25, 2000, Causal Productions (www.causalproductions.com), published on CD-ROM.

[13] Purdue University Campus Map (http://www.eas.purdue.edu/map/campus/)

[14] C. Fraley and A. E. Raftery, How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis , *Computer Journal,* 41:578-588, ? 1998.

[15] C. Fraley and A. E. Raftery, Model-Based Clustering, Discriminant Analysis, and Density estimation , *Journal of the American Statistical Association 97:611-631, 2002.*

[16] IEEE Xplore Website, (http://ieeexplore.ieee.org/Xplore/DynWel.jsp). See also http://dynamo.ecn.purdue.edu/~landgreb/publications.html.

M. Murat Dundar received the B.Sc degree from the Bogazici University, Istanbul, Turkey in 1997. He received his M.Sc degree from Purdue University, USA, in 1999 where he is currently pursuing his Ph.D degree. He has been with the Los Alamos National Laboratory during the summer of 2000, 2001 and 2002 as a graduate research assistant. His research interests are hyperspectral data analysis, statistical pattern recognition and machine learning.

M. Murat Dundar

David A. Landgrebe

David Landgrebe (S 54,—M 57—SM 74—F 77—LF 97) received the BSEE, MSEE, and PhD degrees from Purdue University, West Lafayette, IN.

He is Professor Emeritus of Electrical and Computer Engineering at Purdue University. His area of specialty in research is communication science and signal processing, especially as applied to Earth observational remote sensing.

Dr. Landgrebe was President of the IEEE Geoscience and Remote Sensing Society for 1986 and 1987 and a member of its Administrative Committee from 1979 to 1990. Dr. Landgrebe is a Life Fellow of the Institute of Electrical and Electronic Engineers, a Fellow of the American Society of Photogrammetry and Remote Sensing, a Fellow of the American Association for the Advancement of Science, a member of the Society of Photo-Optical Instrumentation Engineers and the American Society for Engineering Education, as well as Eta Kappa Nu, Tau Beta Pi, and Sigma Xi honor societies. He received the NASA Exceptional Scientific Achievement Medal in 1973 for his work in the field of machine analysis methods for remotely sensed Earth observational data. In 1976, on behalf of the Purdue's Laboratory for Applications of Remote Sensing, which he directed, he accepted the William T. Pecora Award, presented by NASA and the U.S. Department of Interior. He was the 1990 individual recipient of the William T. Pecora Award for contributions to the field of remote sensing. He was the 1992 recipient of the IEEE Geoscience and Remote Sensing Society s Distinguished Achievement Award.