REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

Public Reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comment regarding this burden estimates or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188,) Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE 11/29/01	3. REPORT TYPE AND DATES COVERED
	15	Final Progress Report November 1997-May 2001 31 st

4. TITLE AND SUBTITLE
Unsupervised Classification System for Hyperspectral Data Analysis

5. FUNDING NUMBERS
Grant DAAG55-98-1-0016

6. AUTHOR(S)
PI. Dr. Luis O. Jimenez
CoPIs Dr. Miguel Velez and Dr. Shawn Hunt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
ECE Department
PO Box 9042
University of Puerto Rico at Mayaguez
Mayaguez, PR 00681-9042

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
Department of the Army
US Army Research Office
~~AMXRO-AAA~~ 37839-PH-DPS
.7

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

There is an increasing interest in remote sensed vision systems for surveillance, object recognition, target identification, and cartography. Other fields that could benefit from such systems are land use and land cover administration, estimation of water sedimentation, and the creation of maps. Of particular interest are automatic systems that are robust with respect to analyst's knowledge in areas such as image analysis and computer vision. Remote Sensed image analysis focuses on obtaining information from radar, multispectral and hyperspectral images. Hyperspectral data enables the analyst to detect more materials, objects, and regions with more accuracy than previously possible. As the number of bands of high spectral resolution data increases, the capability to detect more detailed classes should also increase and the classification accuracy should increase as well. The curse of dimensionality has been known for more than three decades. There is a need for the development of algorithms for detection, and classification that utilize the amount of information and separability that hyperdimensional data offers while simultaneously avoiding the difficulties inherent in hyperdimensional space. The present report will summarize the research done in the areas of clustering, parameter estimation of hyperspectral data, band subset selection, data compression and unsupervised decision fusion mechanism.

14. SUBJECT TERMS
Unsupervised classification, Hyperspectral data, Subset Selection, Lossless compression, Parameter estimation, Decision Fusion, Fuzzy Clustering, Neural Network Clustering, Regularization.

15. NUMBER OF PAGES
44

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev.2-89)
Prescribed by ANSI Std. 239-18
298-102

Enclosure 1

20020131 148

MASTER COPY: PLEASE KEEP THIS "MEMORANDUM OF TRANSMITTAL" BLANK FOR REPRODUCTION PURPOSES. WHEN REPORTS ARE GENERATED UNDER THE ARO SPONSORSHIP, FORWARD A COMPLETED COPY OF THIS FORM WITH EACH REPORT SHIPMENT TO THE ARO. THIS WILL ASSURE PROPER IDENTIFICATION. NOT TO BE USED FOR INTERIM PROGRESS REPORTS; SEE PAGE 2 FOR INTERIM PROGRESS REPORT INSTRUCTIONS.

MEMORANDUM OF TRANSMITTAL

U.S. Army Research Office
ATTN: AMSRL-RO-BI (TR)
P.O. Box 12211
Research Triangle Park, NC 27709-2211

☐ Reprint (Orig + 2 copies)
☐ Technical Report (Orig + 2 copies)
☐ Manuscript (1 copy)
☒ Final Progress Report (Orig + 2 copies)
☐ Related Materials, Abstracts, Theses (1 copy)

CONTRACT/GRANT NUMBER: Grant DAAG55-98-1-0016

REPORT TITLE: Unsupervised Classification System for Hyperspectral Data Analysis

is forwarded for your information.

SUBMITTED FOR PUBLICATION TO (applicable only if report is manuscript):

Sincerely,

Luis O. Jimenez
Associate Professor
ECE Department
University of Puerto Rico at Mayaguez

Unsupervised Classification System for Hyperspectral Data Analysis

Luis O. Jimenez, Miguel Velez, and Shawn Hunt
Laboratory of Applied Remote Sensing and Image Processing
University of Puerto Rico at Mayaguez

1. Foreword

Hyperspectral remote sensing imaging involves collecting hundreds of multiple, contiguous, narrowband, spectra. This represents a collection of data nearly two orders of magnitude greater than current monochromatic and multispectral imaging. Specific applications of interest to the DoD include: *environment* - pollution detection; *geology* - mineral detection, surface materials, major rock types, altered rocks; *hydrology* - water quality, point and no point pollution; *archaeology* - further characterization of known area, localize dig; *agriculture* - type, structure, texture and moisture of soils; *forestry* - vegetation type mapping, quantification of biomass, stress detection; *oceanography* - mapping littoral areas, water characteristics; *marine biology* - characterizing surface environment; *endangered species* - characterizing known environment; *surveillance* - object recognition, target identification.

There is an increasing interest in remote sensed vision systems for surveillance, object recognition, target identification, and cartography. Other fields that could benefit from such systems are land use and land cover administration, estimation of water sedimentation, and the creation of maps. Of particular interest are automatic systems that are robust with respect to analyst's knowledge in areas such as image analysis and computer vision. Remote Sensed image analysis focuses on obtaining information from radar, multispectral and hyperspectral images. Hyperspectral data enables the analyst to detect more materials, objects, and regions with more accuracy than previously possible.

As the number of bands of high spectral resolution data increases, the capability to detect more detailed classes should also increase and the classification accuracy should increase as well. The curse of dimensionality has been known for more than three decades. There is a need for the development of algorithms for detection, and classification that utilize the amount of information and separability that hyperdimensional data offers while simultaneously avoiding the difficulties inherent in hyperdimensional space.

The present report will summarize the research done in the areas of clustering, parameter estimation of hyperspectral data, band subset selection, data compression and unsupervised decision fusion mechanism.

2. Table of Contents

3. List of Illustrations and Tables

3.1 List of Illustrations

3.2 List of Tables

4. Statement of Problem Studied

4.1 Unsupervised Classification System

There is an increasing interest in remote sensed vision systems for surveillance, object recognition, target identification, and cartography. Of particular interest are automatic systems that are robust with respect to analyst's knowledge in areas such as image analysis and computer vision. Remotely Sensed image analysis focuses on obtaining information from radar, multispectral and hyperspectral images. Hyperspectral data enables the analyst to detect more materials, objects, and regions with more accuracy than previously possible.

As the number of bands of high spectral resolution data increases, the capability to detect more detailed classes should also increase and the classification accuracy should increase as well. Parallel with such expectations there are some problems that need to be solved before being able to retrieve the potentially large amount of information from hyperspectral data. The problems that need to be solved are based on hyperdimensional space properties and their implications for remote sensing image classification. The curse of dimensionality has been known for more than three decades. In combinatorial optimization over many dimensions, it is seen as an exponential growth on the computational effort with the number of dimensions. In statistics, it manifests itself as a problem with parameter or density estimation due to the paucity of data. Local neighborhoods are almost surely empty of data, requiring the window of estimation to be large and producing the effect of losing detailed density estimation. This implies that there is not enough data to produce well estimated parameters yielding rank-deficient problems. Most of present detection and classification algorithms do not take into consideration the properties of hyperdimensional data. What is called for is the development of algorithms for detection, and classification that utilize the amount of information and separability that hyperdimensional data offers while simultaneously avoiding the difficulties inherent in hyperdimensional space.

The present report will summarize the research done in the areas of clustering, parameter estimation of hyperspectral data, band subset selection, data compression and unsupervised decision fusion mechanism.

4.2 Clustering (Dr. Luis O. Jimenez)

Clustering deals with the task of splitting a set of data points into a number of more-or less homogeneous classes (clusters), with respect to a similarity measure. There are different approaches to clustering based on different theoretical systems. Among the different cluster approaches are statistical approach, fuzzy clustering, neural network clustering and fuzzy logic-based neural network clustering to mention a few of them. In this research project we explored the applications of Fuzzy C-Means, Neural Network clustering and Fuzzy Neural Network to hyperspectral data clustering and the effect of dimensionality on the results.

4.2.1 Fuzzy C-Means Clustering

The Fuzzy C-Means clustering algorithm is very similar to the Hard C-Means in its iterative nature and in the computation of the cluster centers. On the other hand this two clustering schemes are different in the output they yield. Hard C-Means yields well-defined clusters with very well defined characteristics, and belonging to a cluster implies complete exclusion to the others. Fuzzy C-Means clustering yields a partitioning matrix:

$$U = \{u_{ij}\}; \tag{1}$$

where u_{kl} is the grade of membership of the j^{th} data point to the i^{th} cluster; such that ,

$$0 \leq u_{ij} \leq 1 \tag{2}$$

and,

$$\sum_{i=1}^{C} u_{ij} = 1 \forall j = 1,2,...,N, \tag{3}$$

where C is the total number of fuzzy clusters, N is the total number of data points and

$$2 \leq C < N. \tag{4}$$

This implies that each data point belongs to every cluster but with a degree of membership. The sum of the membership grades of one data point to all the different classes is normalized to be equal to 1. The focus of the algorithm is to optimize an objective function, which acts as a performance index of the clustering algorithm. The form for the objective function in this algorithm is:

$$J(u_{ij}, \mu_k) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^{m} * \| x_j - \mu_i \|^2 \tag{5}$$

To perform this minimization problem we first differentiate equation (5) with respect to μ_j, fixing u_{ij}, for $i = 1,2,...,C$ and $j = 1,2,...,N$ and then we differentiate equation (5) with respect to u_{ij} fixing μ_j, for $i = 1,2,...,C$ and we apply the conditions in equation (5), thus we get:

$$\mu_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}}, \forall i = 1,2,...,C \tag{6}$$

and,

$$u_{ij} = \frac{\left(1/\| x_j - \mu_i \|^2\right)^{1/m-1}}{\sum_{k=1} \left(1/\| x_j - \mu_k \|^2\right)^{1/m-1}}; \forall i = 1,...,C, j = 1,...,N \tag{7}$$

where x_j represents the j^{th} data point, m is called exponential weight and it influences the fuzziness of the membership, or partition matrix. Equation (6) and equation (7) have to be solved using an iterative method, which is what we call the Fuzzy C-Means (FCM) Clustering algorithm.

The steps for the FCM are:

1. Select a number of clusters C such that $2 \leq C < N$ and exponential weight m such that $1 \leq m < \infty$. Choose an initial partition matrix $U^{(0)}$, most of the time since we have no a priori knowledge of the data characteristics this initial partition

matrix is chosen completely arbitrary, satisfying equation (3). Also choose termination criterion ε and set iteration index $l = 0$.

2. Calculate the fuzzy cluster centers $\mu_i^{(l)}$ using $U^{(l)}$ and equation (6).
3. Calculate the new partition matrix $U^{(l+1)}$ by using $\mu_i^{(l)}$ and equation (7).
4. Calculate a difference relation between the current and previous iteration objective function values. For this case if:

$$\xi = (J(l\text{-}1)\text{-}J(l))/\, J(l\text{-}1) \geq \varepsilon, \tag{8}$$

then set $l = l + 1$ and go back to step 2. If $\xi \leq \varepsilon$,, then stop.

4.2.2. Neural Network Clustering

Neural network has to discover by itself the relationships or similarities between the data points. There are different types of learning rules for Neural Network unsupervised learning. The one used in this project is called the competitive learning rule, or winner takes all rule, specifically the Kohonen Learning Rule. Here learning is based on the clustering of the input data into similar groups and separation of dissimilar data points. The weights (W) of this network can be viewed as the centroid for the clusters in which we are going to fit the data points. The learning for the Kohonen Network is given by:

$$w_i^{k+1} = w_i^k + \alpha^k (X - w_i^k) \tag{9}$$

where;

$$w_i = \frac{W_i}{\|W_i\|} \tag{10}$$

This method is also called Vector Quantization. The weights of this system can be viewed as unit vectors. The system measures the cosine of the angle between the input vector and the weight vectors, and compare them. The weight vector that is more aligned to the input vector wins, and is updated, the other weight vectors remain unchanged. The system keeps on performing this task until the changes in the weight vectors reach a predetermined minimum.

4.2.3. Fuzzy Logic-Based Neural Network Clustering

Competitive learning is one of the major clustering techniques in neural networks. Making a generalization of the conventional competitive learning rules we can obtain a fuzzy competitive learning rule based on the FCM algorithm. The fuzzy competitive learning rule is derived by minimizing the objective function J in (5) applying the Gradient Descendent Method. The network to be implemented is one very similar to the Kohonen Network presented earlier with the exception

that the weights in this Network will be fuzzy weight vectors. Again this weights will correspond to the centroids of the data clusters. The learning rule for this network is given by;

$$\mu_i^{k+1} = \mu_i^k + \eta\gamma_m\left[x_j - \mu_i\right]; j = 1,2,...,n; i = 1,2,...,C \tag{11}$$

where;

$$\gamma_m = (u_{ij})^m[1 - m(m-1)^{-1}(1-u_{ij})] \tag{12}$$

with m and u having the same restrictions as in the FCM algorithm. We can see that this corresponds to the FCM algorithm with the difference that instead of classifying the entire data set and then updating the weights, the weights are updated as each input vector is evaluated by the system. It is important to see that it would correspond to the Crisp Kohonen Network if the fuzzy exponential weight (m) was equal to one.

4.3 Use of regularization to fix the parameter estimation process (Dr. Luis O. Jimenez)

This section presents the effect of the dimensionality of hyperspectral data in the estimation of the covariances and its consequence in the classification process. The required number of samples to train a quadratic classifier increases quadratically with the number of features (Jimenez and Landgrebe 1998). This is critical for hyperspectral data. Due to problems in the covariance matrix estimation the data points in a normal distribution have a tendency to be declared as an outlier with respect to the training samples. The definition of an outlier is as follows. **X** is an outlier of a normal distribution with mean μ_i and covariance Σ_i if

$$(\mathbf{X} - \mu_i)^T \Sigma_i^{-1}(\mathbf{X} - \mu_i) \geq T, \tag{13}$$

Where T is a value that takes into consideration the dimensionality of the data and is related to the Chi Square distribution. In a realistic setting using statistical pattern recognition the mean μ_i and covariance Σ_i are replace by the maximum likelihood estimates that are computed based on the training samples. The following experiment will show the consequence of not having an adequate number of training samples in the detection of outliers in hyperspectral data

Using hyperspectral data from the AVIRIS sensor that correspond to soil, 200 pixels were used for training samples and 200 used for testing samples. We rendered an experiment where we choose a T to reject the 1% less likely of the data as outliers. In Figure 1 we observe that using the ML estimate of the covariance around 1% of the training samples were rejected as outliers. This results is consistent with our expectations. For the testing samples the amount of data points that are classifiy as outliers increases with the dimensionality until 100% were declared outlier. When we used the identity matrix instead of the ML estimate the problem is mitigated for testing and augmented for training. For a well estimated covariance matrix only 1% of the training and testing samples should be declared outliers.

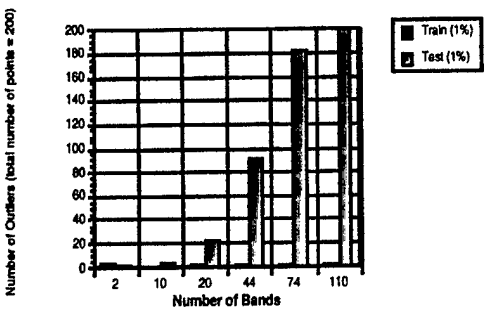

Figure 1a. Number of outliers in testing and training samples using the Maximum Likelihood covariance matrix.

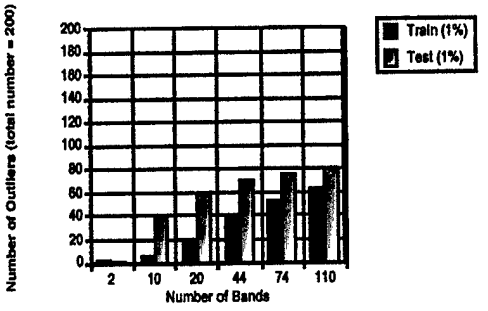

Figure 1b. Number of outliers in testing and training samples using the Identity matrix as the covariance matrix.

The main purpose of this part of the investigation was to apply regularization to the estimation of the covariance matrix when the size n_i of the training samples is not greater than the dimension d of the feature vector or if n_i is not appreciably larger than d. The regularization schema is as follows. A new regularized covariance matrix will be compute using the Maximum Likelihood estimate from the training sample and the identity matrix as is expressed in equation (14).

$$\hat{\Sigma}_i^{Reg}(\gamma) = (1-\gamma)\Sigma_i^{MLE} + \gamma\, c_i \mathbf{I}_p \tag{14}$$

Where $\hat{\Sigma}_i^{Reg}$ is the regularized covariance matrix, Σ_i^{MLE} is the maximum likelihood estimated, $\mathbf{I}_d$ is the corresponding identity matrix, d is the dimension of the feature vector [Duda, Hart and Stork, 2001]. The parameter of regularization γ could have values between 0 and 1. When $\gamma = 0$ we the regularized covariance is the Maximum Likelihood estimate. When $\gamma = 1$ the regularized covariance is equal to the identity matrix multiply by a constant. The effect of this in the identification of outliers will be shown in the results.

4.4 Band Subset Selection (Dr. Miguel Velez)

Observations from hyperspectral imaging sensors lead to high dimensional data sets from hundreds of images taken at closely spaced narrow spectral bands. High storage and transmission requirements, computational complexity, and statistical modeling problems combined with problem prior physical insight motivate the idea of dimension reduction using band selection. A standard approach for dimension reduction is principal component analysis [Geladi, P. and H. Grahn]. In this approach, the original hyperspectral image is transformed by means of linear transformations into a set of uncorrelated or orthogonal "images." A well-known fact for multispectral and hyperspectral images is that most of the spatial information content is summarized by the first few principal components [Geladi and Grahn][Schowengerdt]. A disadvantage of this approach is the inherent transformation of the original hyperspectral image (physical meaningful spectral data) into linear combinations of bands with, in many instances, little or no physical relation with the spectral information content on the original image.

An alternative dimensionality reduction approach that keeps the physical insight from spectral knowledge is band subset selection. In band subset selection, we select a subset of bands that in some sense summarizes most of the information contained in the original hyperspectral data cube. Band selection can be based on physical insight or optimization of some information content or of class separability measure [Pásztor and Csillag][Price][San Miguel-Ayanz and Biging][Van den Broek et. al]. From a classification point of view, band selection is a feature subset selection problem. By using band selection, we can reduce not only the cost of recognition by reducing the number of bands that need to be collected (as in reconfigurable sensors or in sensor design), but in some cases it can also provide better classification accuracy due to finite sample size effects [Jimenez and Landgrebe, 98][Jimenez and Landgrebe, 2000][Zongker and Jain].

The optimal band subset selection problem is a combinatorial optimization problem requiring the use of search methods to solve it. The main objective of this research work was the development of fast and reliable algorithms for band selection. In this research, we developed band selection methods based on subset selection methods presented in [Golub and Van Loan, 1997] to determine the subset of most independent columns in a matrix. Two approaches were studied based on singular value decomposition (SVD) and rank revealing QR (RRQR) factorization. In our work, we evaluated

- which one of the potential algorithms worked better in our application. We showed in [Vélez-Reyes, et al, 2000] that band selection based on SVD was more robust than band selection using RRQR in terms of closeness to the principal components.
- comparison of the SVD method to other methods that try to approximate the principal components presented in the literature. We showed in [Velez-Reyes et al, Sept. 2001] that SVD-based band selection was much better in selecting an approximation than the other methods because it deals with the band selection in a multivariable fashion compared to the other approaches that use a greedy approach.

Detailed presentation of these results is given in [Vélez-Reyes and Jiménez, Seattle 98][Vélez-Reyes et al, San Diego 97][Velez-Reyes et al, Orlando 2000][Velez-Reyes and Linares, Sept. 2001] (copies included). Applications of the band selection methods developed in this research to classification, compression and biomedical applications are presented in [Jimenez-Rodriguez et al, Sept. 2001][Jimenez et al, Orlando April 1999][Hunt and Velez-Reyes][Rodriguez-Diaz et al, July 2000]

In the following sections, we will discuss the general problem of dimension reduction and then formulate the band selection problem and motivate the idea of approximating principal components as a way to select bands. We then present a summary of the results of this research project. The reader is referred to the publications for more details.

4.4.1 Band Subset Selection and Optimal Dimension Reduction

From a statistical modeling perspective, as the number of bands increases, the discrimination capacity of hyperspectral data should increase as well. However, the need for training samples for a classifier can increase exponentially with the number of bands depending on the classifier being used [Chan and Hansen]. This is the so-called curse of dimensionality. Therefore, it is of interest to develop methodologies to reduce the dimensionality of the hyperspectral image data but at the same time retain as much as possible of their class discriminatory information.

Dimension reduction is the general problem of projecting a data set into a lower dimensional space. Let $\mathbf{x}$ be an n-dimensional random vector (a pixel) with zero mean and covariance matrix Σ_X. We wish to consider a linear dimension-reducing transformation of $\mathbf{x}$ to a random variable $\mathbf{y}$ given by

$$\mathbf{y} = \mathbf{A}^T\mathbf{x}$$

where $\mathbf{A}$ is a $n \times p$ matrix with $p<n$ and $\mathbf{A}^T\mathbf{A}=\mathbf{I}_p$ (orthonormal). Thus $\mathbf{y}$ is a p-dimensional random variable. From a classification perspective, we would like to select the $\mathbf{A}$ matrix such that in the lower dimensional space the classes are as separated as possible leading to improve performance of the training algorithms and the classifier.

The selection of the appropriate projection matrix $\mathbf{A}$ can be formulated as a constrained optimization problem

$$\max_{\mathbf{A}^T\mathbf{A}=\mathbf{I}} \Phi(\mathbf{A},\mathbf{X}) \tag{15}$$

where $\mathbf{X}$ is the image and Φ is a properly selected objective function. The selection of the objective function is a key step in the design of the dimension reduction algorithm. It is desired to select the cost function Φ that will result in a projection with large statistical separability and small class variability. Another alternative to select Φ is to keep as much as possible of the spatial information available on the image. For this later case, principal components are the optimal solution.

A dimension reduction problem of particular interest for applications where hyperspectral sensors can be used is band subset selection. The band subset selection problem can be framed in framework discussed previously by further restricting the projection matrix $\mathbf{A}$ as follows

$$\mathbf{A} = \mathbf{P}\begin{bmatrix}\mathbf{I}_p \\ \mathbf{0}\end{bmatrix}$$

where $\mathbf{P}$ is a permutation matrix. The net effect of this constraint is that the dimension-reducing transformation $\mathbf{A}$ now selects a subset of p of the original variables $\mathbf{x}$ as follows

$$\mathbf{y} = \mathbf{A}^T\mathbf{x} = \begin{bmatrix}\mathbf{I}_p & \mathbf{0}\end{bmatrix}\mathbf{P}^T\mathbf{x} = \begin{bmatrix}x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_p}\end{bmatrix}$$

The selection of a subset of bands have several interesting advantages compared to more general projections since we are retaining the physical meaning of the data (no combinations) in order to (a) maximize human understanding, (b) combine spectral data with other data types, and (c) exploit physical modeling/simulation. Also this information can be used in optimizing sensor design.

The selection of the p-optimal bands leads to a combinatorial optimization problem with a very large dimension solution space. For instance, selecting 10 out of 224 bands (as in AVIRIS) results in searching approximately 7.148×10^{16} possibilities. This problem can be tackled using standard search mechanisms for combinatorial optimization problems that are quite time consuming. The researchers have proposed an automated band selection algorithm [Vélez-Reyes and Jiménez,

Settle 98][Velez-Reyes et al, San Diego 97][Velez-Reyes et al, Orlando 2000] that is based on algorithms to select the most independent columns of a matrix. The determination of the most independent columns of a matrix is still a combinatorial optimization problem. However, good approximated solutions based on singular value decomposition (SVD) [Golub, G. and C.F. Van Loan] and rank revealing QR (RRQR) factorizations [Chan and Hansen] are presented in the linear algebra literature. It can be shown [Golub, G. and C.F. Van Loan][Chan and Hansen], that bands selected using this method, under certain circumstances, are a good approximation to the principal components and, therefore, will summarize most of the spatial variability information contained in the data cube [Vélez-Reyes and Jiménez, Settle 98][Velez-Reyes et al, San Diego 97]. The reduced computational time is achieved because these matrix decompositions that can be computed in polynomial time.

4.4.2 Principal Component Analysis of Hyperspectral Imagery

Let **X** be the hyperspectral image arranged in matrix form with singular value decomposition (SVD)

$$\mathbf{X}=\mathbf{U}\Sigma\,\mathbf{V}^T$$

where $\mathbf{U}=[\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n]$ and $\mathbf{V}=[\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n]$ are orthonormal matrices of the left and the right singular vectors respectively, $\Sigma=\text{diag}\{\sigma_1, \sigma_2, ..., \sigma_n\}$, and σ_i is the i-th singular value of **X**. The singular values are ordered according to magnitude $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n$. The matrix **X** is an $m \times n$ matrix where n is the number of bands and m is equal to the total number of pixels in the image. The columns of the matrix **X** are constructed by stacking the columns of the image at each channel.

The j-th principal component is given by

$$\mathbf{PC}_j=\sigma_j \mathbf{u}_j = \mathbf{X}\,\mathbf{v}_j, \quad j=1,2,...,n \tag{16}$$

The percentage of variability that is explained by each individual principal component is given by

$$\%Var_{j\text{-}th\ \text{component}} = \frac{\sigma_j^2}{\sum_{i=1}^{n}\sigma_i^2}$$

For hyperspectral data, due to redundancy, it is often the case that at one or many points in the singular value spectra that $\sigma_p >> \sigma_{p+1}$ for some $p<n$ which implies that the matrix **X** is near rank deficient. Therefore, most of the spatial information about the hyperspectral image is summarized in a p-dimensional subspace. It turns out that in most hyperspectral applications $p<<n$ resulting in a significant reduction of dimensionality. For later reference, we will denote by $\mathbf{U}_1$ the matrix formed by the first p left singular vectors, i.e. $\mathbf{U} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}$. The total percentage of variability explained by the first p principal components is given by

$$\%Var = \frac{\sum_{i=1}^{p}\sigma_i^2}{\sum_{i=1}^{n}\sigma_i^2}$$

Data reduction using principal components is achieved by transforming the original image into a set of orthogonal images images (or principal components), and keeping only the first p. A clear disadvantage of this approach as shown by (16) is the inherent transformation of the original hyperspectral image **X** (physical meaningful spectral data) into linear combinations of bands with, in many instances, little or no physical relation with the spectral information content on the original image.

4.4.3 Matrix Factorization Based Band Subset Selection

To simplify our presentation, we will reformulate the band subset selection problem as the problem of reordering the columns of the matrix **X** using a permutation matrix, **P**. Let $\overline{\mathbf{X}}$ be the matrix generated by multiplying **X** by **P**

$$\overline{\mathbf{X}} = \mathbf{XP} = \left[\underbrace{\overline{\mathbf{X}}_1}_{\substack{selected \\ bands}} \quad \overline{\mathbf{X}}_2 \right]$$

Notice that the only difference between $\overline{\mathbf{X}}$ and **X** is the ordering of their columns. The problem of band selection can be posed as finding a permutation matrix **P** such that the m×p matrix $\overline{\mathbf{X}}_1$ has the bands with the desired properties.

Notice that the singular value decompositions of **X** and $\overline{\mathbf{X}}$ are related by

$$\overline{\mathbf{X}} = \mathbf{XP} = \mathbf{U\Sigma V^T P} = \mathbf{U\Sigma}\left(\mathbf{P^T V}\right)^T = \mathbf{U\Sigma}\overline{\mathbf{V}}^T \tag{17}$$

so the permutation matrix only has an effect on the order of the right singular vectors and, hence, both matrices have the same principal components. A question of importance in this analysis is in what sense $\overline{\mathbf{X}}_1$ can be a good approximation to $\mathbf{U}_1$. In order to evaluate how close the selected bands are to the principal components, we decided to use the canonical correlation between the selected bands and the corresponding principal components. Canonical correlation is a multidimensional extension of the cosine of the angle between two vectors to the cosine of the angle between two subspaces [Wickens]. In our application, there are two spaces of interest, the p-dimensional space generated by the selected bands and the p-dimensional space generated by the first p principal components. The canonical correlation is the cosine of the smallest angle between vectors generated by all possible linear combinations of the selected bands and vectors generated by all possible linear combinations of the principal components. This angle in a sense measures the similarity between the two subspaces. We think this is the best measure of proximity between selected bands and principal components.

An important relation between this angle and the singular values of $\overline{\mathbf{X}}_1$ and **X** is presented next [Golub and Van Loan][Chan and Hansen].

$$\sin\theta\left(\Re(\mathbf{U}_1), \Re(\overline{\mathbf{X}}_1)\right) \le \frac{\sigma_{p+1}(\mathbf{X})}{\sigma_p(\overline{\mathbf{X}}_1)}$$

where $\Sigma(\mathbf{A})$ is the range space and $\sigma_i(\mathbf{A})$ is the i-th singular value of matrix **A**. Ideally, we would like that both spaces to be aligned which would imply that $\sin\theta=0$. A sufficient condition for this to happen would be that $\sigma_p(\overline{\mathbf{X}}_1) >> \sigma_{p+1}(\mathbf{X})$. It is shown in [Golub and Van Loan][Chan and Hansen] that this later condition will occur if a gap condition is met in the original HSI data (i.e. $\sigma_p(\mathbf{X}) >> \sigma_{p+1}(\mathbf{X})$) and a set of sufficiently uncorrelated bands is selected which will guarantee that $\sigma_p(\overline{\mathbf{X}}_1) \approx \sigma_p(\mathbf{X})$. This is the approach we followed in the band selection algorithms presented in [Vélez-Reyes and Jiménez, Seattle 98][Vélez-Reyes et al, San Diego 97][Velez-Reyes et al, Orlando 2000].

To try to meet the condition $\sigma_p(\overline{\mathbf{X}}_1) \approx \sigma_p(\mathbf{X})$, we can use the following result from [Golub and Van Loan] that the p-th singular value of **X** and of $\overline{\mathbf{X}}_1$ are related by

$$\sigma_p(\mathbf{X}) \geq \sigma_p(\overline{\mathbf{X}}_1) \geq \frac{\sigma_p(\mathbf{X})}{\sigma_{\max}(\overline{\mathbf{V}}_{11}^{-1})} \tag{18}$$

where $\overline{\mathbf{V}}_{11}$ is the upper pxp block of $\overline{\mathbf{V}}$. To obtain a sufficiently independent set of columns (bands), the permutation matrix **P** must be chosen such $\sigma_{\max}(\overline{\mathbf{V}}_{11}^{-1}) \approx 1$. The SVD and RRQR subset selection algorithms try to determine good choices (not necessarily optimal) for **P**.

4.4.3.1 SVD Subset Selection

Notice that

$$1 \geq \sigma_{\max}(\overline{\mathbf{V}}_{11}) \geq \sigma_{\min}(\overline{\mathbf{V}}_{11}) = \frac{1}{\sigma_{\max}(\overline{\mathbf{V}}_{11}^{-1})}$$

the leftmost inequality comes from the fact the $\mathbf{V}_1$ is an orthonormal matrix and the interlacing property of singular values [Golub and Van Loan]. If $\overline{\mathbf{V}}_{11}$ is well-conditioned (i.e. $\kappa(\overline{\mathbf{V}}_{11}) = \sigma_{\max}(\overline{\mathbf{V}}_{11}) / \sigma_{\min}(\overline{\mathbf{V}}_{11}) \approx 1$), then $\sigma_{\max}(\overline{\mathbf{V}}_{11}^{-1}) \approx 1$.

A heuristic solution [Golub and Van Loan] to this problem can be obtained by computing the QR with column-pivoting factorization of the matrix $\mathbf{V}_1$ that is composed by the first p-right singular vectors of **X**. To understand the reasoning behind this heuristic algorithm, let

$$\mathbf{V}_1^T \mathbf{P} = \begin{bmatrix} \overline{\mathbf{V}}_{11}^T & \overline{\mathbf{V}}_{12}^T \end{bmatrix} = \mathbf{Q}\begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \end{bmatrix}$$

where **Q** is orthonormal, **P** is the permutation matrix, and $\mathbf{R}_{11}$ is an upper triangular matrix, be the desired factorization. Notice that $\mathbf{R}_{11}$ is nonsingular and that $\overline{\mathbf{V}}_{11}^T = \mathbf{Q}\mathbf{R}_{11}$ and therefore $\kappa(\overline{\mathbf{V}}_{11}) = \kappa(\mathbf{R}_{11})$. Heuristically, [Golub and Van Loan] column pivoting tends to produce a well-conditioned $\mathbf{R}_{11}$ (i.e. $\kappa(\mathbf{R}_{11}) \approx 1$) and so the overall process tends to produce $\sigma_{\max}(\overline{\mathbf{V}}_{11}^{-1}) \approx 1$. Thus we obtain the SVD band selection algorithm based on the SVD subset selection algorithm in [Golub and Van Loan] and is summarized next.

Algorithm: SVD Subset Selection

1. Construct a matrix representation **X** of the hyperspectral image. Each column of **X** is constructed by stacking the columns of each single-band image.
2. Subtract the mean from each band.
3. Compute the SVD of **X**.
4. Compute the QR factorization with pivoting of the matrix $\mathbf{V}_1^T$ where $\mathbf{V}_1$ is formed by the first p right singular vectors of **X**. Save the pivot matrix **P** that results from this factorization.
5. Compute $\overline{\mathbf{X}} = \mathbf{XP}$.
6. Take the first p columns of $\overline{\mathbf{X}}$ as the selected bands.

4.4.3.2 Subset Selection using the RRQR Factorization

As suggested in [Chan and Hansen], RRQR can be used to determine the most independent columns of the matrix **X**. The motivation to look into RRQR is because SVD subset selection algorithm requires the computation of the SVD of **X**, which is a computational intensive procedure when compared to the computation of RRQR factorization. For and m×n matrix, the number of flops in the computation of the SVD is in the order of mn^3 while for the RRQR factorization is in the order of mn^2 [Golub and Van Loan]. The reasoning behind this approach is described next.

Let

$$\mathbf{XP} = \begin{bmatrix} \overline{\mathbf{X}}_1 & \overline{\mathbf{X}}_2 \end{bmatrix} = \mathbf{Q} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{R}_{22} \end{bmatrix}$$

where $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$ are p×p and (n-p)×(n-p) upper triangular matrices respectively, be a QR with pivoting factorization of **X**. Notice that

$$\overline{\mathbf{X}}_1 = \mathbf{Q} \begin{bmatrix} \mathbf{R}_{11} \\ \mathbf{0} \end{bmatrix}$$

and therefore $\sigma_i(\overline{\mathbf{X}}_1) = \sigma_i(\mathbf{R}_{11})$. From (18), we get that

$$\sigma_p(\mathbf{R}_{11}) = \sigma_p(\overline{\mathbf{X}}_1) \geq \frac{\sigma_p(\mathbf{X})}{\sigma_{\max}(\overline{\mathbf{V}}_{11}^{-1})} \tag{19}$$

so that

$$\frac{\sigma_p(\mathbf{R}_{11})}{\sigma_p(\mathbf{X})} \geq \frac{1}{\sigma_{\max}(\overline{\mathbf{V}}_{11}^{-1})} \tag{20}$$

Therefore by selecting **P** such that $\sigma_p(R_{11}) \approx \sigma_p(X)$, we get a subset $\overline{X}_1$ with the desired linear independence properties. Trying to make $\sigma_p(R_{11}) \approx \sigma_p(X)$ is the objective in the pivoting strategies of the Type I algorithms for the computation of the RRQR factorization [Chandrasekaran and Ipsen]. The RRQR band selection algorithm presented next is based on the RRQR factorization presented in [Chan and Hansen][Chandrasekaran and Ipsen].

Algorithm: RRQR Band Selection

1) Construct a matrix representation **X** of the hyperspectral image. Each column of **X** is constructed by stacking the columns of each single-band image.
2) Subtract the mean from each band.
3) Compute the QR with pivoting factorization **XP=QR** of **X**.
4) Let $\mathbf{R}^{(0)}=\mathbf{R}$, and $\mathbf{P}^{(0)}=\mathbf{P}$
   a) For l=0 to p-1 do
      i) Compute the largest singular value and corresponding right singular vector $\mathbf{w}^{(l)}$ of $\mathbf{R}^{(l)}$.
      ii) Find the permutation matrix that permutes the largest element in absolute value of $\mathbf{w}^{(l)}$ to the top.
      iii) Apply this permutation to the columns of $\mathbf{R}^{(l)}$ and the last n-l columns of $\mathbf{P}^{(l)}$ to form $\mathbf{P}^{(l+1)}$. Retriangularized $\mathbf{R}^{(l)}$ without column pivoting. Set $\mathbf{R}^{(l+1)}$ equal to the matrix resulting from the elimination of the first column and first row of the retriangularized $\mathbf{R}^{(l)}$.
5) Compute $\overline{\mathbf{X}} = \mathbf{X}\mathbf{P}^{(p-1)}$.
6) Take the first p columns of $\overline{\mathbf{X}}$ as the selected bands.

The RRQR factorization is a refinement of the QR with pivoting factorization that tries to achieve $\sigma_p(R_{11}) \approx \sigma_p(X)$. Since $\mathbf{R}^{(l)}$ is an upper triangular matrix, its largest singular value and singular vector can be easily estimated therefore not requiring the computation of its full singular value decomposition at each step.

4.5 Lossless Compresion Method for hyperspectral data (Dr. Shawn Hunt)

AVIRIS images are hyperspectral images consisting of 224 bands in the visible through infrared wavelengths, 380 to 2500 nanometers. Each spectral band is contiguous and approximately 10 nanometers wide. Because of this, one band is generally highly correlated with its neighbors. The precision of each element depends on the amount of processing done. The raw data is 12 bits (since 1995, 10 bits before that), but the precision increases as the processing is done. Data that has been geometrically and radially corrected is distributed in 16 bit format.

The large number of bands in hyperspectral images means that they will be close spectrally, and thus have a large correlation. This same fact makes the selection of a subset of bands difficult because of the large number of combinations possible. For example, if we are to select two bands to predict the others we have 24,976 possible combinations. If we are to select three bands, the number of combinations becomes 1,848,224. An exhaustive search then becomes impractical for all but the

smallest problems. A lengthy search may be tolerated if it is only to be done once for a large number of images, or if the compression results are changed drastically. In the case presented here, however, at least six bands are neede to achieve compression comparable to using one contiguous band, as shown later. The $1.64 \cdot 10^{11}$ possible combinations make it impossible to do an exhaustive search, and another method must be found.

Which subset of the 224 bands should be used to predict the rest? Since our goal is to have simple codes and high compression, we want to know which elements should be used as input to the predictor in order for the prediction error to have a low first order entropy. A direct measure of this can be done using conditional entropies. For example, the conditional entropy of x given y, H(x|y), is the average number of bits needed to code x given that y is known, where

$$H(x \mid y) = -\sum_i \sum_k p(x_i, y_k) \log(p(x_i \mid y_k)). \tag{21}$$

Thus, if we were to choose between y and z for predicting x, we would choose the one that gives the lower conditional entropy.

For the band selection problem, say we want to know which set of M bands are optimal for predicting the others. This can be done as follows. The conditional entropy of each band given a set of M bands is calculated. This is then repeated for all possible combinations of M bands. The set of bands that give the lowest average entropy are optimal.

The problem with calculating the conditional entropies, as mentioned above, is time complexity. When using AVIRIS images, the large number of bands and high precision becomes significant. Going back to the two band case, 224 entropies are calculated for each of the 24,976 possible combinations for a total of 5.6 million entropies. To calculate these conditional entropies, conditional densities (or equivalently joint densities) are estimated. For the two band case, third order joint densities are needed. If the data is represented by 12 bits, it can take any one of 4096 possible values. Third order joint densities means $4096^3 = 6.9 \cdot 10^{10}$ probabilites must be estimated. Further agravating the problem is the large amount of data needed for good estimates.

A conditional entropy can be viewed as the entropy of the prediction error if an 'optimal' predictor were used. As calculating the optimal prediction bands using entropy is not tractable, and even if it was, there is no easy way of finding or implementing the 'optimal' predictor, two simplifications will be made. First, linear predictors will be used, and second, the mean squared error will be used instead of entropy as a performance measure. In other words, minimize the expected value of the error (MSE) of the linear predictor:

$$e_i = b_i - y_i, \tag{22}$$

where e_i is the prediction error, b_i is the spectral value of band i, and y_i is the output of the predictor. The linear predictor is selected to minimize

$$E\{b_i - y_i \mid S\}, \tag{23}$$

where S is the set of bands used for prediction. Of course, since the purpose is to have the minimum entropy, the results will be compared to using entropies when possible.

4.6 Decision Fusion For Hyperspectral Data (Dr. Luis O. Jimenez)

Decision fusion is one of the most widely used procedures of data fusion. Different classification methodologies are used and a local decision is performed at each one of them. These decisions are combined in a Decision Fusion Center [Klein]. The Decision Fusion Center has a set of algorithms to integrate the individual and local decisions of each sensor. The algorithms are based on different techniques such as Boolean logic, voting schemes, Fuzzy logic, statistical approaches and Neural Network. One of the most common approaches used for decision fusion is the Parallel Fusion Network [Iyengar et al]. The classification systems observe common patterns. These systems do not communicate with each other but they provide complementary information about the pattern to be classified. A set of classification system is complementary if each one observes some information that the others do not. From an image processing perspective, each pixel is a pattern that will be discriminated.

4.6.1 Unsupervised Decision Fusion for Hyperspectral Data

Most of the high dimensional vector space of hyperspectral data is empty. Over fitting of the data occurs due to the lack of enough data to estimate well the parameters in a high dimensional space. In order to eliminate some of the redundancy, band subset selection algorithms are applied. By taking different subsets of bands of a hyperspectral image, applying the clustering algorithm to each one and fusing the results we might be able to use the data in all its spectral range without losing valuable information, and at the same time dealing with the curse of dimensionality. We exploited the fact that hyperspectral sensors produce redundant information by grouping bands in subsets to produce local classifications at every group. A criterion to group the bands can be to choose the most independent subsets. Figure 2 shows that scheme of local classifications at every subset.

The number of bands in the local classifications should be high enough to have good classification. At the same time the number of bands, that is the dimensionality of the local classification, should be made adequate enough to increase the ratio of the number of unlabeled samples per number of bands. That ratio will produce better estimation of the parameters required for first and second order statistics. Those estimations will exploit the amount of information hyperspectral data have, avoiding some of the problems of high dimensionality previously outlined.

Let $\mathbf{X}$ be the vector that contains all the output of the hyperspectral sensor at every band. It has the measurement of all d bands in the sensor data. The vector $\mathbf{X}$ is subdivided in sub-vectors $\mathbf{x}_i$ that contain the measurements of the i^{th} subset of bands that will be used in local classifications. Its structure is as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_N \end{bmatrix}^T, \text{ where } \mathbf{x}_i = \begin{bmatrix} x_{i,1} & \cdots & x_{i,d_i} \end{bmatrix}^T \tag{24}$$

where there are N groups of bands. The i^{th} group of bands is constructed following the band subset selection algorithm explained before. d_i is the number of bands in the i^{th} group, with the property that $\sum_{i=1}^{N} d_i = d' \leq d$. Where d is the total number of bands. We can now feasibly search subsets of bands that are mostly independent and will provide acceptable local unsupervised classification.

These subsets of bands are fed to a clustering algorithm. The results of each one of them are classification maps, u_i that are fused in a classification center, as shown in Figure 2.

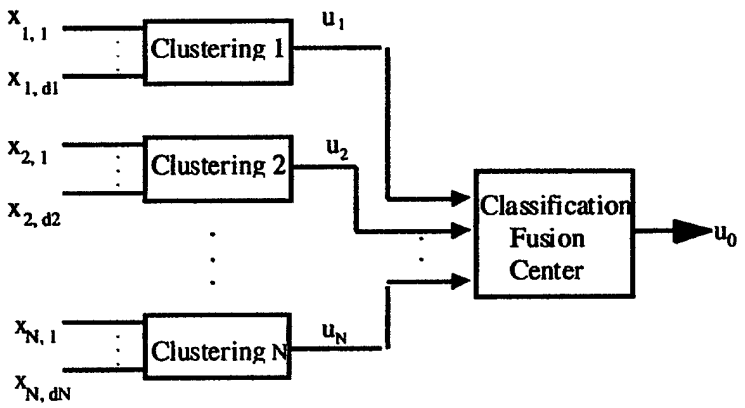

Figure 2. Unsupervised classification and decision fusion scheme for hyperspectral data.

How are the subsets of bands chosen? First we choose the d_1 most independent according to the algorithm of band subset selection. This will construct the subset 1 of d_1 bands. We applied a clustering algorithm on it that will result in a classification u_1. Then we applied the band subset selection algorithm again in the remaining $d-d_1$ bands and choose the d_2 mostly independent bands for the second subset. u_2 is the result of applying the unsupervised clustering algorithm to that particular subset. We continue subsequently until we have the appropriate number N of groups of subsets of bands. u_i could be seen as the classification result of the pixel **X** in the i^{th} subset of bands.

Several ways of using data fusion in hyperspectral unsupervised classification were studied in this research project. These results helped us in our understanding of hyperspectral imaging.

4.6.2 Classification Fusion Center

All the decision fusion mechanisms discussed in this section are based on a pixel-by-pixel basis. One of the most known fusion techniques is majority voting. This scheme fuses classification based on pixel classification with most occurrences. That is, the class with major incidence will decide the cluster of the pixel being analyzed. Other decision fusion is weighted majority voting. This method uses the same principle of majority voting but adds weights to the fused pixels. These weights are based on the degree of independency of the band subsets. As mentioned before each subset of bands selected has a degree of independence. The first subset correspond to the d_1 bands more independents. The second subset corresponds to the subset that has the second largest degree of independence. The idea is to bias decision to the clustering results of the most independent subset of bands, since its classifications are based on more information content. Squared weights are also applied to bias even more the decision fusion towards the most independent bands classifications.

Other methods tested are based on pixel probability density functions (pdf) estimations, taking into consideration the estimation of the density distribution. The first pdf method tested was maximum pdf fusion. The fused image is obtained by getting the maximum pdf, pixel by pixel, for each classification, and then classifying to the pixel with greater pdf value. The second pdf method is achieved by calculating pdf mean per pixel, and classifying each pixel to the corresponding classification of the pdf mean value. Assuming f(X/wk) is the probability density function of pixel **X** given class k, and $[x_1, x_2, x_3,\ldots,x_N]$ is the pixel **X** at subsets i to N, the decision fusion rules for M classes can be summarized in the following way [Kittler][Jimenez, Morales, Creus]:

4.6.2.1. Majority voting,

Let

$$C_{ij} = \begin{cases} 1, & f(\mathbf{x}_i/w_j) > f(\mathbf{x}_i/w_k), \quad \forall_i \\ 0, & \text{else} \end{cases} \tag{25}$$

Then assign pixel **X** to the k^{th} class if,

$$\sum_{i=1}^{N} C_{ik} \geq \sum_{i=1}^{N} C_{ij}, \quad \forall j, \tag{26}$$

4.6.2.2. Max pdf,

Assign pixel **X** to the k^{th} class if,

$$\max_{i=1}^{N}\left[f(\mathbf{x}_i/w_k)\right] = \max_{i=1}^{N}\left\{\max_{j=1}^{M}\left[f(\mathbf{x}_i/w_j)\right]\right\}, \tag{27}$$

4.6.2.3. Max average,

Assign pixel **X** to the k^{th} class if,

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i/w_k) = \max_{i=1}^{M}\left\{\frac{1}{N}\sum_{i=1}^{N} f(x_i/w_k)\right\}, \tag{28}$$

4.6.2.4. Linearly weighted majority voting,

Let

$$C_{ij} = \begin{cases} (N-i+1), & f(x_i/w_j) > f(x_i/w_k), \quad \forall_i \\ 0, & \text{else} \end{cases} \tag{29}$$

Assign pixel **X** to the k^{th} class if,

$$\sum_{i=1}^{N} C_{ik} \geq \sum_{i=1}^{N} C_{ij}, \quad \forall j, \tag{30}$$

4.6.2.5. Quadratic weighted majority voting,

Let

$$C_{ij} = \begin{cases} (N-i+1)^2, & f(x_i/w_j) > f(x_i/w_k), \quad \forall_i \\ 0, & \text{else} \end{cases} \tag{31}$$

Assign pixel **X** to the k^{th} class if,

$$\sum_{i=1}^{N} C_{ik} \geq \sum_{i=1}^{N} C_{ij}, \quad \forall j, \tag{32}$$

5. Summary of Most Important Results

This section presents the most important results of the research accomplished in this project. The first section is related with clustering results using Fuzzy C-Means Clustering, Neural Network Clustering and Fuzzy Neural Network Clustering methods applied to hyperspectral data. Section 2.2 has results of applying the regularization method to enhance the covariance estimation. It shows the effect of regularization in the detection of outliers. Section 2.3 presents results of the Band Subset Selection mechanism. Section 2.4 presents the results of the compression algorithm and finally the Decision Fusion mechanism results are shown in section 2.5.

5.1 Clustering Results

This section shows the results of our earliest experiments of the effect of dimensionality on the results of Fuzzy C Means clustering, Neural Network clustering and Fuzzy Neural Network clustering. The algorithms were applied on a segment of an AVIRIS frame with 220 bands. Figure 3 shows the result of the Fuzzy C means in high dimensionality. Figure 4 and 5 show the results for NN Clustering and Fuzzy Neural Network clustering respectively under the same conditions.

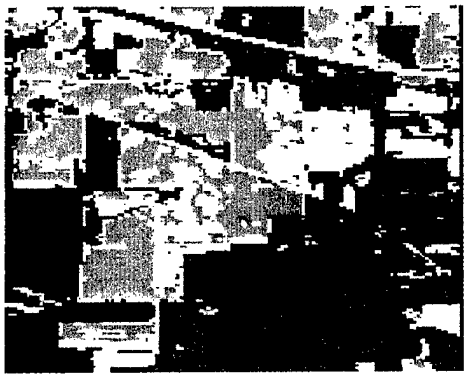

Figure 3. Fuzzy C-Means Clustering algorithm result.

Figure 4. Neural Network Clustering algorithm result

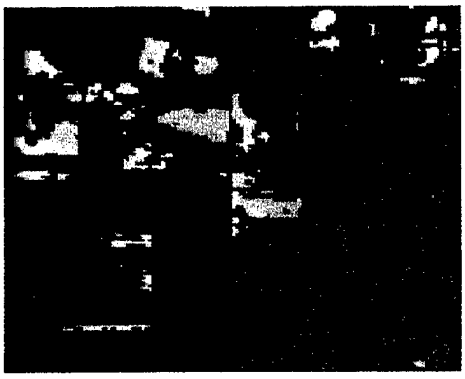

Figure 5. Fuzzy-Neural Network Clustering algorithm result.

According to these preliminar results we could observe that Fuzzy C-means is a more stable algorithm under high dimensionality than the other two. It is able to detect more patterns that are more accord to the a priori information of a testbed.

5.2 Use of regularization for parameter estimation process

This section presents the result of using a regularized covariance matrix estimation versus using the maximum Likelihood Estimate with a relative small number of training samples. Two hundred of training samples and two hundred of testing samples from the same soil class were used. They were obtained from a segment of a frame from the AVIRIS sensor with 220 bands. Figure 6a shows the number of outliers in the set of 200 training samples (in green) and the number of outliers in the set of 200 testing samples (in red) when using the Maximum Likelihood estimate of the covariance matrix in equation (13). It is shown that the number of outliers increases with the increment of the number of bands. Figure 6b shows the number of

outliers when the regularize covariance matrix was used for a small γ (= .02). Obviously the number of outliers in the test samples with the increment of bands were reduce using a regularization method.

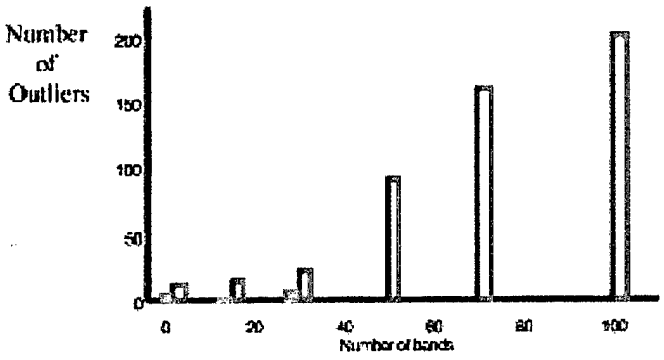

Figure 6a. Number of outliers versus number of bands using Σ^{MLE}

Figure 6b. Number of outliers versus number of bands using Σ^{REG} with γ = .02

5.3 Band Subset Selection Results

To evaluate the performance of the heuristic algorithms in approximating the optimal solution for the subset selection problem, several numerical experiments were conducted with data from the LANDSAT TM simulator and data from the

AVRIRIS sensor from the NW Indian pine test site. LANDSAT data was used since with only 7 bands the subset selection problem can be solved exactly for all possible combinations and subset sizes. With the AVIRIS data set, we only show results for subsets of 1 and 2 bands. All computations were done using the MATLAB™ environment.

5.3.1 LANDSAT TM Simulator Example

The band selection methods were used for band selection using a multispectral image taken over the town of Añasco in Puerto Rico by the LANDSAT TM simulator. Since it has only 7 bands, all combinatorial problems associated with band selection can be solved exactly and allow us to better analyze the results from our band selection algorithms.

5.3.1.1 Approximating Principal Components

The first comparison is to evaluate how well the selected bands approximate the corresponding principal components. Table 1 shows the singular values for the LANDSAT Image and the individual and cumulative percentage of variability. Notice that with three PCs we have over 99% of the total variability. A large gap can be identified between the second and third singular values. Table 2 shows the bands most aligned with the principal components and the corresponding canonical correlation. The canonical correlation is equal to the cosine of the angle between the subspaces. Tables 3 and 4 show the bands selected using SVD and the RRQR algorithms and the corresponding correlation. Notice that the bands selected using the SVD approach are always more correlated to the PCs than those selected using the RRQR factorization. Notice that both algorithms did very well when selecting two bands due to the gap between the second and third singular values. We highlighted the three-band case because in the next section we study it for class discrimination.

Table 1: Summary of Singular Values and Corresponding Variability for LANDSAT Image.

PC_i	σ_i	% $Var_{i\text{-th}}$	% Var
1	18,981	60.69	60.69
2	14,792	36.86	97.55
3	3,176	1.70	99.25
4	1,876	0.55	99.80
5	837	0.12	99.92
6	517	0.05	99.96
7	469	0.04	100.00

Table 2: Canonical Correlation of the Optimal Bands with the Corresponding Principal Components for LANDSAT

p	Selected Bands	Canonical Correlation
1	{5}	0.8905
2	{3, 5}	0.9963
3	{1,4,5}	0.9147
4	{1,3,5,7}	0.9892
5	{1,2,4,6,7}	0.9902
6	{1,2,3,4,5,7}	0.8613

Table 3: Canonical Correlation between the Principal Components and the bands selected using SVD.

p	Selected Bands	Canonical Correlation
1	{6}	0.7863
2	{3,6}	0.9962
3	{1,4,6}	0.9136
4	{1,3,6,7}	0.9890
5	{1,2,4,6,7}	0.9902
6	{1,2,3,4,6,7}	0.7961

Table 4: Canonical Correlation between the Principal Components and the bands selected using RRQR.

p	Selected Bands	Canonical Correlation
1	{6}	0.7863
2	{3,6}	0.9962
3	{1,3,6}	0.8672
4	{1,3,6,7}	0.9890
5	{1,2,3,6,7}	0.9314
6	{1,2,3,4,6,7}	0.7961

5.3.1.2 Class Discrimination Performance

Our interest in looking at band selection based on matrix factorization was to use it as the feature reduction pre-processor in a classifier as shown in Figure 7. The performance of the classifier depends on separability among classes so we would like to select those based that result in higher separability. In this section, we show how well our band selection algorithm did in selecting bands with good discrimination performance. In the experiment presented, we will select 3 bands, C means with covariance was used as the clustering method, the initialization method was based on eigenvalues and eigenvectors, and there were a total of 8 clusters identified. Once the data was grouped in eight clusters, we proceed to select the three bands that gave the best class separability in the reduced dimension space. Four criterion were used to evaluate class separability: Bhattacharyya (B), Jeffreys-Matusita (JM), Divergence (D), and Transformed Divergence (TD) distances [Chandrasekaran and Ipsen].

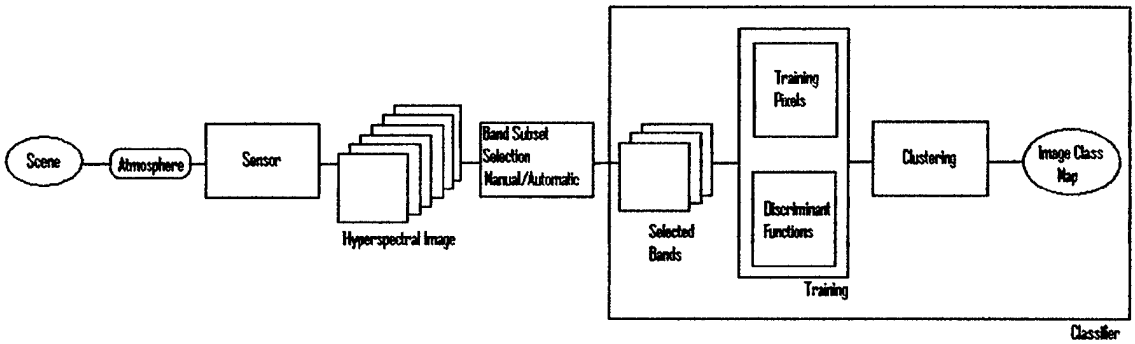

Figure 7. Band Selection as the Feature Extraction Process in Classification

- Bhattacharyya Distance

$$B_{ij} = \frac{1}{8}(M_i - M_j)^T \left\{\frac{\Sigma_i + \Sigma_j}{2}\right\}^{-1} (M_i - M_j) + \frac{1}{2}\ln\left\{\frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}}\right\}$$

- Jeffreys-Matusita Distance

$$JM_{ij} = 2\left(1 - e^{-B_{ij}}\right)$$

where B_{ij} is the Bhattacharyya Distance

- Divergence

$$D_{ij} = \frac{1}{2} Tr\left\{(\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1})\right\} + \frac{1}{2} Tr\left\{(\Sigma_i^{-1} - \Sigma_j^{-1})(M_i - M_j)(M_i - M_j)^T\right\}$$

- Transformed Divergence

$$TD_{ij} = 2\left(1 - e^{-\frac{D_{ij}}{8}}\right)$$

where D_{ij} is the divergence.

Since evaluation of these distances required evaluating the distance among several classes, we show results for selecting bands to maximize average distance and maximize the minimum distance. The results are tabulated in Table 5. For the 3-band case, the combination selected using SVD was {1,4,6} while using RRQR factorization was {1,3,6}. The bands selected using SVD were the optimal for maximizing the minimum distance while in the average case it was in the top 5 combinations for most costs. The bands selected using RRQR was number 6 in maximizing the minimum distance but not in the top 10 for the average case. This result points to the better performance of the SVD over the RRQR algorithm and to the fact that it also gives bands with good discrimination capacity.

Table 5: Ranking of Solutions Obtained using SVD and RRQR Algorithms (Ranking = 1 is Optimal Solution).

Distance	Ranking of Bands Selected using SVD		Ranking of Bands Selected using RRQR	
	Minimum distance	Average Distance	Minimum distance	Average Distance
Bhattacharyya	1	5	6	21
Jeffreys-Matusita	1	5	6	21
Divergence	1	19	6	26
Transformed Divergence	1	3	6	17

5.3.2 AVIRIS Example

The subset selection method was applied to the analysis of the AVIRIS data set from the Northwest Indian Pine Test site in Indiana. Only computations to compare with principal components were done here. Since the AVIRIS sensor has 220 bands solving the combinatorial problems that arise in determining the optimal discriminating bands require use of search methods and we are currently working on their implementation to use in analysis of this property in hyperspectral data. To give you an idea of the difficulties that arise in solving the exact problem for the AVIRIS case, if 2 bands are selected there are 24,090

possible combinations of 220 bands, if 3 are selected there are 1,750,540 possible combinations, and for 4 bands there are 94,966,795 possible combinations.

Figure 8 shows the variances of the first 100 principal components. Only in the first 10 principal components one can identify gaps. This will prove key in the performance of the RRQR algorithm as we shall see later. The canonical correlation of the selected bands with the principal components for bands selected using SVD and RRQR is shown in Figure 9. Notice that after 9 bands, the RRQR correlation behave very erratically and mostly low correlation. The SVD approach shows a more stable behavior although degraded due to the presence of no gaps after the 10^{th} singular value. The canonical correlation for the SVD bands is always higher that those for RRQR as seen in the LANDSAT case.

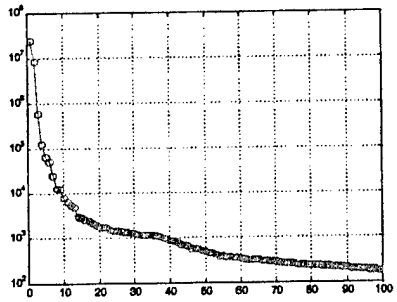

Figure 8. First 100 Eigenvalues of the Covariance Matrix for the AVIRIS Indian Pine Test Image.

Figure 10 shows the number of flops taken by SVD vs. RRQR. As expected RRQR took fewer flops. In the SVD approach the bulk of the computation goes into computing the singular value decomposition of the image. The QR factorization of the $\mathbf{V}_1$ matrix of the first p right singular vectors is negligible. The RRQR is computed sequentially so It is worthy to point out that SVD is implemented internally in MATLAB and optimized by MATLAB staff. We used a RRQR routine not well optimized and still were able to get better performance.

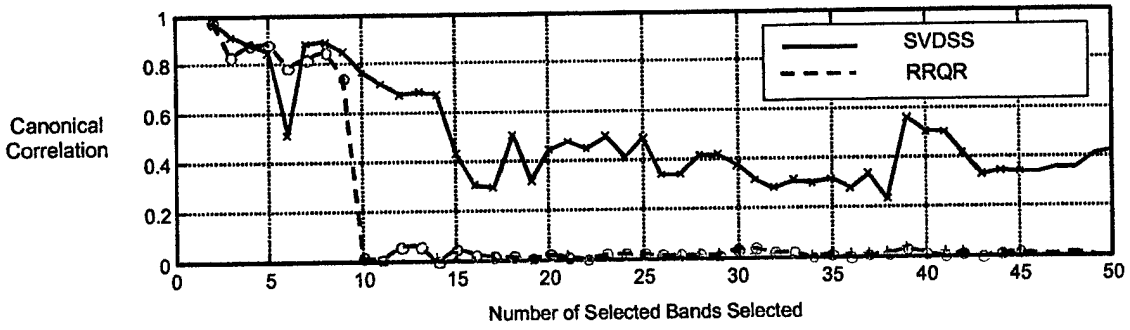

Figure 9. Canonical correlation analysis of the selected bands: (a) canonical correlation for SVD and RRQR, (b) Singular Values.

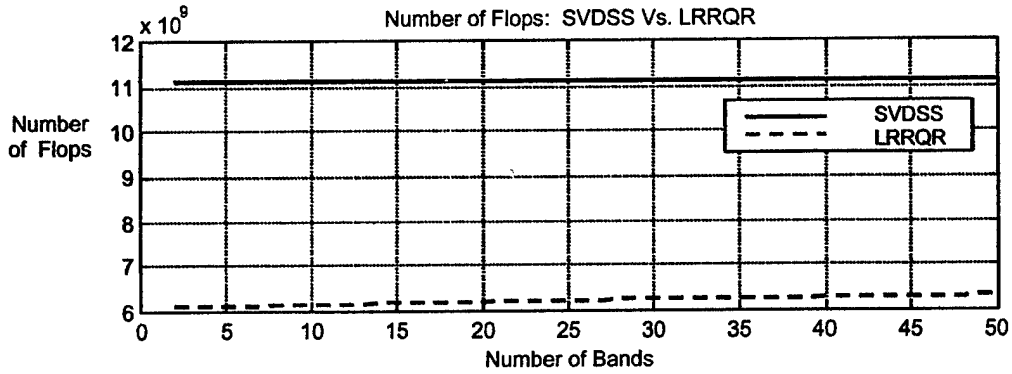

Figure 10. Number of flops taken by RRQR and SVD band selection algorithms.

5.3.3 Final Comments on Comparison of SVD and RRQR Band Selection Algorithms

In these sections, we presented a theoretical basis for hyperspectral image band selection using SVD and RRQR factorization matrix decompositions and compared their performance in their capacity to approximate principal components and select bands with good class discrimination properties. Numerical experiments using LANDSAT and AVIRIS images illustrate the good performance of the proposed approach for band selection and its closeness to the optimal solution. Band selection is inherently a combinatorial optimization problem and the proposed band selection methods compute a reasonable solution as the experiment showed in polynomial time.

5.3.4 Comparison to Other Approaches

In this section, we study how well our band selection algorithm based on SVD subset selection compares to other approaches proposed in the literature for band selection in hyperspectral imagery based on principal component analysis. We briefly introduce the approaches and present experimental results comparing all approaches using AVIRIS data.

5.3.4.1 Sensitivity-Based Band Selection

The first two approaches studied take advantage of the fact that the absolute sensitivity of the variance of the i-th principal component to the variance of the j-th band is given by [Golub and Van Loan]

$$\frac{\partial \lambda_i}{\partial \operatorname{var}(x_j)} = v_{i,j}^2$$

where λ_i is the variance of the i-th principal component and v_{ij} is the j-th entry of the i-th right singular (or loading) vector. The two algorithms differ in the way this sensitivity information is used. The first algorithm select those bands for which the variances of the first p-principal components are most sensitive. The second algorithm rejects those bands for which the variance of the smallest n-p principal components are most sensitive. The two algorithms are summarized next.

Algorithm: Forward Sensitivity Subset Selection(FS^3)

1. Construct a matrix representation **X** of the hyperspectral image. Each column of **X** is constructed by stacking the columns of each single-band image.

2. Subtract the mean from each band.
3. Compute the SVD of **X**.
4. Set p = number of bands to select and K=ϕ
5. for i = 1 to p,
    i. Select band j_i if
        1. $j_i = \arg \max_{j \in \{1,2,\ldots,n\}-K} |v_{i,j}|$
        2. K=K+{j_i}
    ii. where v_i is the i-th loading vector for **X**.
6. Set $\overline{\mathbf{X}}_i = \left[\mathbf{x}_{j_1} \mid \mathbf{x}_{j_2} \cdots \mathbf{x}_{j_p}\right]$ where $j_i \in K$.

Algorithm: Backward Sensitivity Subset Selection(BS^3)

1. Construct a matrix representation **X** of the hyperspectral image. Each column of **X** is constructed by stacking the columns of each single-band image.
2. Subtract the mean from each band.
3. Compute the SVD of **X**.
4. Set p = number of bands to select and K={1,2,…,n}
5. for i = p+1 to n,

    Reject band j_i if

    $$j_i = \arg \max_{j \in K} |v_{(n+p+1-i),j}|$$

    K=K-{j_i}

    where v_i is the i-th loading vector for **X**.
6. Set $\overline{\mathbf{X}}_i = \left[\mathbf{x}_{j_1} \mid \mathbf{x}_{j_2} \cdots \mathbf{x}_{j_p}\right]$ where $j_i \in K$.

Algorithms FS^3 and BS^3 correspond to algorithms B4 and B2 described in [Jollife][King and Jackson] respectively. Algorithm BS^3 was applied to band subset selection in [Pásztor et al][Csillag et al].

5.3.4.2 Principal-Component-Regression Band Subset Selection

This approach is based on linear regression of the HSI bands on the principal components. This approach in presented as algorithm B3 in [Jollife][King and Jackson].

Algorithm: PC Regression Subset Selection($PCRS^2$)

1. Construct a matrix representation **X** of the hyperspectral image. Each column of **X** is constructed by stacking the columns of each single-band image.
2. Subtract the mean from each band.
3. Compute the SVD of **X**.

4. for i = 1 to n

$$r_i = \left\| \left(\mathbf{I} - \mathbf{U}_1 \mathbf{U}_1^{\mathrm{T}} \right) \mathbf{x}_i \right\|_2$$

5. Set p = number of bands to select and K=ϕ
6. for i = 1 to p,

   Select band j_i if

$$j_i = \arg \max_{j \in \{1,2,\ldots,n\} - K} r_j$$

   K=K+{j_i}

7. Set $\overline{\mathbf{X}}_1 = \left[\mathbf{x}_{j_1} \mid \mathbf{x}_{j_2} \cdots \mathbf{x}_{j_p} \right]$ where $j_i \in$ K.

This algorithm is kind of a greedy strategy for band selection. As we shall see later its main defect is that it does not take into consideration the interdependencies of the HSI bands. This is because two bands with small sums of squares r_i maybe highly dependent but based on this strategy both bands could be selected by the algorithm.

5.3.4.3 Experimental Results

To compare the performance of the algorithms in approximating the principal components several numerical experiments were conducted using an AVIRIS data set of the Cuprite Mining District, Nevada taken in 1995. This image has the following characteristics: 400 samples, 350 lines, and 50 bands in the spectral range of 1.99080 to 2.47900 μm (bands: 172 to 221). Although the AVIRIS sensor provides information in 220 bands, Cuprite 95 is a subset of 50 bands. A view of this image is shown in Figure 10. The selected spectral range is useful for mineral discrimination. Here, the size of the image was manageable and allowed us to find the subset of bands with the smallest canonical correlation with the principal components for comparison purposes.

Figure 12 shows the value of the singular values for the CUPRITE image. We can see that only in the first 7 principal components one can identify gaps. Also 97% of the total variability is summarized by the first 10 principal components. Because of the image small size, we were capable of computing the optimal solution to the band subset selection with the largest canonical correlation for subsets up to 6 bands and this is shown in Table 6. The canonical correlations of the selected bands with the principal components for bands selected using FS3, BS3, PCRS2, and SVDSS are shown in Tables 7 to 10 respectively.

Figure 11. 1995 AVIRIS Image from CUPRITE Mining District in Nevada.

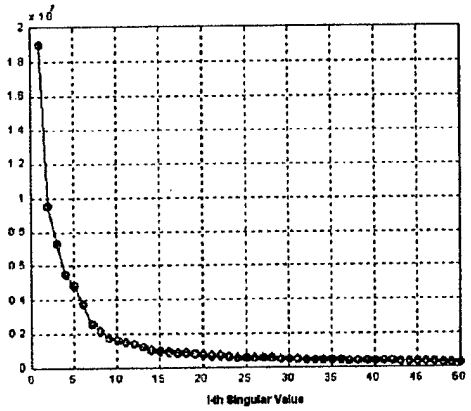

Figure 12. Singular Values Spectrum for the CUPRITE 95 AVIRIS Image.

Table 6: Canonical Correlation of the Optimal Bands.

p	Selected Bands	Canonical Correlation
1	{24}	0.933
2	{1, 50}	0.9223
3	{9, 42, 50}	0.8101
4	{1, 8, 25, 50}	0.9402
5	{1, 9, 25, 36, 49}	0.8905
6	{1, 9, 24, 36, 42, 50}	0.9313

Table 7: Canonical Correlation of the Bands Selected using FS^3.

p	Selected Bands	Canonical Correlation
1	{50}	0.4615
2	{46, 50}	0.0426
3	{46, 49, 50}	0.2500
4	{46, 48, 49, 50}	0.0604
5	{18, 46, 48, 49, 50}	0.0817
6	{18, 19, 46, 48, 49, 50}	0.0471

Table 8: Canonical Correlation of the Bands Selected using BS^3.

p	Selected Bands	Canonical Correlation
1	{20}	0.8818
2	{20, 27}	0.2518
3	{20, 25, 27}	0.0721
4	{20, 25, 27, 36}	0.0390
5	{20, 25, 27, 36, 50}	0.0680
6	{19, 20, 25, 27, 36, 50}	0.0106

Table 9: Canonical Correlation of the Bands Selected using $PCRS^2$.

p	Selected Bands	Canonical Correlation
1	{24}	0.9330
2	{24, 25}	0.2832
3	{24, 33, 34}	0.0653
4	{24, 33, 34, 35}	0.0106
5	{26, 31, 32, 33, 50}	0.0208
6	{18, 19, 20, 32, 33,50}	0.0090

Table 10: Canonical Correlation of the Bands Selected using SVDSS.

p	Selected Bands	Canonical Correlation
1	{50}	0.4615
2	{6, 50}	0.8934
3	{5,48,50}	0.5875
4	{6,37,49,50}	0.7232
5	{9,37,45,49,50}	0.8501
6	{6,18,37,47,49,50}	0.7825

Figure 13 shows the canonical correlations for the selection of up to 20 band subsets for all four algorithms. Here we do not show the solution for the optimal case because its computation requires large computational requirements. We obtain similar results as we see in Tables 7-10 where in almost all cases the SVDSS algorithm selected the bands with the largest canonical correlation. Furthermore, notice that after 7 bands there are no gaps in the singular value spectrum but still the SVDSS is capable of obtaining bands with more than 60% canonical correlation for most cases. This robustness is further studied in [Velez-Reyes et al, Orlando 2000].

Algorithms FS^3, BS^3, and $PCRS^2$ fail in getting a reasonable approximation to the principal components. SVDSS gets the best approximations to the principal components and, as shown in [Linares 2001], the results shown here are within the top 3% of all possibilities. This can be explained by studying the construction of the algorithms. Algorithms FS^3, BS^3, and $PCRS^2$ do the band selection by looking at one possibility at the time. As we already discussed for $PCRS^2$ this approach ignores the interdependencies of the bands. The SVDSS approach is kind of a sensitivity-based approach to band selection since it uses the loading vectors, $\mathbf{V}_1$ associated with the first p principal components. Notice that the j-th row of $\mathbf{V}_1$ is the sensitivity of the variance of the first p principal components to the variance of the j-th band. When the QR factorization with pivoting is applied to $\mathbf{V}_1^T$, the selection of pivots [Golub and Van Loan] is such that it will select a group of highly independent rows with large norm. In other words, it tries to select bands that have a significant effect on all the principal components (rows of large magnitude) but their effects on the principal components are relatively independent (band interdependencies).

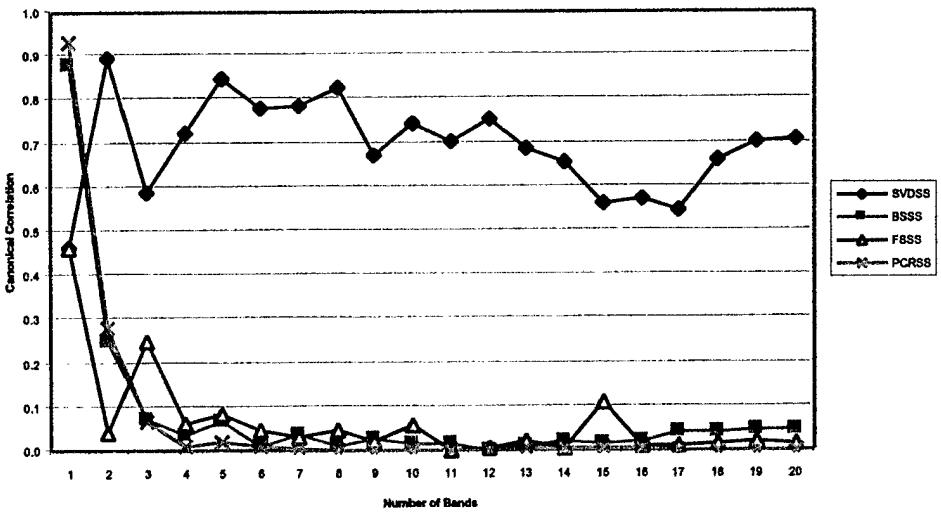

Figure 13: Canonical Correlation of Bands Selected using SVD and other PC-based Methods.

5.3.4.4 Comments on Comparison with other Approaches

In this section, we presented a comparison of three methods for band subset selection that use information from the HSI principal components presented in the literature and the SVD band selection method developed in this research. It was shown that successful methods for band subset selection using principal components should take into consideration the interdependencies across bands. Numerical experiments with the CUPRITE 1995 AVIRIS HSI are used to illustrate this point. The SVDSS gave the best performance of all algorithms. This algorithm also shows robust performance even in cases where the gap condition needed for high canonical correlation was not met.

5.4 Compression Method for hyperspectral data

The first part of the research was to investigate the correlation properties of hyperspectral data. Next, this information was used to develop lossless compression algorithms. The correlation properties of hyperspectral images depend on the sensor used. The research here used AVIRIS images. These images consist of 224 contiguous spectral bands, from 400nm to 2500nm, each about 10nm wide. The correlation, or more appropriately, the statistical relation between bands was studied in a number of ways. The correlation between the bands is shown in Figure 14. Since this information is going to be used for developing compression algorithms, it is more appropriate to look at the magnitude of the correlation. When used for prediction, a correlation of .7 and -,7 is the same. Figure 15 shows the same data as Figure 14, but with magnitude only. Some important results are clearly visible in the figures. The spectral bands near 1400nm and 1900nm go almost to zero because of atmospheric absorption. Because these bands have little or no ground data, they will have low correlation to other bands, and even between themselves. These bands, starting at about 110 and 150 in the figures, will be difficult to compress, because they are mostly noise, and are not useful in predicting other bands.

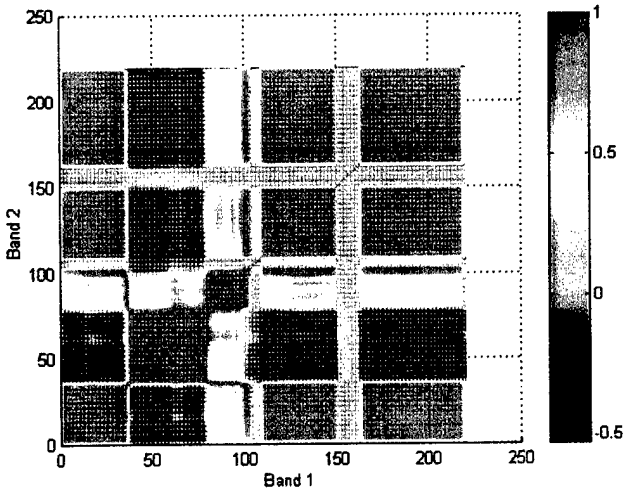

Figure 14. Correlation coefficients of AVIRIS data

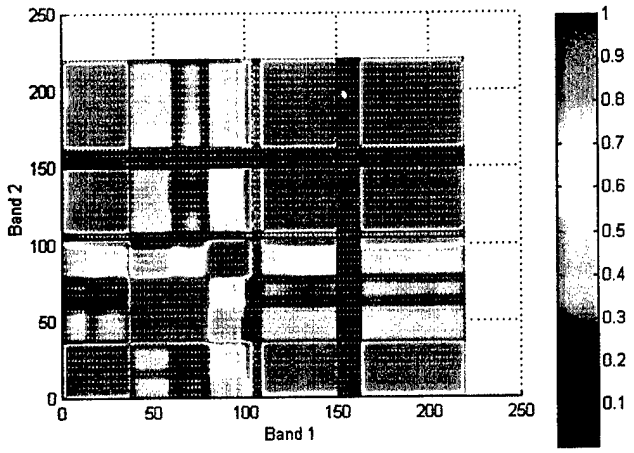

Figure 15. Magnitude of Correlation coefficients of AVIRIS data

On the other hand, we can see that there are sets of bands that have high correlation. The bands from 1 to 36, 37 to 80, 81 to 103, 109 to 149 and 164 to 219 all have high correlation between them. Also, bands from 1 to 36 have a relatively high correlation to the bands from 109 to 149 and 164 to 219. The bands from 37 to 80 have high correlation among them, but do not have a high correlation to any other bands. The same is true for the bands from 81 to 109. It would seem reasonable then, that if we were to select bands for predicting the others, there would be bands from the set 1-36, 37-80, and 81-109. As shown below, this is exactly what the band selection algorithm does.

The correlation of the bands is being investigated for use in lossless compression algorithms, where reduction of redundancy is the goal. In lossless compression work, some bands are used to predict others, get a prediction error, then code the error. In work done on reducing spectral redundancy presented in the literature[Wu and Memon], typically only one band is used to predict the adjacent band. To see the performance of this approach, the correlation of adjacent bands is shown in Figure 16. For comparison, the average correlation coefficient for each band with respect to the others is shown in Figure 17.

Correlation coefficients give a good estimate of how well bands can predict others, but what we really want to know is the information content of some bands given others. This will tell us how many bits are needed to code the data, and so how much compression can be obtained. This information content is given by the entropy, another statistical

measure, more amenable to compression work. The entropy of a band calculated with log base 2 is the information in bits per pixel. The conditional entropy is the information content of a band given that another band is known. In other words, the conditional entropy is the number of bits needed to code each prediction error if a perfect predictor were available. These results are shown in Figure 18. The x axis is the given band, and the y axis is the average entropy of all other bands.

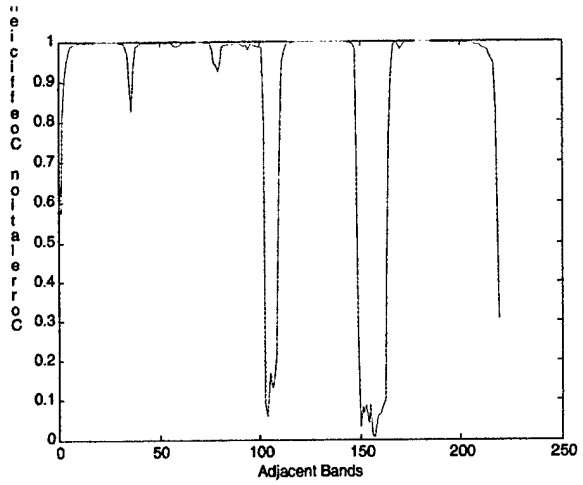

Figure 16. Correlation coefficients of adjacent bands in AVIRIS data

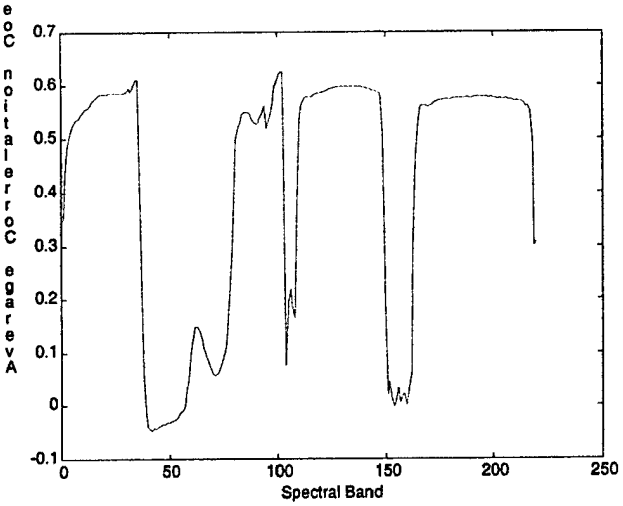

Figure 17. Average Correlation coefficient of each band in AVIRIS data

It can be seen that the bands close to 30 and 40 are the best bands to use for prediction. Specifically, bands 41, 42 and 29 have the first through third lowest average conditional entropies. If we were going to use one band for prediction, it would probably be band 41. Notice that we would probably choose a band in the set 1-36 if we made the choice based on correlation coefficients (see Figure 17).

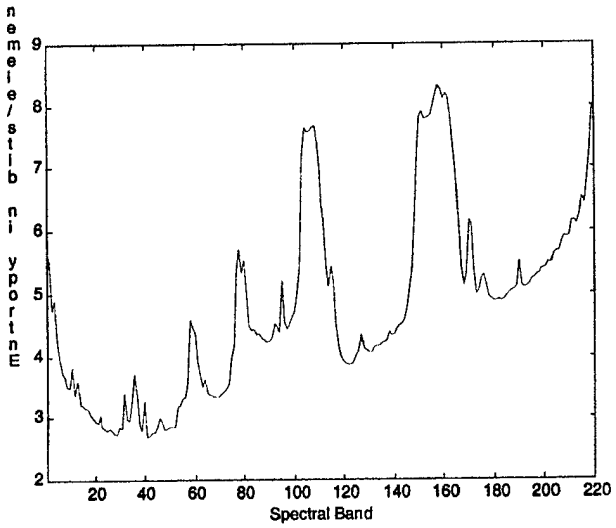

Figure 18. Average conditional entropy of AVIRIS data

It would be nice to have conditional entropies conditioned on more than one band, but the calculation of this is too computationally intensive. To calculate the conditional entropies in Figure 18, conditional probabilities were computed. Since AVIRIS data has 14 bit precision, each data can have 2^{14}=16,384 possibilities, and there are $16{,}384^2$=268,345,456 probabilities to be calculated. Conditional entropies conditioned on two bands would need $16{,}384^3$=4.398^{12} probabilities calculated. Since this is not possible, different simplifications were studied. The first simplification was to use a linear predictor to calculate the prediction error, then calculate the entropy of the error. This can be use for a limited number of bands, but when using more than two or three bands for the prediction, the number of combinations used becomes unwieldy. The number of combinations is:

One band : 224^2=50,176 combinations

Two bands : 5,299,800 combinations

Three bands : 385,118,800 combinations

Using the results from the previous year, we can base our compression work on look for combinations of bands that will be good spectral predictors, or bands that can predict well the remaining bands. The initial idea was to use nonlinear predictors for the data. This turned out to be a bad idea for two reasons. First, the nonlinear predictors are difficult to calculate, and lead to slow compression and decompression. Second, as shown in [Hunt], the nonlinear predictor converges to a linear predictor when predicting large amounts of data. In this case, there is little advantage to using the nonlinear predictors. The work then concentrated on band selection for compression using linear predictors. At present, algorithms in literature use one band for predicting the adjacent band. The compression is done sequentially, one band at a time. Because the bands must be available when decompressing, only bands that have been compressed can be used in the prediction. Using more than one sequential band for prediction does not increase the performance enough to offset the increase in complexity. This strategy achieves acceptable performance, but the compression and decompression are sequential. In many applications such as pattern recognition, only a subset of the total number of bands is used. This proves bad for the compression algorithms because if the last band compressed need to be decompressed, then all bands must be decompressed. The work here concentrated on finding subsets of bands that would equal the performance of using sequential bands. In this manner, no matter which band needs to be decompressed, only the subset of bands must be decompressed first [Hunt, and Vélez].

As mentioned above, finding a good subset of bands is a combinatorial problem. Instead of going through an exhaustive search, the research here used algorithms presented by Vélez [Vélez-Reyes and Jiménez], to quickly find a good subset. The results of this work is shown below.

Table 11: Fast band selection results.

Bands	lin. pred. entropy	SSE
29	7.830790	243,796,908,522
29-42	7.262668	31,454,577,590
29-42-89	6.776586	13,437,745,488
14-29-42-64	6.620856	8,834,423,644
9-29-36-42-66	6.543167	7,293,346,614
1-29-37-42-70-123	6.046576	6,382,762,255

Table 12: Principal Component results.

number of PC used	lin. pred. entropy	SSE
1	7.989636	105,173,135,736
2	7.109786	9,275,148,970
3	6.476988	3,252,692,941
4	6.300056	2,445,712,891
5	6.157913	1,860,244,668
6	5.919479	1,456,170,791

Table 11 shows the entropy error of using from one to six bands for prediction. Using six bands, an entropy error of 6.05 was achieved, equaling the performance of using one band to predict its adjacent band. Table 12 shows the error entropy and SSE using the principal components as predictors. This was done as a comparison, since the principal components give the smallest SSE from among all linear combinations of bands.

The analysis of these results based on the correlation work done previously is interesting. Above it was shown that the band with the lowest conditional entropy is band 41. The band selector here chose band 29. This is reasonable if we go back to the correlation coefficients. Bands from 1-36 have high correlation between them, and also a high correlation to the bands from 109-149 and 164-219. The bands from 37-80 have a high correlation among them, but have a low correlation to the other bands. Based on this, it is better to choose a band in the set 1-36. Also, when choosing two bands, the algorithm selects 29 and 42. The band 29 has high correlation with the bands mentioned above, and band 42 has a high correlation with bands in the set 37-80. Going back to Figure 15, we can see that the only set of bands, besides the noise bands, that do not have a high correlation with bands 29 and 42 are the bands from 81 to 103. In table 11 we can see that when three bands are selected, band 89 is selected along with bands 29 and 42.

5.5 Unsupervised Decision Fusion Results

This section shows the results obtained after applying the decision fusion algorithms. We group the bands in 10 subsets of 20 bands, and classify 10 clusters. The results were fused using the five mechanism previously discussed: majority voting, linear weighted voting, square weighted voting, maximum pdf, and max mean pdf. Figure 19 shows

a frame section of an AVIRIS image of Kennedy Space Flight Center in Florida. This segment shows two main areas from different settings. The region at the left is characterized for being and urban area and the one at the right is mainly a swamp with a river and a road. Figures 20 to 24 show the classification results using decision fusion. Figures 20, 21, and 22 show the voting schemes based results: majority voting, linear weighted voting and the square weighted voting schemes. They produce results that are similar. Figures 23 and 24 shows the methods based on probability density function: maximum pdf and max mean pdf. They are similar among themselves but they are different with respect to the voting schemes in the first group. The pdf methods are able to classify better the urban area in the region at the left of the image. The spatial structures are better defined in that classification maps. At the same time the pdf based fusion mechanism localized some urban structures in the swamp region at the right of the image. Further experiments need to be done in other hyperspectral images in order to detect bias in the decision fusion algorithms. This understanding will help us modify and enhance the algorithm.

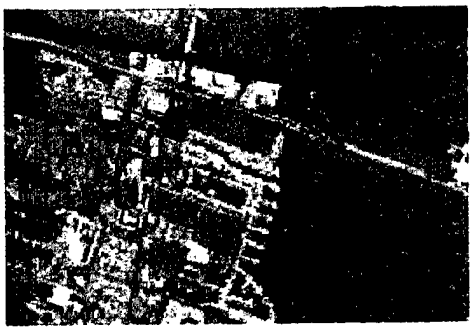

Figure 19. Segment of Original Image

Figure 20. Majority Voting

Figure 21. Linear Weighted Voting

Figure 22. Square Weighted Voting

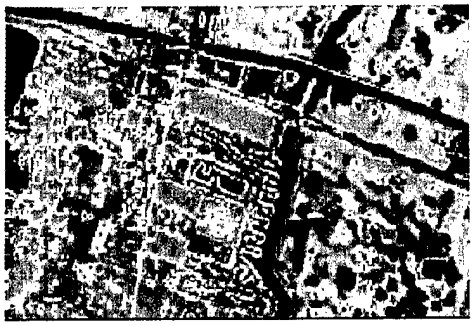

Figure 23. Maximum pdf

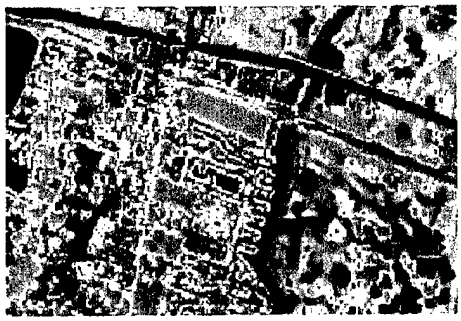

Figure 24. Max Mean pdf

6 Listing of all publications and technical reports supported under this grant

6.a Papers Published in peer review journals.

Jimenez, Luis O., Morales, Anibal and Creus, Antonio, "Classification of Hyperspectral Data Based on Feature and Decision Fusion Approaches using Projection Pursuit, Majority Voting and Neural Networks," IEEE Transactions on Geosciences and Remote Sensing, **Vol. 37, No. 3**, pp. 1360-1366, May 1999.

6.b Papers published in conference proceedings.

S.D. Hunt, "Nonlinear Predictors for Lossless Compression of AVHRR Imagery," Proceedings of the SPIE, vol.3717, pp. 198-203, April 1999.

S. D. Hunt and M. Velez-Reyes, "Band selection for lossless image compression." In S.S. Shen and M.R. Descour, editors, *Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VII*, Proceedings of SPIE Vol. 4381, April 2001.

L. O. Jiménez, M. Velez-Reyes, Y. Chaar, F. Fontan, and C. Santiago, "Partially Supervised Detection Using Band Subset Selection in Hyperspectral Data." In *Algorithms for Multispectral and Hyperspectral Imagery V*, Proceedings of SPIE, Vol. 3717, Orlando Florida, April 1999.

L.O. Jiménez-Rodríguez, M. Vélez-Reyes, J. Rivera-Medina, and H. Velásquez, "Unsupervised Decision Fusion for Hyperspectral Data." In *Image and Signal Processing for Remote Sensing VII*, Proceedings of SPIE, Vol. 4541, September 2001.

M. Vélez-Reyes and L. O. Jiménez, "Subset Selection Analysis for the Reduction of Hyperspectral Imagery." In *Proceedings IEEE International Geosciences and Remote Sensing Symposium*, July 6-10, Seattle, WA, 1998.

M. Vélez-Reyes, L. O. Jiménez, F. Pagán, and G. Fernández, "Subset Selection for the Analysis of Hyperspectral Data." In *Proceedings of the 1997 International Symposium on Spectral Sensing*. Held in San Diego CA, 1997.

M. Vélez-Reyes, L.O. Jiménez, D.M. Linares, and H.T. Velázquez, "Comparison of matrix factorization algorithms for band selection in hyperspectral imagery." In *Algorithms for Multispectral, Hyperspectral and Ultraspectral Imagery VI*, Proceedings of SPIE, Vol. 4049, Orlando Florida, 2000.

M. Velez-Reyes and D.M. Linares, "Comparison of Principal-Component-Based Band Selection Methods for Hyperspectral Imagery." In *Image and Signal Processing for Remote Sensing VII*, Proceedings of SPIE, Vol. 4541, September 2001.

7. **List of all participating scientific personnel showing any advanced degrees earned**

Name	Degree Earned	Expected Date of Graduation
Ileana Carrasquillo	BSEE	Dec 2001
Leila Rodriguez	BSEE	Dec 2001
Hector Velasquez	MSCE	May 2002
Jorge Rivera	MSCE	May 2002
Felix Fontan	MSEE	May 2002
Fernando Gallo	MSEE	May 2002
Emmanuel Arzuaga	MSCE	May 2002
Yahia Massalmah	MSEE	May 2003
Diego Rivera	MSEE	May 2003

8. Bibliography

Chan, T.F. and P.C. Hansen, "Some applications of the rank revealing QR factorization." In *SIAM J. Sci. Stat. Comput.*, Vol. 13, No. 3, pp. 727-741, 1992.

Chandrasekaran, S. and I. C.F. Ipsen, (1994) "On rank-revealing factorizations." In *SIAM J. Matrix Anal. Appl.*, Vol. 15, No. 2.

T. Chen, D. Staelin, and R. Arps, ``Information Content Analysis of Landsat Image Data for Compression," *IEEE Transactions on Geoscience and Remote Sensing*, vol.GE-25, no.4, pp.499-501, July 1987.

Csillag, F, L. Pásztor, and L.L. Biehl, "Spectral Band Selection for the Characterization of Salinity Status of Soils." In *Remote Sensing of the Environment*, Vol. 43, 1993.

Duda, R.O., Hart P.E., Stork, D.G., Pattern Classification, 2nd Edition, John Wiley & Sons, Inc., 2001, pp. 113-114.

Geladi, P. and H. Grahn, *Multivariate Image Analysis*. John Wiley & Sons, 1996.
Schowengerdt, R.A., *Remote Sensing: Models and Methods for Image Processing*, 2nd Edition. Academic Press, 1997.

Golub, G. and C.F. Van Loan, *Matrix Computations*, 3rd Edition, John Hopkins University Press, Baltimore, MD, 1997.

S.D. Hunt, "Nonlinear Predictors for Lossless Compression of AVHRR Imagery," Proceedings of the SPIE, vol.3717, pp. 198-203, April 1999.

S. D. Hunt and M. Velez-Reyes, "Band selection for lossless image compression." In S.S. Shen and M.R. Descour, editors, *Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VII*, Proceedings of SPIE Vol. 4381, April 2001.

S.S. Iyengar, L. Prasad, H. Min, *Advances in Distributed Sensor Technology*, New Jersey: Prentice Hall, 1995 , pp. 67-76.

Jimenez, L. O. and D. A. Landgrebe, "Supervised Classification in High Dimensional Space: Geometrical, Statistical, and Asymptotical Properties of Multivariate Data," *IEEE Transactions on System, Man, and Cybernetics*, Vol 28, No. 1, pp. 39-54, 1998.

L. O. Jiménez, M. Velez-Reyes, Y. Chaar, F. Fontan, and C. Santiago, "Partially Supervised Detection Using Band Subset Selection in Hyperspectral Data." In *Algorithms for Multispectral and Hyperspectral Imagery V*, Proceedings of SPIE, Vol. 3717, Orlando Florida, April 1999.

Jimenez, Luis O., Morales, Anibal and Creus, Antonio, "Classification of Hyperspectral Data Based on Feature and Decision Fusion Approaches using Projection Pursuit, Majority Voting and Neural Networks," IEEE Transactions on Geosciences and Remote Sensing, **Vol. 37, No. 3**, pp. 1360-1366, May 1999.

Jiménez, L. O. and D.A. Landgrebe, "Hyperspectral Data Analysis and Feature Reduction Via Projection Pursuit," In *IEEE Transactions on Geoscience and Remote Sensing*, 2000.

L.O. Jiménez-Rodríguez, M. Vélez-Reyes, J. Rivera-Medina, and H. Velásquez, "Unsupervised Decision Fusion for Hyperspectral Data." In *Image and Signal Processing for Remote Sensing VII*, Proceedings of SPIE, Vol. 4541, September 2001.

Jollife, I.T., "Discarding Variables in a Principal Component Analysis I: Artificial Data." In *Applied Statistics*, Vol. 21, 1972.

Kil, D. and F.B. Shin. *Pattern Recognition with Applications to Signal Characterization*. AIP Press, 1996.

King, J.R. and D.A. Jackson, "Variable Selection in Large Environmental Data Sets using Principal Component Analysis." In *Environmetrics*, Vol. 10, 1999.

J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On Combining Classifiers," IEEE Transactions on Pattern Analysis and Machine Intelligence, **Vol. 20, No. 3**, pp. 226-239, March, 1998.

L. A. Klein, *Sensor and Data Fusion Concepts and Applications*. Washington: SPIE Optical Engineering Press, 1993, pp. 31.

Linares, M., *Evaluation of Matrix Factorization Algorithms for Band Selection in Hyperspectral Imagery*, Master of Science Thesis, University of Puerto Rico Mayagüez Campus, 2001.

Pásztor, L. and F. Csillag, "Band selection procedure for reduction of high resolution spectra." In *Astronomical Data Analysis Software and Systems III*, ASP Conference Series, Vol. 61, 1994.

Price, J.C., "Band selection procedure for multispectral scanners." In *Applied Optics*, Vol. 33, No. 15, 1994.

E. Rodriguez-Diaz, M. Velez-Reyes, Univ. of Puerto Rico/Mayaguez, R. K. Chin, C. A. DiMarzio, Northeastern Univ., "Wavelength selection for imaging hemoglobin in skin." In *Subsurface Sensing Technologies and Applications II*, Proceedings of SPIE, Vol. 4129, San Diego California, July 2000.

San Miguel-Ayanz, J. and G.S. Biging, "An iterative classification approach for mapping natural resources from satellite imagery." In *Int. J. Remote Sensing*, Vol. 17, No. 5, 1996.

Van den Broek, W.H.A.M., D. Wienke, W.J. Melssen, and L.M.C. Buydens, "Optimal wavelength selection by a genetic algorithm for discrimination purposes in spectroscopy." In *Applied Spectroscopy*, Vol. 51, No. 8, 1997.

M. Vélez-Reyes and L. O. Jiménez, "Subset Selection Analysis for the Reduction of Hyperspectral Imagery." In *Proceedings IEEE International Geosciences and Remote Sensing Symposium*, July 6-10, Seattle, WA, 1998.

M. Vélez-Reyes, L. O. Jiménez, F. Pagán, and G. Fernández, "Subset Selection for the Analysis of Hyperspectral Data." In *Proceedings of the 1997 International Symposium on Spectral Sensing*. Held in San Diego CA, 1997.

M. Vélez-Reyes, L.O. Jiménez, D.M. Linares, and H.T. Velázquez, "Comparison of matrix factorization algorithms for band selection in hyperspectral imagery." In *Algorithms for Multispectral, Hyperspectral and Ultraspectral Imagery VI*, Proceedings of SPIE, Vol. 4049, Orlando Florida, 2000.

M. Velez-Reyes and D.M. Linares, "Comparison of Principal-Component-Based Band Selection Methods for Hyperspectral Imagery." In *Image and Signal Processing for Remote Sensing VII*, Proceedings of SPIE, Vol. 4541, September 2001.

Wickens, T.D, *The Geometry of Multivariable Statistics*, Lawrence Erlbaun Assoc, Inc, 1995.

X. Wu, and N. Memon, "Context-Based lossless interband compression- extending CALIC," *IEEE Transactions on Image Processing*, **Vol. 9** No. 6, June 2000.

Zongker, D. and A. Jain, "*Algorithms for Feature Selection*". Proceedings of ICPR'96. IEEE, 1996.