12-13-2003

Limitations of Principal Component Analysis for Dimensionality-Reduction for Classification of Hyperspectral Data

Anil Meerasa Cheriyadat

LIMITATIONS OF PRINCIPAL COMPONENT ANALYSIS FOR DIMENSIONALITY-REDUCTION FOR CLASSIFICATION OF HYPERPSECTRAL DATA

By

Anil Cheriyadat

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Electrical Engineering
in the Department of Electrical & Computer Engineering

Mississippi State, Mississippi

December 2003

LIMITATIONS OF PRINCIPAL COMPONENT ANALYSIS FOR

DIMENSIONALITY-REDUCTION FOR CLASSIFICATION

OF HYPERSPECTRAL DATA

By

Anil Cheriyadat

Approved:

Lori M. Bruce Associate Professor of Electrical and Computer Engineering (Director of Thesis)	Roger L. King Giles Distinguished Professor of Electrical and Computer Engineering
James E. Fowler Associate Professor of Electrical and Computer Engineering	Nicholas H. Younan Graduate Program Coordinator of Electrical and Computer Engineering

A. Wayne Bennett
Dean of Bagley College of Engineering

Name: Anil Cheriyadat

Date of Degree: December 13, 2003

Institution: Mississippi State University

Major Field: Electrical Engineering

Major Professor: Dr. Lori Mann Bruce

Title of Study: LIMITATIONS OF PRINCIPAL COMPONENT ANALYSIS FOR DIMENSIONALITY-REDUCTION FOR CLASSIFICATION OF HYPERSPECTRAL DATA

Pages in Study: 84

Candidate for Degree of Master of Science

It is a popular practice in the remote-sensing community to apply principal component analysis (PCA) on a higher-dimensional feature space to achieve dimensionality-reduction. Several factors that have led to the popularity of PCA include its simplicity, ease of use, availability as part of popular remote-sensing packages, and optimal nature in terms of mean square error. These advantages have prompted the remote-sensing research community to overlook many limitations of PCA when used as a dimensionality-reduction tool for classification and target-detection applications. This thesis addresses the limitations of PCA when used as a dimensionality-reduction technique for extracting discriminating features from hyperspectral data. Theoretical and experimental analyses are presented to demonstrate that PCA is not necessarily an appropriate feature-extraction method for high-dimensional data when the objective is

classification or target-recognition. The influence of certain data-distribution characteristics, such as within-class covariance, between-class covariance, and correlation on PCA transformation, is analyzed in this thesis.

The classification accuracies obtained using PCA features are compared to accuracies obtained using other feature-extraction methods like variants of Karhunen-Loève transform and greedy search algorithms on spectral and wavelet domains. Experimental analyses are conducted for both two-class and multi-class cases. The classification accuracies obtained from higher-order PCA components are compared to the classification accuracies of features extracted from different regions of the spectrum. The comparative study done on the classification accuracies that are obtained using above feature-extraction methods, ascertain that PCA may not be an appropriate tool for dimensionality-reduction of certain hyperspectral data-distributions, when the objective is classification or target-recognition.

DEDICATION

Mom & dad

ACKNOWLEDGMENTS

I sincerely express my gratitude to my advisor Dr. Lori Mann Bruce, for her constant guidance and support all through my research and academics at Mississippi State University. I am indebted to her for developing in me an interest and motivation to conduct productive research. I thank Dr. Bruce for her tireless efforts in reviewing my thesis. I would like to thank Dr. Roger King for teaching me the basics of remote-sensing. I also thank him for serving on my committee. I thank Dr. James Fowler, for serving on my committee. I would like to acknowledge Mr. Cody Gray for providing me the experimental data. I sincerely thank my family, without whose support and encouragement I could not have completed this work. Finally, I thank all my friends for their suggestions and for all those good moments.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

1.1 OVERVIEW

With the advent of high-resolution sensors, a large amount of information is now available for the remote-sensing and biometrics communities, which permits close analysis and understanding of the characteristics of the objects under investigation. The large amount of information collected by these sensors has also necessitated the development of efficient algorithms to reduce the data volume for the purpose of storage, transmission, and computationally efficient analysis. The Hyperion sensor mounted on the Earth observation satellite, the first hyperspectral sensor in space, is an example of advanced hyperspectral-sensor technology [1]. The high-resolution data offered by this sensor permits detailed land-cover classification and identification.

The complexities involved in processing these vast datasets in their raw form render any direct utilization of the data virtually impracticable. In the case of hyperspectral or ultraspectral sensors, the spectral reflectance from the target material is sampled at hundreds to thousands of contiguous spectral bands, respectively. In mathematical terms, each sampled spectral band is referred to as a dimension. Based on the reflectance measurements made by the sensor at each spectral band, the reflectance

data corresponding to the target material is allocated to a discrete point in a N-dimensional space, where N is the number of spectral bands. For example, in an Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), which captures the image in 224 contiguous spectral bands, each pixel, which represents a part or the whole of the target, forms a discrete point in a 224-dimensional space. As shown in Figure 1.1, the discrete point in a 224-dimensional feature space is represented by the vector whose elements are formed from the spectral bands measured at that pixel location.

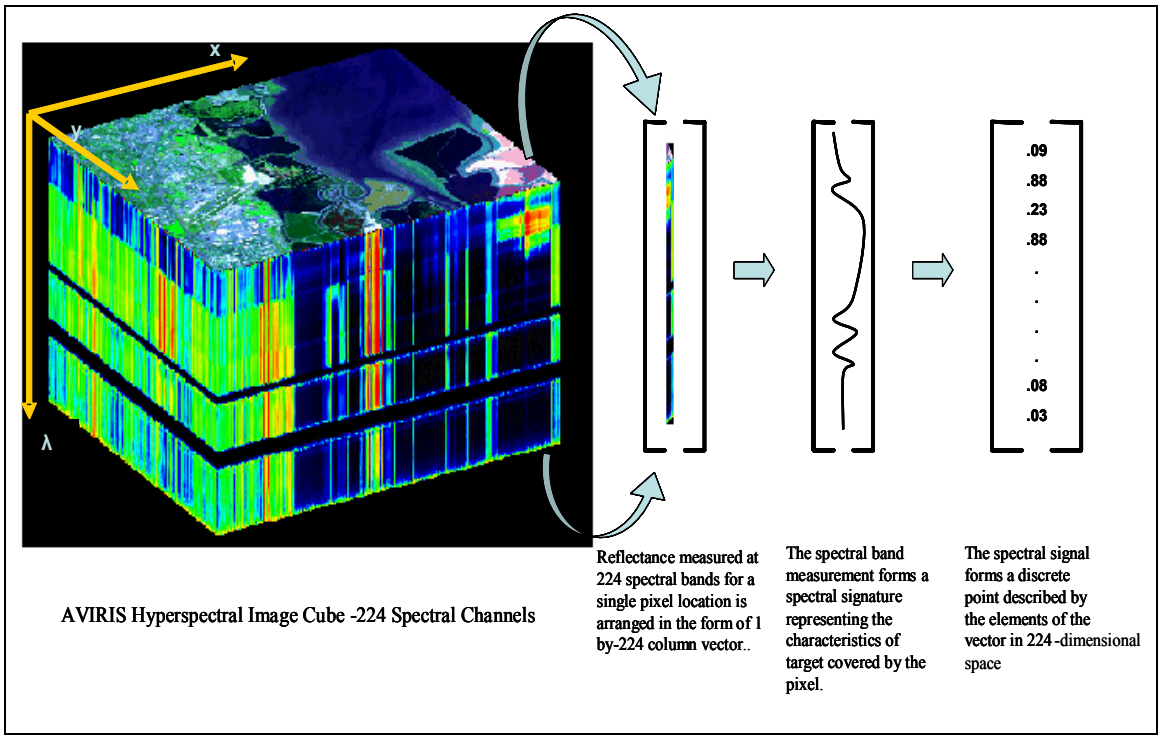

Figure 1.1 Vector Representation of Hyperspectral Image (AVIRIS Image cube courtesy [2]).

Similar issues pertaining to the high-dimensionality of data exist in other applications such as biometrics. For example, in the case of face recognition, a typical image of size 256 by 256 pixels corresponds to a discrete point in a 65,536-dimensional space, as shown in Figure 1.2. The unidentified image is recognized by measuring the

closeness of the corresponding discrete point with respect to a discrete point of the known target.

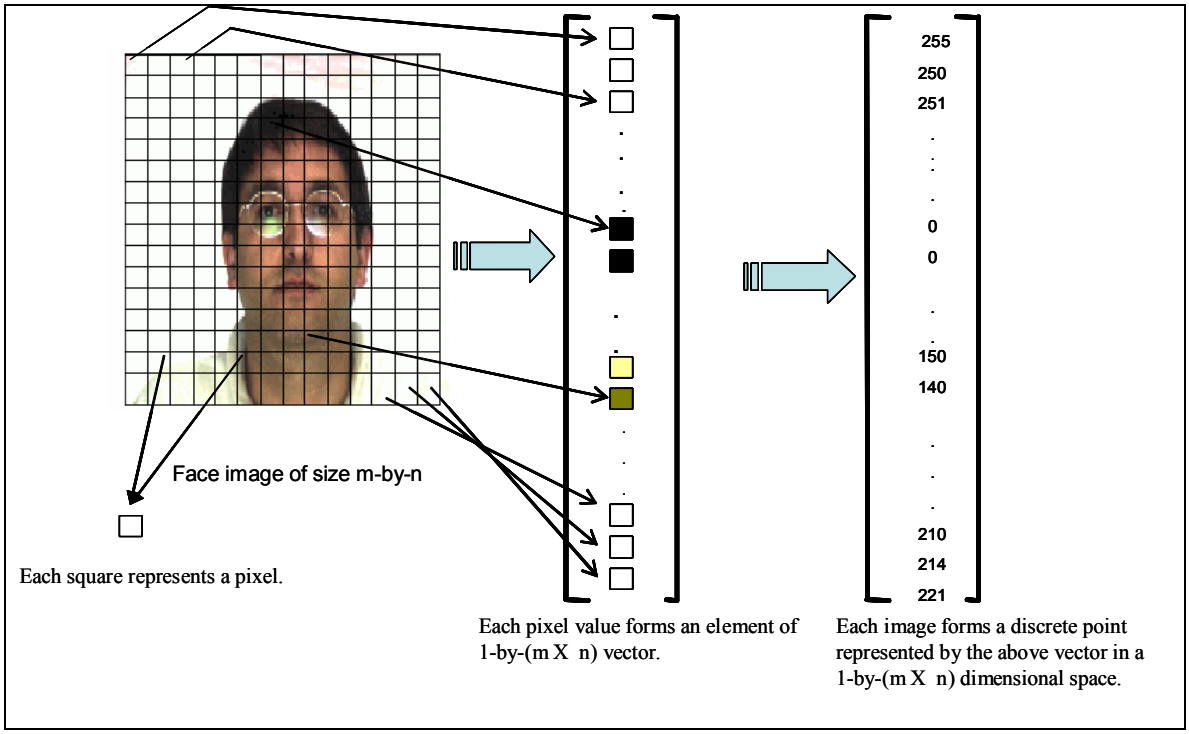

Figure 1.2 Vector Representation of a Face Image (Face image courtesy [3]).

The analysis of data in a hyper-dimensional space for the purpose of target identification or classification is computationally taxing. Although high-dimensional data increases the accuracy of target classification and identification, from a supervised-classification perspective, this accuracy is dependent on the number of training data available. Limited availability of training data, particularly in the case of remote-sensing applications, limits the precision with which object characteristics can be estimated. The limitations imposed by the training data are further exacerbated, as the dimensionality of the data increases.

To circumvent these issues, numerous data pre-processing methods have been developed in order to extract only the relevant information pertaining to the applications involved [4-17]. In the case of classification and target-recognition applications, relevant information pertaining to the discrimination or identification of the target is retained, while the necessary details are discarded. This process is referred to as dimensionality-reduction. The unwanted details are those dimensions that do not provide any discriminatory information. Previous studies by Landgrebe reveal that high-dimensional space is mostly empty [4]. Experiments conducted to study the behavior of multivariate data show that, as dimensionality increases, the multivariate data tend to move away from the origin. The implication of this characteristic on a hyperellipsoidal data-distribution is that the data tends to reside on the outer shell. As stated by Landgrebe "this [high-dimensional data characteristic] implies that the high-dimensional data set can be projected to a lower-dimensional subspace without losing significant information in terms of separability among statistically different classes" [4].

In recent years, researchers have proposed different algorithms [5 -14] to establish this lower-dimensional structure present in high-dimensional data without sacrificing significant information pertinent to the application. Some of the commonly used dimensionality-reduction methods for hyperspectral data are principal component analysis (PCA), greedy search [6], wavelet-based feature-extraction [7],[11], decision-boundary feature-extraction [8], feature-level fusion using best-bases selection algorithms [9],[10],[11], and decision-level fusion of features obtained using best-bases selection [12],[13],[14]. Of the many dimensionality-reduction methods, PCA is one of the most commonly implemented methods for hyperspectral data. Many variations of PCA, such

as multi-scale PCA [15], PCA in combination with LDA [16], and segmented PCA [17], are used for the purpose of dimensionality-reduction.

1.2 POPULARITY OF PCA

PCA has gained immense popularity in the remote-sensing community. Some of the most commonly used remote-sensing software packages for data analysis and interpretation such as ENVI [18] and ERDAS IMAGINE [19] use PCA for data analysis. The option for performing PCA can be found under the TRANSFORM menu in ENVI, while ERDAS IMAGINE use PCA as part of its spectral-enhancement options. Around 19 articles in *IEEE Transactions on Geoscience and Remote Sensing* during the years of 1993 to 2003 reported the use of PCA for various data analysis and interpretation purposes. For the last three years alone, at least 45 research papers related to PCA were presented at the IEEE Geoscience and Remote Sensing Symposium. Around 1194 IEEE journals and conference proceedings during the years of 1993 to 2003 reported research topics that used PCA for data analysis or classification. Some of the common work pertaining to PCA is briefly described here.

PCA is used for land-cover classification of EO-1 Hyperion data [20]. The dominant PCA bands are used as the input to a clustering algorithm. Preston *et al.* report the use of PCA for extraction of statistically reliable features from a 130-dimension data space for seabed classification using multibeam sonar images [21]. A comparative study on the classification performance of different texture features along with PCA features is reported in [22]. Hung *et al.* reports the use of divergence, a measurement of statistical separability, to identify potential features for classification from PCA space [23]. Bajic

reports the use of the largest three PCA components as a classification tool for thermal hyperspectral imagery [24].

Several factors that have led to the popularity of PCA include its simplicity, ease of use, availability as part of popular remote-sensing packages, and theoretical optimal in terms of mean square error when applied as a dimensionality-reduction tool for compression [25]. These advantages have prompted the research community to overlook many limitations of PCA when used as a dimensionality-reduction tool for classification and target-detection applications. The goal of an optimal feature-extraction method for classification and target-detection is not only to reduce the data dimensionality for reducing computational costs, but also to improve classification accuracy. The objective of this thesis is to expose the inability of PCA to extract discriminating features from certain data-distributions resulting in poor classification accuracies. Theoretical and experimental analyses are presented to demonstrate that PCA is not necessarily an appropriate feature-extraction method for high-dimensional data when the objective is classification or target-recognition.

The thesis is organized as follows. The need for dimensionality-reduction and a description of some of the popular dimensionality-reduction methods, including PCA, are detailed in Chapter 2. Chapter 3 presents a detailed theoretical assessment of PCA with respect to certain data-distributions to demonstrate limited ability of PCA to extract features pertinent to classification applications. PCA is applied to two different synthetic datasets, and the observations are analyzed in Chapter 3. Experimental analysis using actual hyperspectral data to corroborate the theoretical assessments is presented in Chapter 4. PCA is compared with other feature-extraction methods such as variations of

the Karhunen-Loève transform, and greedy-search algorithms. Finally, Chapter 5 summarizes the observations made in the previous chapters to establish the fact that PCA is not always suitable for dimensionality-reduction when the objective is classification or target-detection.

CHAPTER II

BACKGROUND

2.1 SUPERVISED IMAGE CLASSIFICATION

Supervised target-recognition systems require training data to design the feature-extraction and classification components as shown in Figure 2.1. One of the main problems in remote-sensing is that typically there are not enough training samples to exploit the entire dimensionality of the data. Hence, feature-extraction algorithms need to be applied to the high-dimensional data in order to reduce its dimension. The main objective of any feature-extraction algorithm for image classification is to reduce the dimensionality without losing significant information that can be utilized to detect a target. In a statistical supervised image-classification scheme, feature-extraction algorithms are developed by analyzing class distribution of the training data. Hence, an ideal choice for a training dataset would be one that covers all the classes involved and contains enough training samples to permit reliable estimations of the data-distribution parameters of the underlying classes. The extracted features are then used as inputs to the classification stage of the system to accurately identify the classes. The classifier is designed based on the reduced features extracted from the training data. In order to accurately and reliably estimate the class-distribution parameters using the training set, a

balance should be maintained between the reduced feature-dimension size and the training-set size.

Recently there have been studies conducted on applying classifiers to subsets of the high-dimensional data in order to avoid limitations associated with small training sets. The decisions made from multiple classifiers are fused to achieve higher classification accuracies [9],[12],[13],[14]. However, even in these applications, a feature-extraction or dimensionality-reduction stage is often included for each of the multiple classifiers.

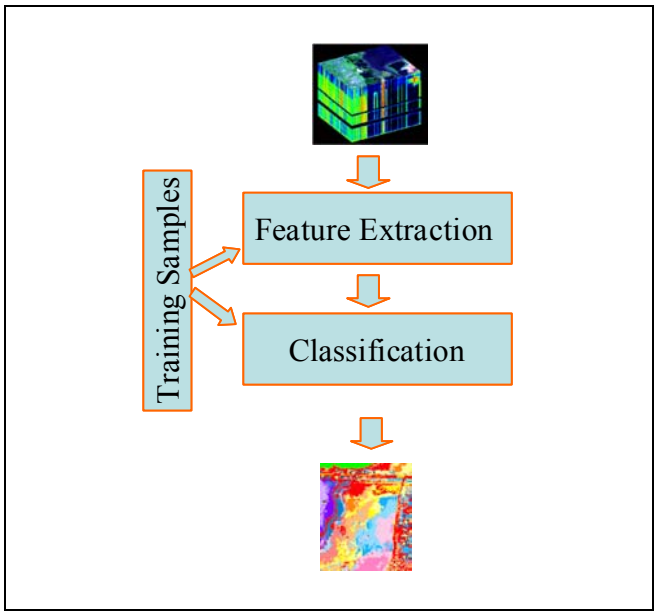

Figure 2.1 A Typical Supervised Classification Scheme.

2.2 DIMENSIONALITY-REDUCTION

The process of selecting or extracting features from high-dimensional data that can be used to discriminate the underlying classes, or identify the target from non-targets, plays a key role in the design of a classifier system. In pattern-recognition terms, this process is also referred to as dimensionality-reduction. The rationale behind performing a dimensionality-reduction may be to remove the redundant information present in the data,

to find out the underlying lower-dimensional structure of the data, or to reduce the computational complexity.

The dimensionality-reduction algorithms can be broadly classified into two categories based on the domain in which they are operating. In the first case, the relevant features are selected from the raw data in its original domain based on some discriminatory criterion. From a classification perspective, the objective of this criterion is to select those features that have a higher discrimination capability, or in other words to enhance the classification capability of the system. Greedy-search methods employed in [6] in the original spectral domain are examples of such an approach. In the second case, the features are transformed into a new domain where the features can be arranged in the order of their importance, which is application specific. For an image-compression application, the order in which features are arranged may be based on information content or entropy. For a classification application, the order of the features should be based on class-discrimination capability. The transformation can be supervised, as in Fisher's linear discriminant analysis (LDA), or unsupervised, as in PCA. Thus dimensionality-reduction is achieved in the transformed domain by retaining only those features pertinent to the application. Figure 2.2 below gives a pictorial comparison of two different dimensionality-reduction approaches.

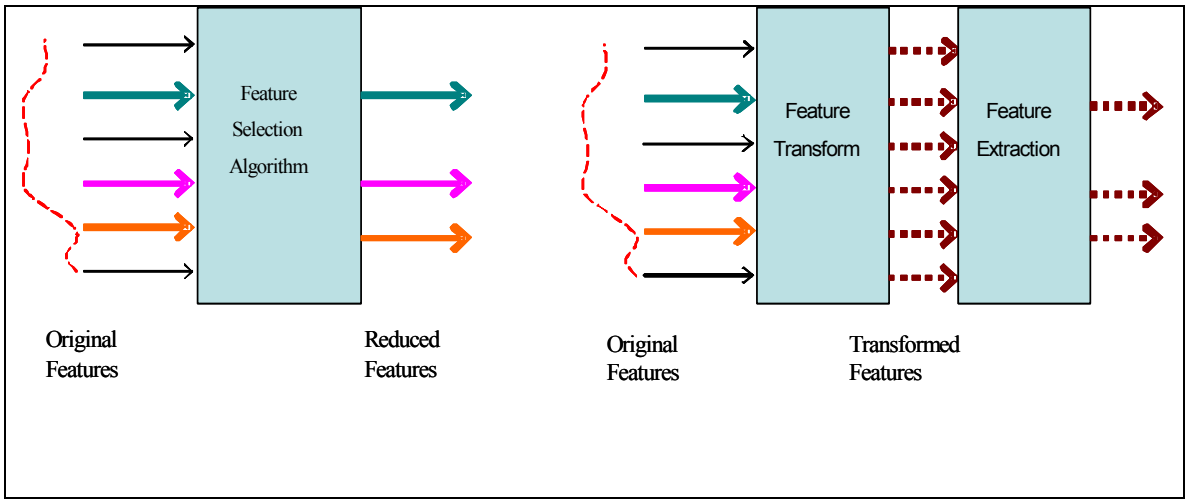

Figure 2.2 Two Approaches for Feature reduction.

2.3 NEED FOR DIMENSIONALITY-REDUCTION

In the case of hyperspectral signal processing or face-recognition tasks, data is distributed in a high-dimensional space, and each observation forms a discrete point in this high-dimensional space. The data points pertaining to a specific class of objects typically form a cluster, while unknown data points are identified and labeled based on their closeness to known data points. This process becomes computationally complex and statistically unreliable as the dimensionality of the data space increases.

In the case of hyperspectral sensors such as AVIRIS or HYPERION, the spectral-band resolutions are on the order of 200 bands. For a 10-bit reflectance data, the number of possible discrete locations in a 200-dimension feature space would be on the order of1024^{200}. Even in the case of face recognition tasks, a 256-by-256 size grayscale image having 8 bits/pixel represents a discrete point from 256^{65536} possible discrete values. Such a large dimension for the feature space poses several limitations. The number of discrete locations in the feature space is unreasonably large such that the computational

burden associated with the huge dimension makes any direct utilization of such information impractical.

Previous studies about the characteristics of high-dimensional space by Jimenez and Landgrebe have shown that higher-dimensional spaces have some interesting properties which are quite different from those most commonly found in two or three dimensional spaces [26]. One important characteristics of higher-dimensional space is reiterated here: "*As dimensionality increases the volume of a hyperellipsoid concentrates on its outer shell"*[26]. This characteristic property of higher-dimensional space implies that as dimensionality increases, data tend to move away from its origin. For higher-dimensional space, data tends to spread around the outer shell, and hence density estimation is difficult. This unique characteristic of higher-dimensional space is demonstrated in Figure 2.3 for multivariate data X with *n* dimensions,

$$X = \{x_i\}, \tag{1}$$

where x_i is a normal independent, identically distributed (*iid*) random variable and i varies from 1 to n. The multivariate data distribution has zero mean and unit variance. It can be observed from Figure 2.3 that as dimensionality increases, the data-distribution tends to move away from the origin. The mean distance from the origin increases as dimensionality increases.

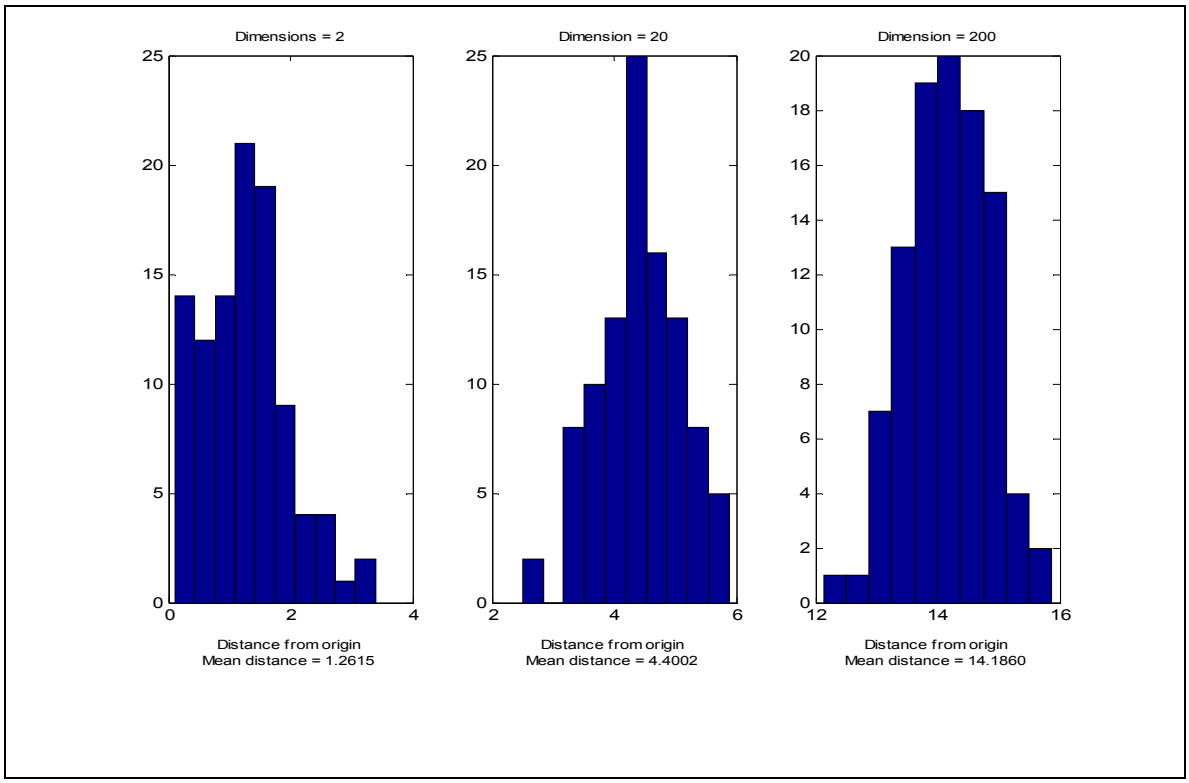

Figure 2.3 Histogram Showing Distances for Normal Multivariate Data from the Origin for Different Dimensions.

In the design of a supervised classification system, generally the statistical class-distribution parameters such as mean and covariance, are estimated from the training samples and substituted for true distribution parameters. The mean parameter relates to the position at which the cluster is found in the feature space while the covariance parameter conveys the information related to the spread or shape of the cluster. In order to estimate the true class-distribution parameters for high-dimensional data, an inordinate number of training samples is needed. That is, to have a statistical estimation close to the true parameters, the class-distribution parameters should be estimated using a large number of training samples. One of the main difficulties in remote-sensing or biometrics is the limitation of the training data. As the dimensionality of the feature space increases, the amount of training data needed for a reliable estimation of class-distribution statistics

increases. Studies done by Hughes on the relationship between sample size, number of dimensions, and classification accuracies show that, for a given training-set size, the classification accuracy peaks for a certain number of features, after which the accuracy degrades as more features are used [27]. This effect is referred as the Hughes phenomena. It is observed from Figure 2.4 that the classification accuracy can be increased by increasing the dimension of the feature space, provided that the class distribution parameters are estimated using an infinite number of training samples.

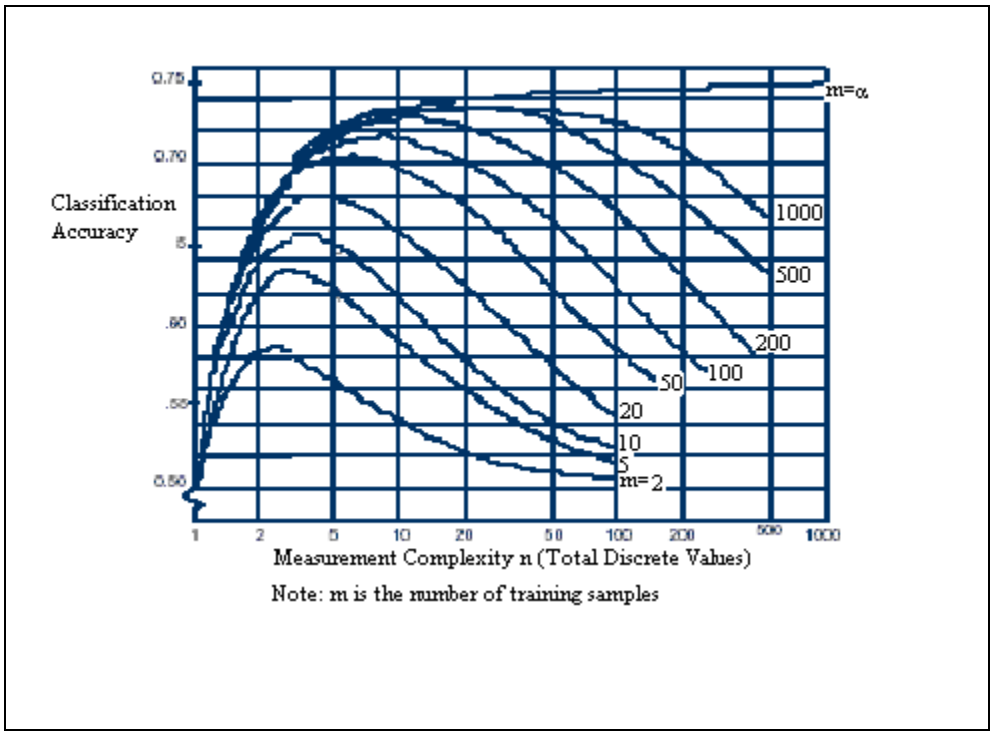

Figure 2.4 Hughes Phenomenon, where *m* is the number of training samples [27].

For example, from Figure 2.4, it is clear that for a sample size of 200 (m=200), the classification accuracy is highest when the measurement complexity is around 8-10, and thereafter, the addition of more features results in lower classification accuracies. This corroborates the fact that, even though the larger dimensions of the spectral

signature make it possible to discriminate the target with better accuracy, the amount of training data needed grows rapidly in proportion to the dimension [4]. Particularly in a remote-sensing scenario, the added cost of overcoming the limitations involved in gathering large amounts of training data may not justify the higher classification accuracies. In fact, it may not be physically possible to obtain enough training data.

The use of smaller training-sample sizes for larger-dimensional data will result in unreliable estimations of data-distribution statistics. This is more prominent with second-order statistical moments like covariance. The unreliable estimation of the class-distribution parameters results in a poor design for the classifier. This degradation in classification performance is more prominent as the features are increased for a given sample size.

In short, the cost and limitations associated with obtaining large training sets in a remote-sensing scenario or face-recognition task, the computational complexities imposed by large dimensions, and the unreliable estimation of class-distribution statistics using smaller training sizes have led to research in dimensionality-reduction for classification. The study on characteristics of higher-dimensional spaces has shown that multivariate data-distribution tends to lie in a lower-dimension [4]. For a classification application, dimensionality-reduction techniques try to establish this lower-dimensional subspace, which retains much of the class-discrimination capabilities. Hence, data dimensionality-reduction techniques to extract features that can improve the statistical separation between the underlying classes play a significant role in classification and target-detection applications.

Although there exist numerous dimensionality-reduction techniques, the suitability of those techniques from a pattern-recognition perspective is not always considered. This thesis is focused on analyzing one such dimensionality-reduction method, specifically PCA, and its suitability for classification applications.

2.4 Overview of Dimensionality-reduction Techniques

There are many different dimensionality-reduction techniques, and these are implemented based on their suitability to the application involved. A few of the most commonly used techniques are discussed here.

Dimensionality-reduction techniques using wavelet transforms are one approach. In this case, the signal is analyzed using an appropriate mother wavelet, and wavelet coefficients at different decomposition levels are used as features. An optimal feature subset is formed by combining wavelet-coefficients extracted from different levels of decompositions. The wavelet-coefficient extraction criterion is contingent upon the application involved. Previous work on dimensionality-reduction for classification purposes using discrete wavelet transforms (DWT) is an example of such an approach [7]. The wavelet-coefficient selection criteria employed by [7] use a class-discriminatory metric, such as area under the receiver operating characteristic (ROC) curve, for extracting the wavelet-coefficient feature set. Saito and Coifman proposed the method of finding the best bases for classification by pruning the wavelet tree based on similar class-separation criteria [11].

Another interesting dimensionality-reduction approach involves the selection of features from the original domain based on certain class-separability criteria. The class-separability criterion can be based on ROC, Battacharya distance (BD), Mahalanobis

distance, or even a simple Euclidean-distance measure. The work in [6] on extracting the best hyperspectral bands for detecting an invasive species, namely Kudzu, is an example use of such an approach. In this case, the spectral bands are used as features. Optimal sets of spectral bands are selected as features by searching through the original spectral domain. The features are selected and combined based on their discrimination capability. The ROC value is used as the class-separation metric, and the features are combined using LDA.

The discrete cosine transform (DCT) is an important dimensionality-reduction technique extensively used in image-compression applications, for example, JPEG compression [28]. The entire image is divided into smaller blocks of size 8-by-8. The pixel values of each block can be row-ordered to form a vector of size 1-by-64. Hence, each block forms a discrete point in a 64-dimensional space. To achieve dimensionality-reduction, the two dimensional DCT is applied to the 8-by-8 matrix. DCT coefficients are considered to be significant if they are above a predetermined threshold level. DCT coefficients considered to be insignificant are zeroed out. As a result, the dimensionality of each block can be reduced from 64-dimension to a lower-dimension. For example, a very low frequency image block (such as an 8-by-8 block containing part of the sky) of 64-dimension can be effectively projected onto a lower 1-dimensional space by retaining only the DC coefficient value of the DCT. Using the inverse DCT transform the 1-dimensional space can be projected back into 64-dimensional space without losing significant information; that is, the mean square error in the reconstructed image is small. Note that the DCT is an unsupervised transform. To perform the DCT on a given signal, no training data is required.

Linear discriminant analysis (LDA) is an example of a linear supervised transformation method for dimensionality-reduction [36]. In this case, the dimensionality is reduced to one less than the number of classes. For example, a three-class problem in a 200-dimensional space is reduced to a 2-dimensional three-class problem. The reduced feature set is optimized based on maximizing a class-separability function. The class-separability function is the ratio of S_B (between-class variance) to S_W (within-class variance). In order to construct the LDA transformation vector, the inverse of the S_W matrix needs to be computed. For cases of higher-dimensional data that do not have sufficient training data to support the large dimensionality, the S_W matrix becomes singular. Hence, the computation of the inverse of S_W is not possible. Thus LDA cannot be implemented directly on hyperspectral datasets for the purpose of dimensionality-reduction unless there is enough training data to complement the large dimension.

Principal component analysis (PCA) is another interesting dimensionality-reduction technique widely used in the field of image classification. The transform is based on the total data-distribution and does not distinguish between S_W and S_B. PCA can be used to project the data from the original high-dimensional space onto a lower-dimensional space, maximizing the variance present in the data. Although theoretically PCA is an optimal linear transformation method for image compression [25], due to the need for training data to estimate the optimum transform, PCA is not typically used for image-compression applications.

Traditional dimensionality-reduction techniques like PCA, LDA, and greedy-search methods implemented on transformed feature space rely on processing the entire dimensional space as a whole and extracting a set of features that can classify all the

classes with a certain accuracy. In recent years, researchers have established more appropriate methods of dimensionality-reduction by decomposing the entire space to a set of best subspaces for class separation and extracting more reliable features from this lower-dimensional subspace. Kumar *et al*. have shown that hyperspectral space can be partitioned into groups based on adjacent group-band pair correlation and discrimination [9]. The spectral features of each group are projected using LDA for each class pair. This method relies on a pairwise classification scheme where a C class problem is divided into a larger set of $(C,2)$ class pairs. Work by Jimenez and Landgrebe has shown that an entire hyperspectral space can be decomposed into subspaces using BD as the criterion [10]. The adjacent bands are grouped together to form band-groups, if they maximize the minimum pairwise class separation estimated using BD. By combining linearly projected features from each subspace, a lower-dimensional feature set is formed such that the minimum pairwise class separation is maximized in this lower-dimensional space. Discriminant analysis performed on this lower-dimensional feature set yielded more reliable classification features. Saito and Coifman introduced a method of finding a local discriminant bases for classification by pruning the wavelet tree based on similar discrimination metrics [11]. Jimenez *et al.* have reported that feature and decision fusion techniques can be integrated to improve the classification potential of hyperspectral data [12]. Benediktsson and Kanellopoulos showed that feature-level fusion when combined with decision-level fusion yields higher classification accuracies [13]. In their proposed method, hyperspectral data is partitioned into uncorrelated groups by analyzing the correlation matrix. Features are extracted from these subgroups using decision-boundary feature-extraction techniques. The decisions made by these extracted features are

combined at the decision level using a logarithmic opinion pool. Jia and Richards have proposed a scheme for efficient image classification and display based on segmented principal component transforms [17]. The hyperspectral data is divided into highly correlated subgroups, and features are extracted from these groups by applying PCA on each subgroup. PCA features from each subgroup are selected based on their BD.

CHAPTER III

PRINCIPAL COMPONENT ANALYSIS

3.1 OVERVIEW OF PRINCIPAL COMPONENT ANALYSIS

PCA transforms or projects the features from the original domain to a new domain (known as PCA domain) where the features are arranged in the order of their variance. The features in the transformed domain are formed by the linear combination of the original features and are uncorrelated. Dimensionality-reduction is achieved in the PCA domain by retaining only those features that contain a significant amount of information. Geometrically, this process can be looked upon as a rotation of the axes of the original vector space to form a set of orthogonal axes for the PCA space. From Figure 3.1, it is clear that PCA axes are formed by the rotation of the data-distribution. The new set of orthogonal axes, namely PCA I and PCA II, are ranked according to the amount of variance they account for in the original data. The direction and magnitude of these principal component axes are computed by performing an eigen decomposition of the total covariance matrix of the multivariate data. The eigenvalues determine the magnitude of the principal component axes, and eigenvectors determine the directions. The amount of variance in the data is represented by the eigenvalues.

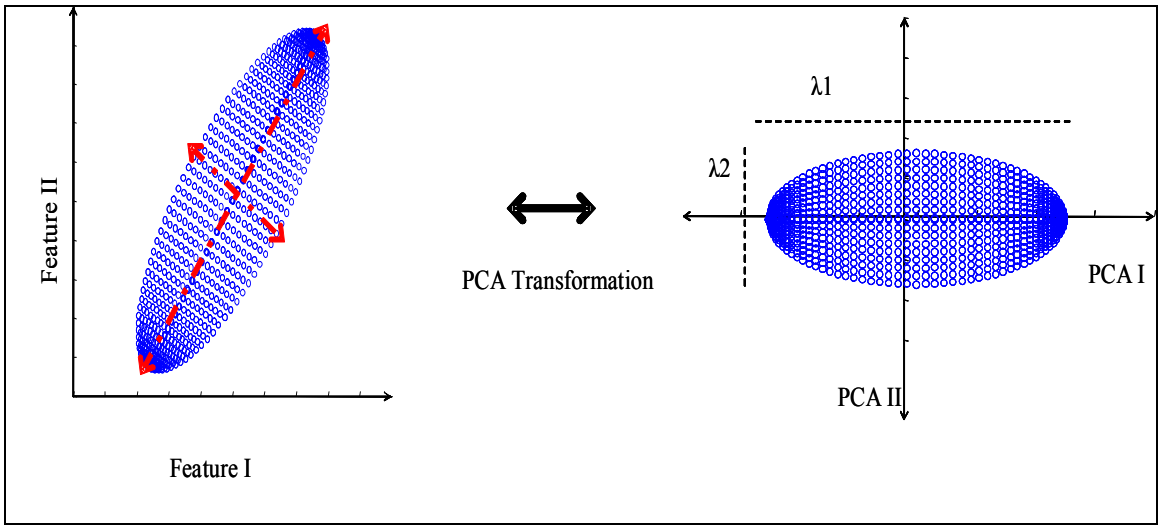

Figure 3.1 Geometrical Representation of PCA Transform.

3.2 General PCA Approach to Dimensionality-Reduction

PCA is a widely used unsupervised dimensionality-reduction technique. It is currently being used as a dimensionality-reduction tool in a wide range of applications, such as document analysis, data mining, content-based image retrieval, face recognition and spectral remote-sensing. PCA is used to reduce a high-dimensional vector to a low-dimensional vector by exploiting the correlation existing in the data. PCA decorrelates the resulting components, and the lower-order components are discarded. For example, in hyperspectral data, PCA tries to capitalize on the large inter-band correlation existing between neighboring bands found in hyperspectral data.

The general approach of using PCA in a multispectral or hyperspectral image-processing application requires the computation of the eigenvectors and eigenvalues of the total covariance matrix calculated from the spectral images. The PCA computation done on an *N*-by-*N* AVIRIS image having 224 contiguous spectral bands is explained below.

Compute the mean 224-band spectral signature, $\vec{m}$, from the N^2 pixels in the AVIRIS spectral image. Compute the covariance matrix of the 224-band spectral image. The covariance matrix Σ is defined as:

$$\Sigma = \sum_{i=1}^{N^2} (\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m})^T \ , \tag{2}$$

where $\vec{x}_i$ is the i^{th} spectral signature, $\vec{m}$ denotes the mean spectral signature and N^2 is the total number of spectral signatures. In order to find the new orthogonal axes of the PCA space, eigen decomposition of the covariance matrix Σ is performed. The eigen decomposition of the covariance matrix is given by

$$\Sigma \vec{a}_k = \lambda_k \vec{a}_k \, , \tag{3}$$

where λ_k denotes the k^{th} eigenvalue, $\vec{a}_k$ denotes the corresponding eigenvector and k varies from 1 to 224.

The eigenvalues denote the amount of variance present in the corresponding eigenvectors. The eigenvectors form the axes of the PCA space, and they are orthogonal to each other. The eigenvalues are arranged in decreasing order of the variance. In order to achieve dimensionality-reduction without losing much of the significant information, only those eigenvalues that constitute a significant level of information are retained. This is done by retaining only those eigenvalues which are above a preset threshold value, or by retaining only the first few eigenvalues that add up to a predetermined percentage of the total variance present in the data. The PCA transformation matrix, A, is formed by choosing the eigenvectors corresponding to the largest eigenvalues. The PCA transformation matrix A is given by

$$A = [\vec{a}_1 \mid \vec{a}_2 \mid ..... \mid \vec{a}_J] \tag{4}$$

where $\vec{a}_1, \vec{a}_2 ... \vec{a}_J$ are the eigenvectors associated with the J largest eigenvalues obtained from the eigen decomposition of the covariance matrix Σ. The data projected onto the corresponding eigenvectors form the reduced uncorrelated features that are used for further classification processes.

3.3 DATA ANALYSIS USING PCA

To better understand how PCA is used for effective data analysis, consider the following scenario where PCA is applied to a target-recognition problem. Assume that the significant feature of a target is its high reflectance value in the red and near-infrared (NIR) regions of the spectrum. The non-target is discriminated by its low reflectance in either the red or NIR region. Figure 3.2 clearly shows how the information in the red and NIR bands have been compressed to form a single dimension, which is the largest principal component (PCA I), calculated using the eigen decomposition of the covariance matrix estimated from the data-distribution. By measuring the data projected onto this new dimension, objects having high reflectance in the red and NIR bands (targets) can be easily discriminated from objects that are having low reflectance values either in the red or NIR band. The amount of information that is retained by this single dimension can be determined from the eigenvalues. For certain data-distributions, application of PCA improves the classification accuracies as well as reduced the computational load in processing higher-dimensional data.

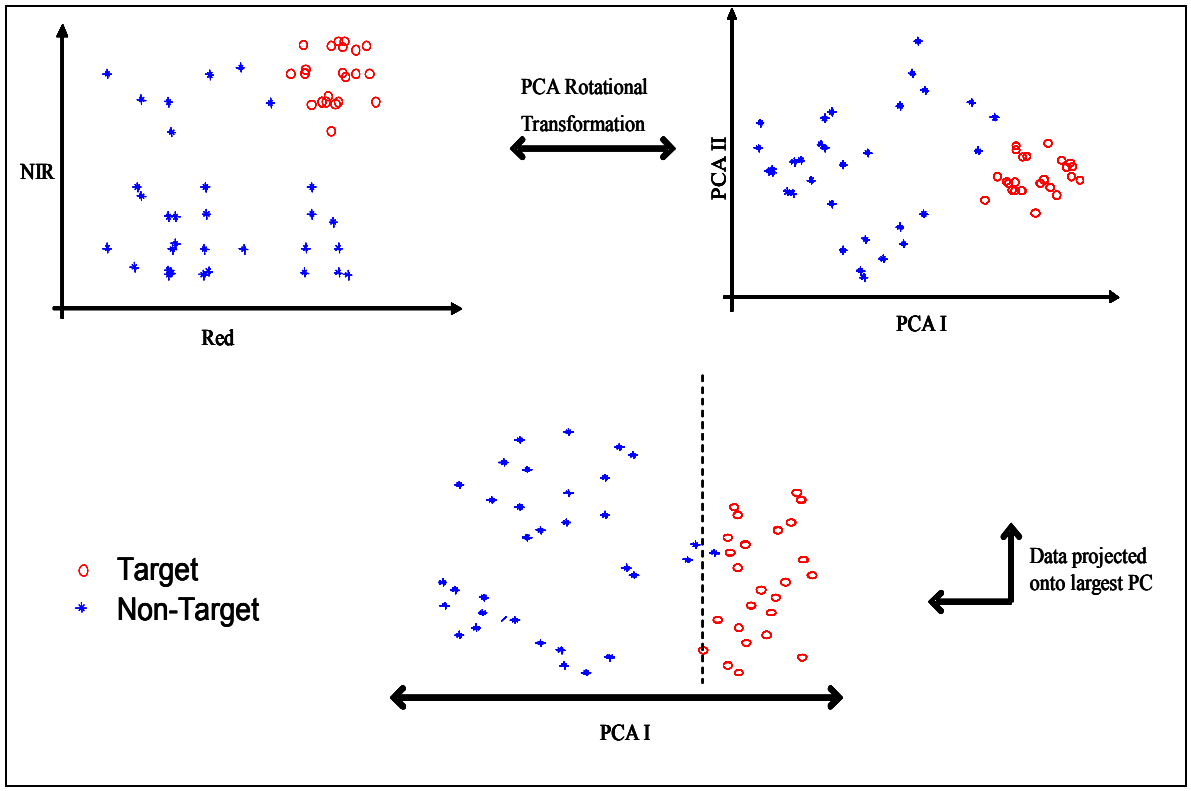

Figure 3.2 Information Compression using PCA.

Next, consider the example where PCA is applied to a 3-dimensional data-distribution as shown in Figure 3.3. The data is generated artificially in order to demonstrate the effectiveness of PCA. In this example, PCA transforms the data in such a way that the data points of the two classes are less clustered, so as to maximize the variance within the data.

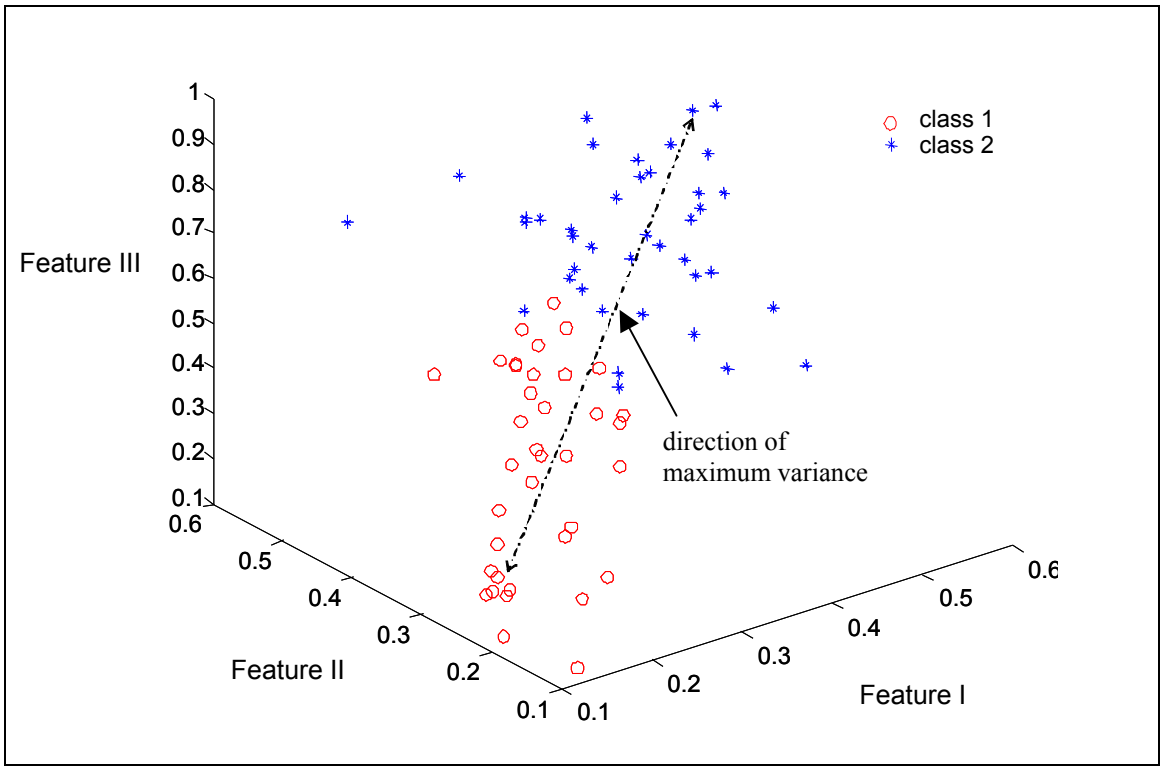

Figure 3.3 An Example 3-Dimensional Data-distribution.

It can be observed from Figure 3.4 that the data points of the two classes are well separated in the new transformed feature space, thus enabling the classifier to accurately identify the underlying classes. Each of the axes of the transformed space, which are also referred as principal components, is arranged in the order of the decreasing variance. The length of the principal components represents the variance present in that direction. It is observed from Figure 3.3 to Figure 3.5 that the first principal component (PCA I) captures much of the variance (dotted line in Figure 3.3) present within the data-distribution. The PCA II and PCA III capture the rest of the variance present in the data. It can be seen from Figure 3.4 that the variance in these directions are relatively small.

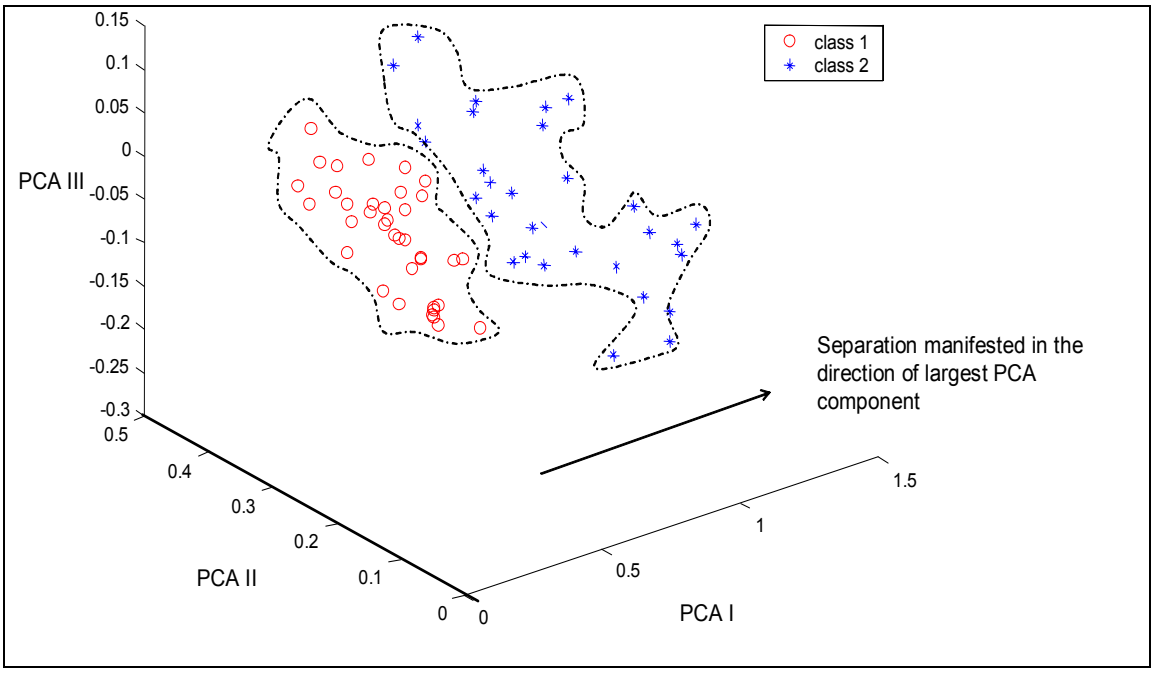

Figure 3.4 PCA Applied on 3-Dimensional Space to Improve Class Separation in Lower-dimensional space.

By ignoring these two components (PCA II and PCA III) and by retaining the largest principal component (PCA I), we can achieve dimensionality-reduction as well as retain much of the information present in the data. In this example, the dimensionality of the data is reduced from three to one. It can be observed from Figure 3.5 that in this example, the data-distribution projected onto its largest principal component retains almost all of the class discriminatory information required to improve the classifier accuracy. However, later in the thesis it will be shown that this is not always the case. In Figure 3.5, the Y-axis represents the sample number. From hereafter in this thesis, the Y-axis for similar one-dimensional projections represents the sample number.

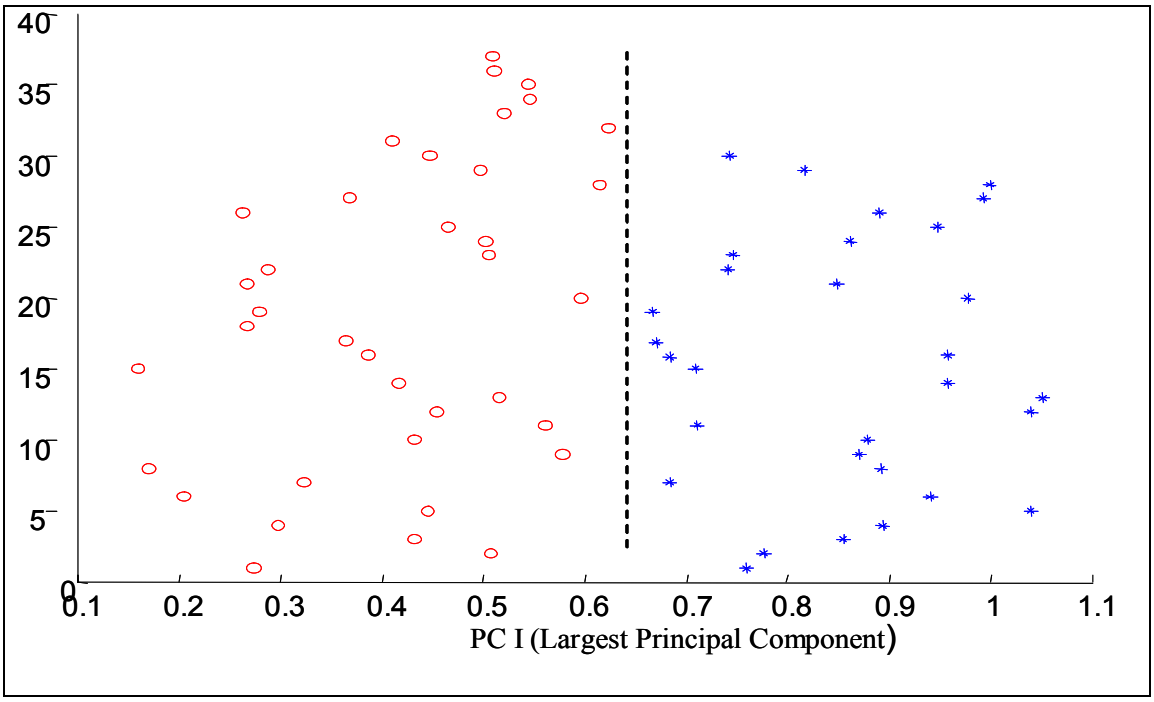

Figure 3.5 Data Projected onto the Largest Principal Component.

PCA results in establishing the lower-dimensional projection that maximizes the variance present in the data.However this gives rise to a few questions which are the basis for this thesis. Does PCA yield discriminating features if the variance it tries to maximize is not oriented in a direction that facilitates class discrimination? In this case, does PCA projection work in favor of or against the classification accuracy? If the PCA projection does indeed separate the classes, can the principal components be arranged based on the quality of information (discrimination capability) rather than on the quantity of information (variance represented by eigenvalues)? By ignoring certain lower-order principal components for dimensionality-reduction, is it certain that any of the suitable classification features are not lost? Thus, the main objective of this thesis is to analyze PCA, which is one of the most commonly used dimensionality-reduction techniques in multispectral and hyperspectral remote-sensing, to understand its appropriateness as a dimensionality-reduction technique from a classification application perspective.

3.4 HYBRID PCA APPROACHES FOR DIMENSIONALITY-REDUCTION OF HYPERSPECTRAL DATA

In addition to the general PCA approach, PCA has been used in combination with other analysis methods. A hybrid method called multiscale principal component analysis (MSPCA) that has the capability of PCA to extract linear information from the variables as well as the multi-resolution capabilities of wavelet analysis is applied for dimensionality-reduction in [15]. MSPCA is appropriate for data that has significant changes over both time and frequency. In this approach, PCA is applied on wavelet coefficients obtained at different scales.

In order to account for some of the limitations of PCA with respect to maximizing the class separation between the underlying classes, a unified PCA/LDA approach is reported in [16] for face-recognition tasks. The features obtained using PCA are further processed using LDA to obtain classification features that can discriminate the underlying classes and hence increase the classification accuracy.

Another approach called segmented principal component transformation (SPCT) is developed for efficient hyperspectral image classification and display in [17]. The hyperspectral data is divided into highly correlated subgroups, and features are extracted from these groups by applying PCA on each subgroup. PCA features from each subgroup are selected based on their BD. This approach helps to reduce the computational load significantly compared to the general PCA approach and also avoids the PCA transform becoming biased due to large variance at certain regions of the spectrum.

Maximum noise fraction (MNF) or noise-adjusted principal component (NAPC) is a transformation method similar to PCA, wherein the principal components are not ranked based on their variance but based on the signal-to-noise quality [29]. The authors

of this paper recognize that PCA does not always produce images that show steadily decreasing image quality with increasing component number. The MNF transform is constructed with the knowledge of signal covariance and the corresponding noise covariance. The transform is constructed such that the resulting projection maximizes the signal to noise covariance ratio. The authors of [29] paper discuss methods for indirect estimation of the noise covariance.

CHAPTER IV

THEORETICAL ASSESSMENT OF PCA

This chapter focuses on the characteristics of hyperspectral data and its influence on PCA as feature-extraction technique. This chapter includes an investigation of the effect of data-distribution on PCA, and a theoretical analysis on why PCA may not be appropriate for feature-extraction when applied to certain data-distributions.

4.1 DATA CHARACTERISTICS AND PCA

In this section, characteristics of hyperspectral data such as correlation, within-class variance, and between-class variance are investigated. The effects of these characteristics on the PCA method are also analyzed.

4.1.1 Data Correlation and Dimensionality-reduction

Hyperspectral data is highly correlated. This fact can be corroborated by the observation that hyperspectral data tends to be distributed in the shape of a hyperellipsoid [4]. The high inter-band correlation can also be corroborated by viewing an inter-band correlation matrix for a hyperspectral dataset. An example inter-band correlation matrix of a 196-band hyperspectral dataset is shown in Figure 4.1. The white areas in the image represent high correlation and the black areas represent low correlation. It can be

observed from Figure 4.1 that the neighboring-band areas that are white in color represent high correlation, and as the bands are apart the correlation decrease which is represented by the black areas in the image. The image contains two black strips, which are the water-band regions of the spectrum, and they are zeroed out for correlation computation. Linear-transformation methods such as PCA try to exploit the correlation present in the data to achieve information compression. As the correlation increases, the data-distribution can be reduced to fewer dimensions without loss of much information.

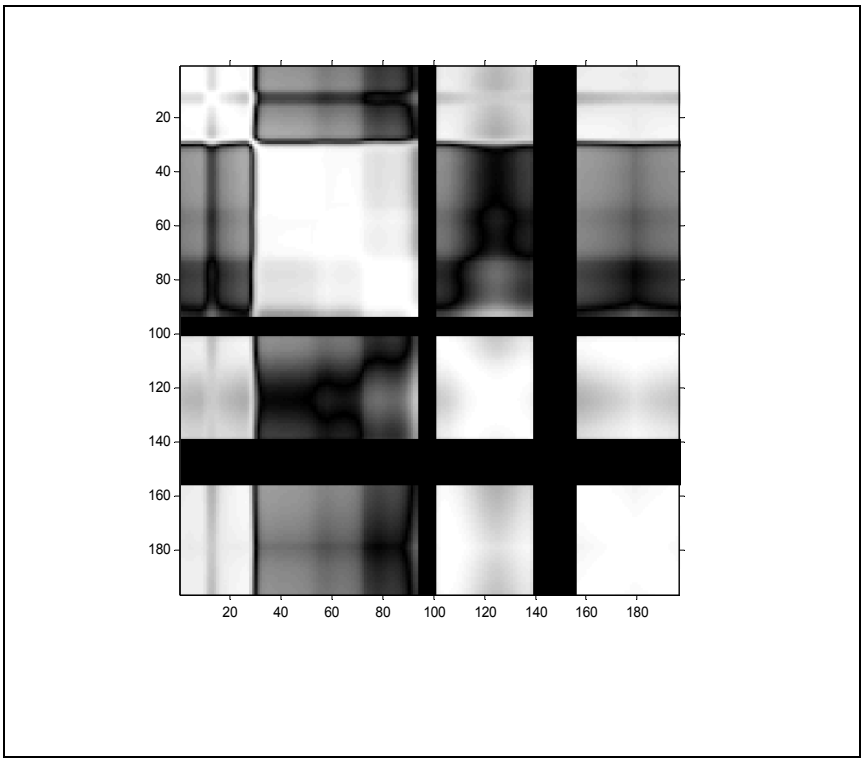

Figure 4.1 Correlation Matrix for 196 band Hyperspectral Data (White = 1 or -1 black ≈0).

Consider two artificially generated data-distributions, Case I and Case II, as shown in Figure 4.2 and Figure 4.3, respectively. The Case I data-distribution is highly correlated, and the 2-dimensional scatter plots show high correlation existing between

features. For such data-distributions, linear-transformation methods that exploit the correlation information may be suitable for achieving dimensionality-reduction.

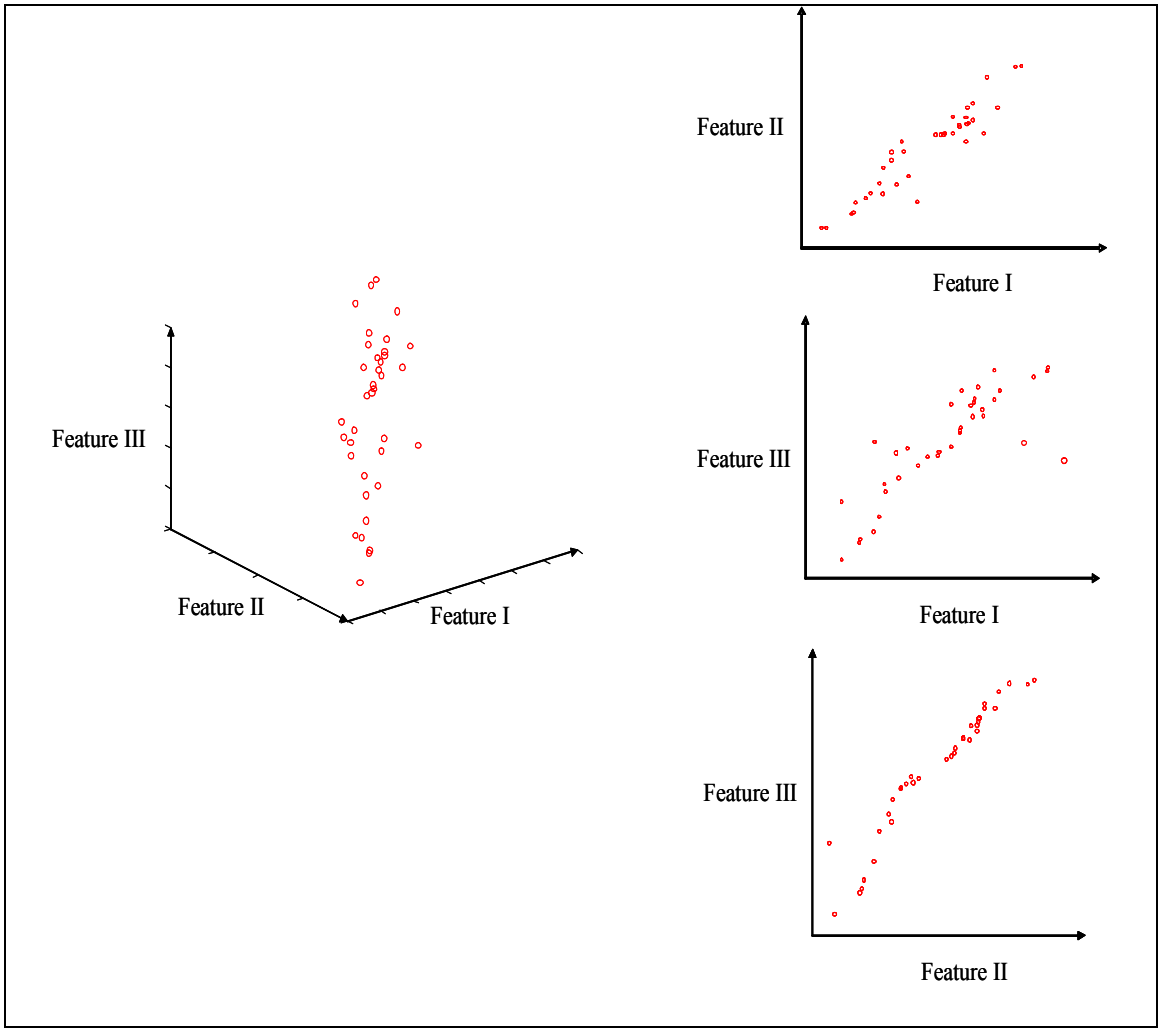

Figure 4.2 Example *Case I* Data-distribution and Scatter Plots.

For the Case II data-distribution, it can be observed from Figure 4.3 that the data is weakly correlated. The 2-dimensional scatter plots representing the inter-feature correlation further demonstrates this fact. Linear transformations like PCA may not yield a high degree of dimensionality-reduction in this case.

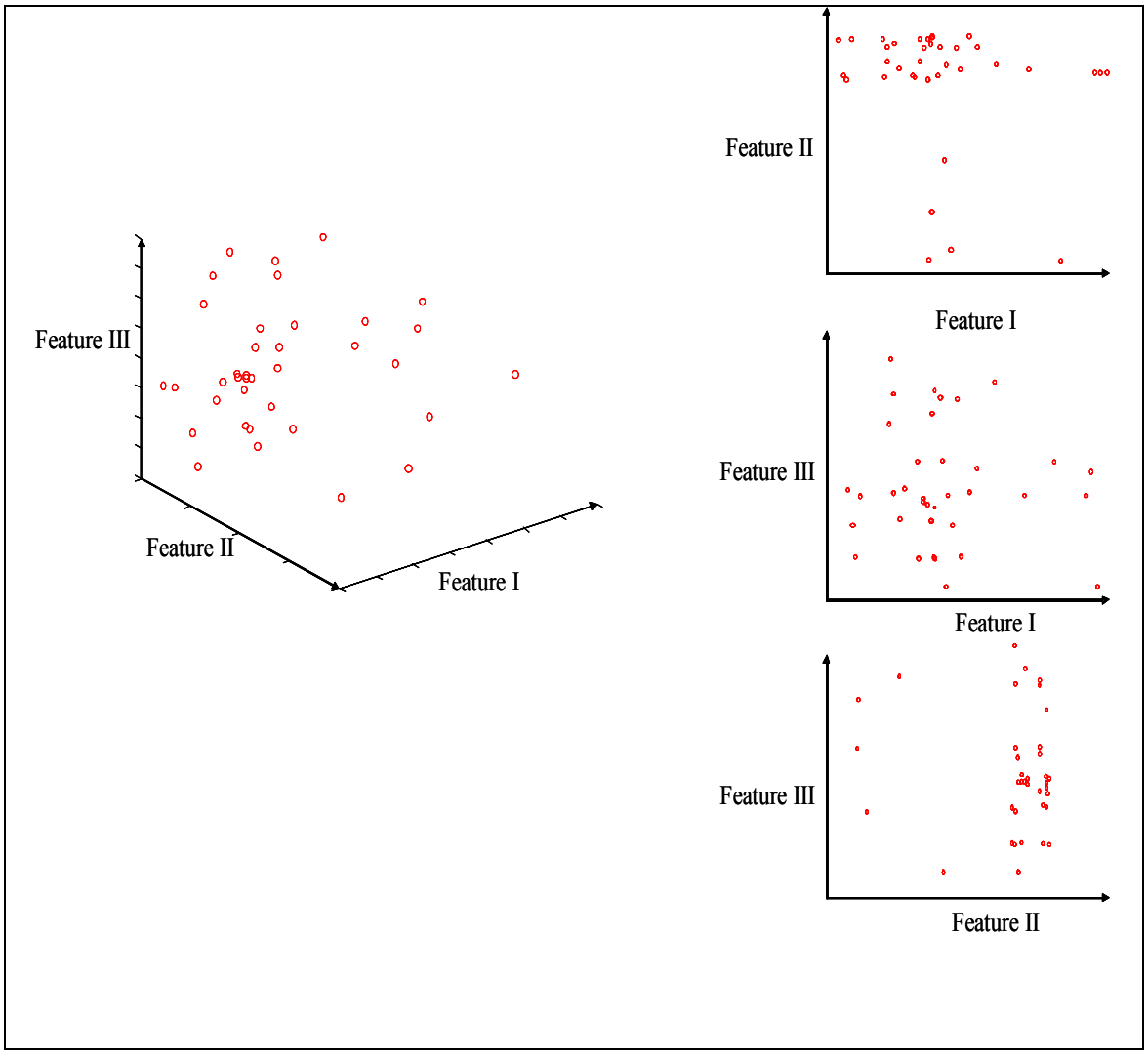

Figure 4.3 Example *Case II* Data-distribution and Scatter Plots.

Table 1 and Table 2 show the correlation coefficients and eigenvalues obtained by applying PCA to the Case I and Case II data-distributions respectively. It is seen that for the Case II data-distribution wherein data correlation is very weak, the information content in each principal component, which is represented by the magnitude of the corresponding eigenvalues, is almost equally distributed. Hence, PCA may not be able to achieve dimensionality-reduction in this case without sacrificing major information present in the original data-distribution. For the Case I data-distribution wherein the data

correlation is high, the information content is gathered around certain principal components. Hence, in this case, PCA will yield a better dimensionality-reduction as well as retaining a high percentage of the total information present in the original data-distribution when only those principal components associated with large eigenvalues are retained.

Table 1

CORRELATION COEFFICIENTS AND EIGENVALUES OBTAINED BY APPLYING PCA FOR *CASE I* DATA-DISTRIBUTION

	Feature I	Feature II	Feature III	Eigenvalues
Feature I	1.0	0.942	0.962	0.0108
Feature II	0.942	1.0	0.922	0.0002
Feature III	0.962	0.922	1.0	0.0001

Table 2

CORRELATION COEFFICIENTS AND EIGENVALUES OBTAINED BY APPLYING PCA FOR *CASE II* DATA-DISTRIBUTION

	Feature I	Feature II	Feature III	Eigenvalues
Feature I	1.0	-0.206	-0.054	0.0422
Feature II	-0.206	1.0	-0.127	0.0338
Feature III	-0.054	-0.127	1.0	0.0215

Figure 4.4 shows the plot of normalized eigenvalues against the principal-component number. It is clear that, for the Case I data-distribution, the dimensionality can be reduced from 3 to 1 by keeping a normalized eigenvalue threshold of 0.1. For Case II data-distribution, setting a threshold between 0.1 and 0.5 will not result in any dimensionality-reduction.

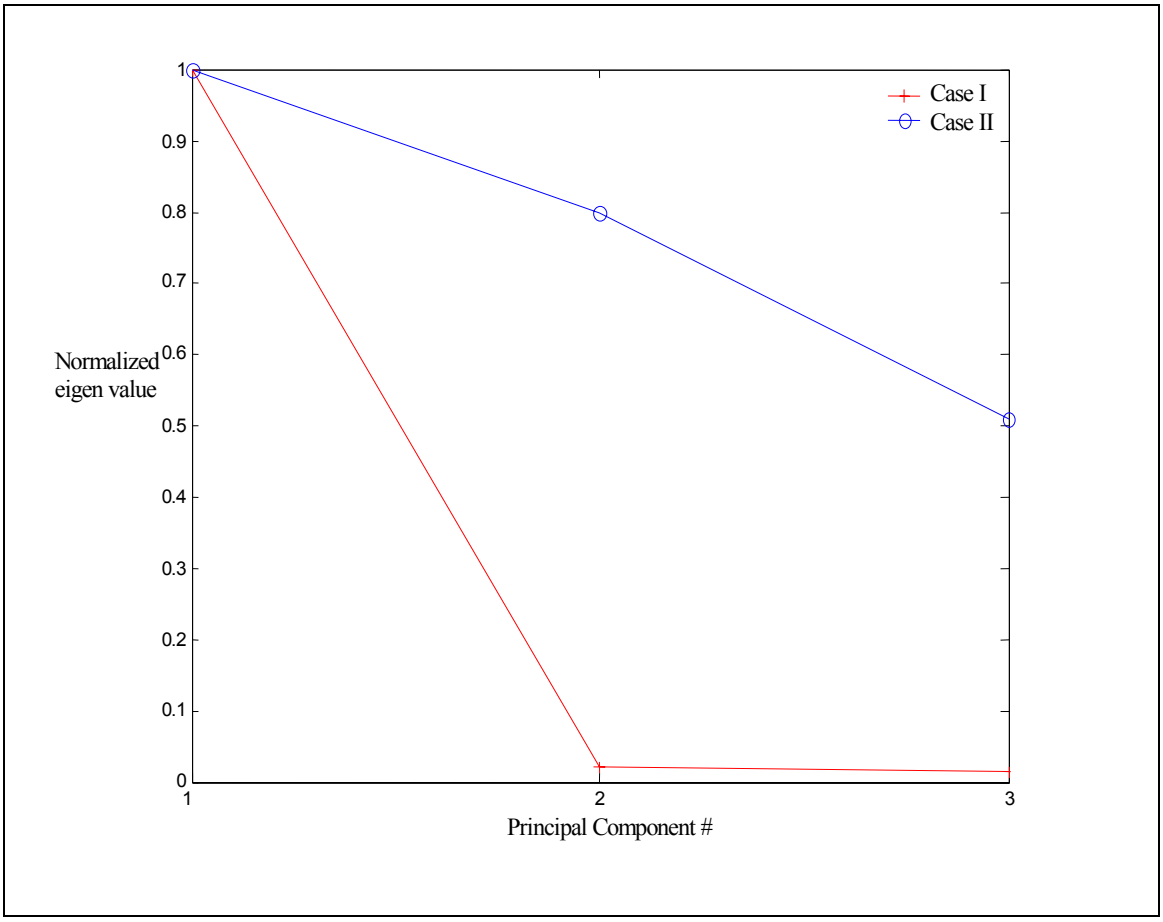

Figure 4.4 Normalized Eigenvalues against Principal Component Number.

4.1.2 Influence of Within-Class Variance of Data on PCA

The within-class variance (S_W) is a crucial characteristic of data-distributions that can adversely affect the feature-extraction capabilities of linear transformations like PCA.

The S_W can be quite prominent in the case of hyperspectral data. The S_W is caused by several factors such as natural variation in the target material, environmental conditions, and sensor angle. Even though some of these naturally occurring variances can be compensated for, S_W can still degrade classification accuracy. If PCA is applied to data-distributions wherein the largest variation in the data is due to within-class variations, then PCA may increase these within-class variations further. Hence, the reduced feature set obtained from a PCA transformation may not represent the separation of the underlying classes; instead it may represent the within-class variations that exist in the data. Hence, in this thesis, it is emphasized that, even though PCA is a popular dimensionality-reduction technique, implementing PCA for feature reduction and selection without clearly understanding the data-distribution statistics may yield adverse results.

The within-class covariance matrix S_W is given by

$$S_w = \sum_{i=1}^{C} \frac{n_i}{n} \Sigma_i , \tag{5}$$

where n_i is the number of training samples for class ω_i, n is the total number of training samples, Σ_i is the covariance matrix of class ω_i, and C is the total number of classes. Σ_i is given by

$$\Sigma_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (\vec{x}_j - \vec{m}_i)(\vec{x}_j - \vec{m}_i)^T , \tag{6}$$

where $\vec{x}_j$ is the jth training sample from class ω_i, and $\vec{m}_i$ denotes the sample mean of class ω_i.

Figure 4.5 shows a 2-dimensional feature space representing two different class distributions. It can be observed that the direction of the maximum variance of the data-distribution (dotted lines) represents the general direction of the within-class variance of each class.

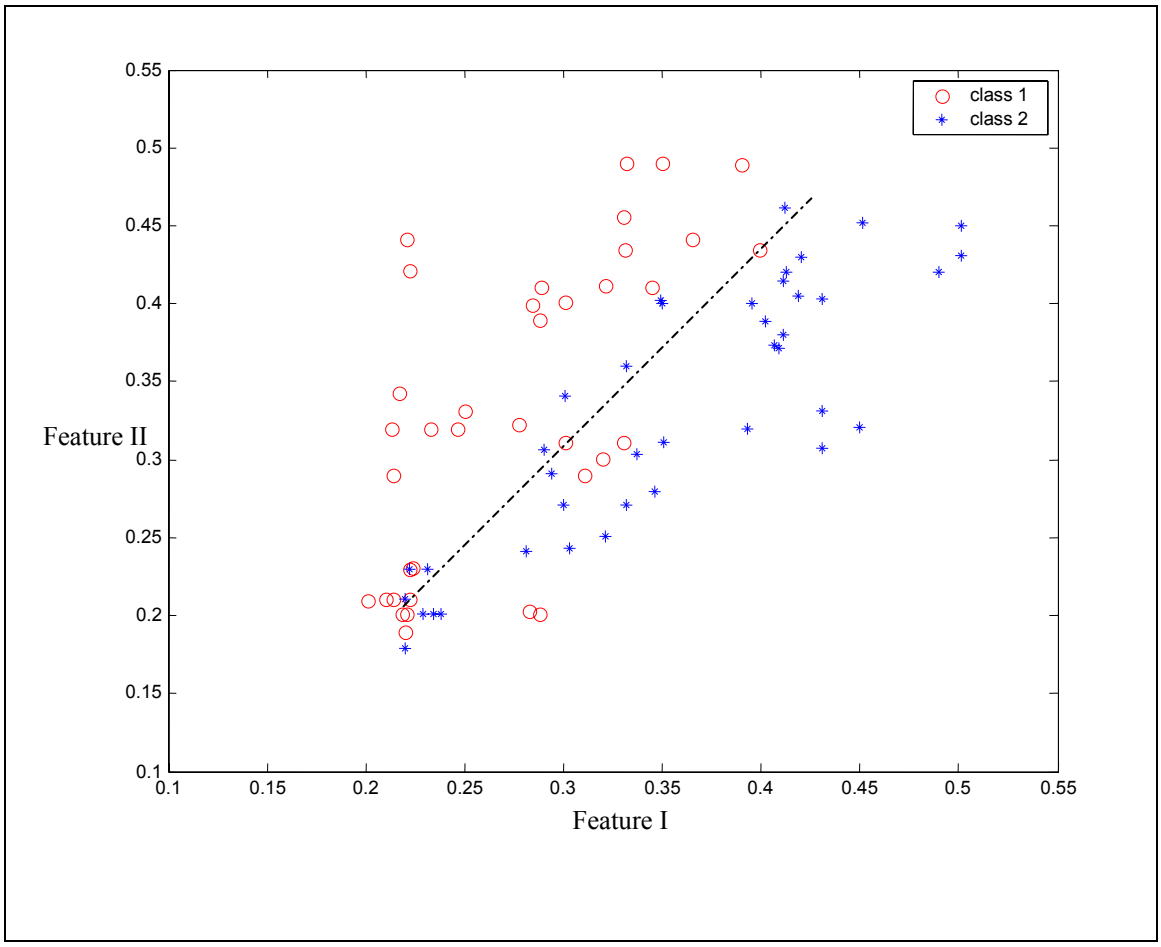

Figure 4.5 Demonstration of Within-Class Variance for a 2-Dimensional Data-distribution.

For aforementioned reasons, PCA applied to this data-distribution yields a data projection in the lower-dimension that is oriented in such a way so as to maximize the within-class variance existing in the data. It can be observed from Figure 4.6 that none of the principal components are suited to discriminating the classes. Note the increase in the

within-class variance in PCA space. This projection will not reveal any features that could maximize the class separation. Hence, in the case of hyperspectral data, PCA transform may be biased towards the large within-class variations existing in certain spectral regions.

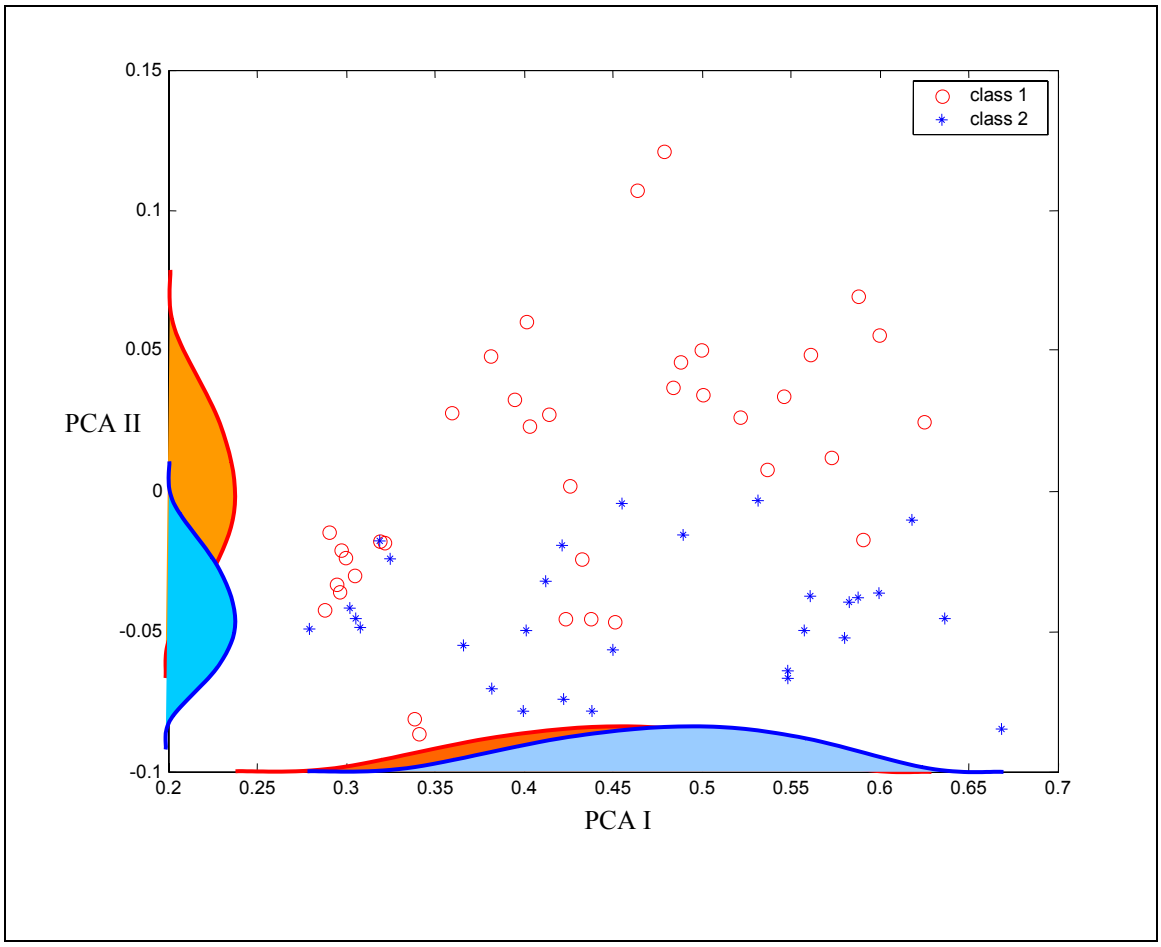

Figure 4.6 Two Dimensional Data-distribution Projected onto PCA space.

4.1.3 Influence of Between-Class Variance of Data on PCA

The between-class variance (S_B) represents the separation between the means of the underlying classes. In the case of hyperspectral data, the dissimilarities in reflectance values of certain spectral bands for different classes give rise to the relatively low

between-class variance. The goal of any linear transformation for feature-extraction purposes should be to maximize this between-class variance and to minimize the within-class variance present in the data.

The between-class variance matrix S_B is given by

$$S_B = \sum_{i=1}^{C} \frac{n_i}{n} (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T, \tag{7}$$

where n is the total number of training samples, n_i is the number of training samples for class ω_i, C is the total number of classes, $\vec{m}_i$ is the sample mean for class ω_i, and $\vec{m}$ is the total sample mean given by

$$\vec{m} = \sum_{i=1}^{C} \frac{n_i}{n} \vec{m}_i. \tag{8}$$

Figure 4.7 shows a 2-dimensional data-distribution representing two classes. The direction in which the maximum variance is oriented is represented by the dotted line in the figure. In this case, the maximum variance is oriented in same direction as the between-class variance existing in the data. From Figure 4.7, it can be observed that none of the features in the original feature space is suited to discriminate the classes. PCA applied to such a data-distribution will result in establishing a lower-dimensional space wherein the class separations are further increased. From Figure 4.8, it can be observed that PCA I is an ideal feature to discriminate the underlying classes.

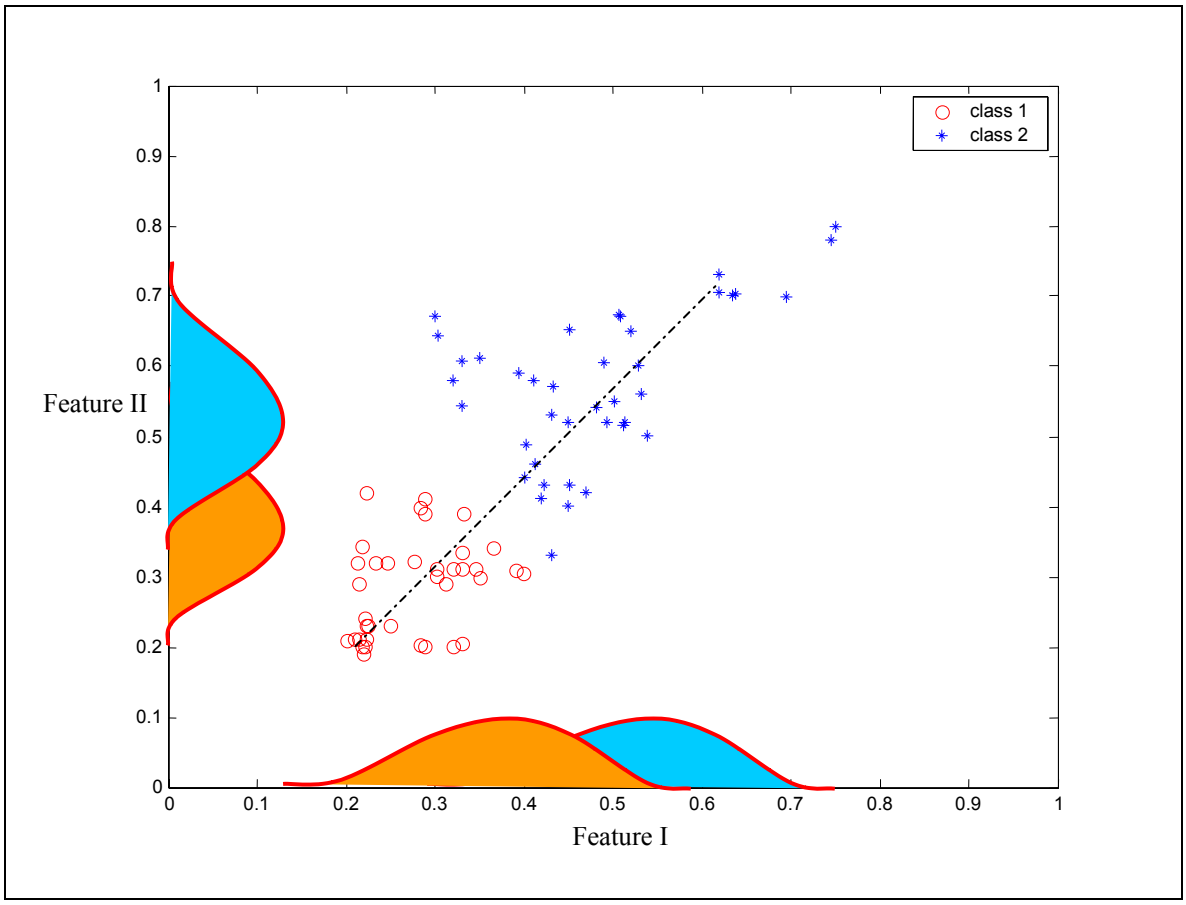

Figure 4.7 Demonstration of Between-Class Variance for a 2-Dimensional Data-distribution.

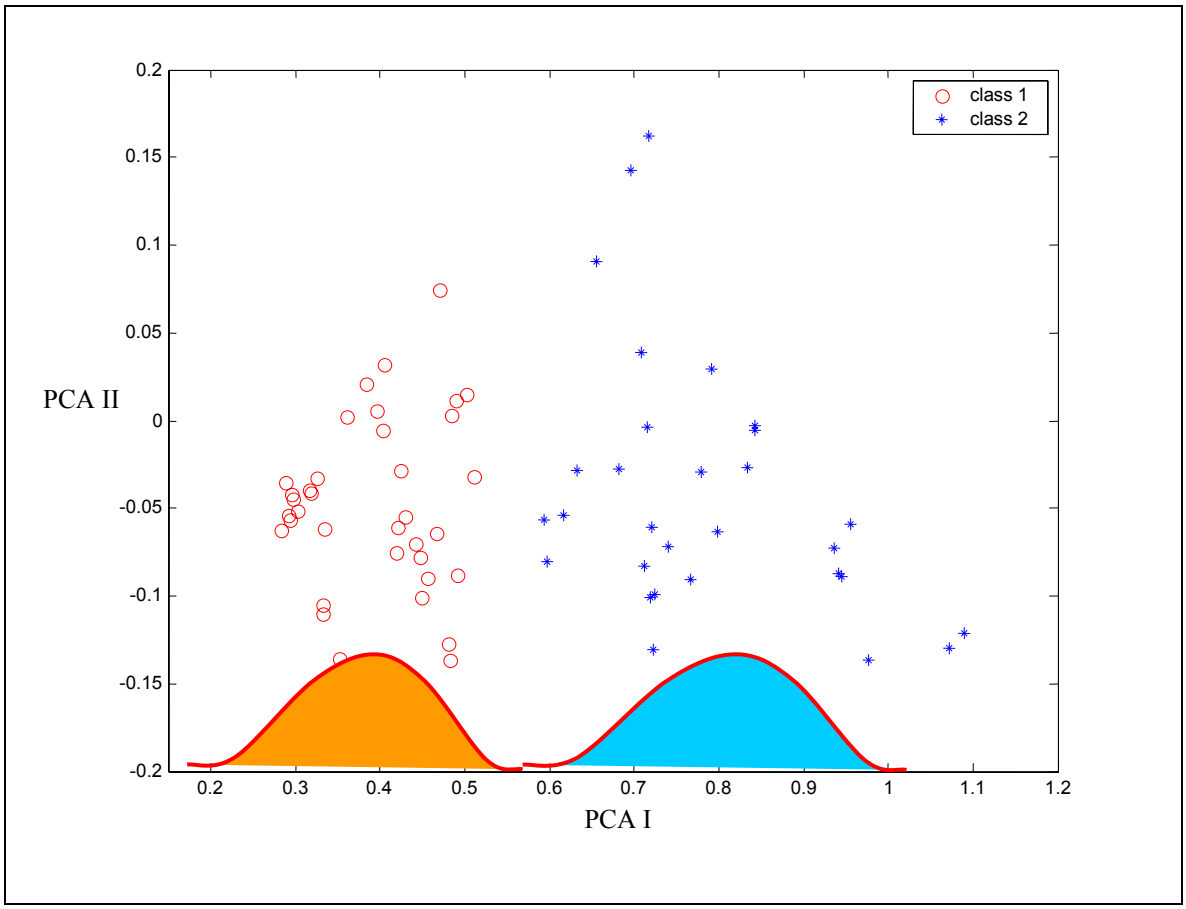

Figure 4.8 An Ideal PCA transformation.

4.2 GOAL OF DIMENSIONALITY-REDUCTION: COMPRESSION VS. CLASSIFICATION

PCA has been theoretically proven to be optimal for information compression [25]. The objective of dimensionality-reduction for information compression is to achieve high ratio of information compression to reconstruction distortion. PCA has the ability to extract the lowest dimensional structure with which the higher-dimensional structure can be reconstructed with least mean square error. The mean square error, e_{ms}, between the original signal and the reconstructed signal is given by the following equation

$$e_{ms} = \sum_{i=1}^{k} \lambda_i - \sum_{i=1}^{J} \lambda_i \qquad (9)$$

$$= \sum_{i=J+1}^{k} \lambda_i \,, \tag{10}$$

where λ_i represents the ith eigenvalue, k denotes the original dimension, and J represents the reduced dimension. The disadvantage of PCA is its requirement for training data to construct the transformation; this is not always feasible in an information-compression scenario.

From a classification perspective, the goal of PCA transformation is not to extract a lower-dimensional structure that can be projected back to its higher-dimensional space with the least error, but rather to extract the lower-dimensional structure that contains information to discriminate between the underlying classes. It can be seen from the equation (9) that, by retaining only the principal components having higher eigenvalues, the goal is to lower the reconstruction error and not to enhance the discrimination capabilities of the principal components.

4.3 Analysis of Effect of Data-distribution on PCA

PCA is a rotational transformation method wherein the axes in the original domain are rotated to form new orthogonal axes in the PCA space. The features projected onto the principal components in PCA space are uncorrelated. Since the original axes are involved in rotation, there is no loss of information due to the transformation itself. Information loss occurs only when some of the new orthogonal axes are discarded for dimensionality-reduction purposes. In PCA, the rule for achieving dimensionality-reduction is determined by measuring the quantity of information in the form of variance and retaining those dimensions that constitute a preset quantity of information. It is already seen that PCA is optimal in a mean squared error sense [25].

From a classification or detection perspective, the important criterion is not the quantity of information that is gathered, but the intelligibility of the information to discern the underlying classes. Consider Figure 4.9 wherein an artificially generated two-class data-distribution is plotted in a 3-dimensional feature space. It can be seen that the maximum variance in the data-distribution is not oriented in a direction that favors class discrimination (see the dotted lines). It can be observed from Figure 4.10 that Feature II in the original feature space itself is a good classification feature that can classify the data points into its classes with 100% accuracy.

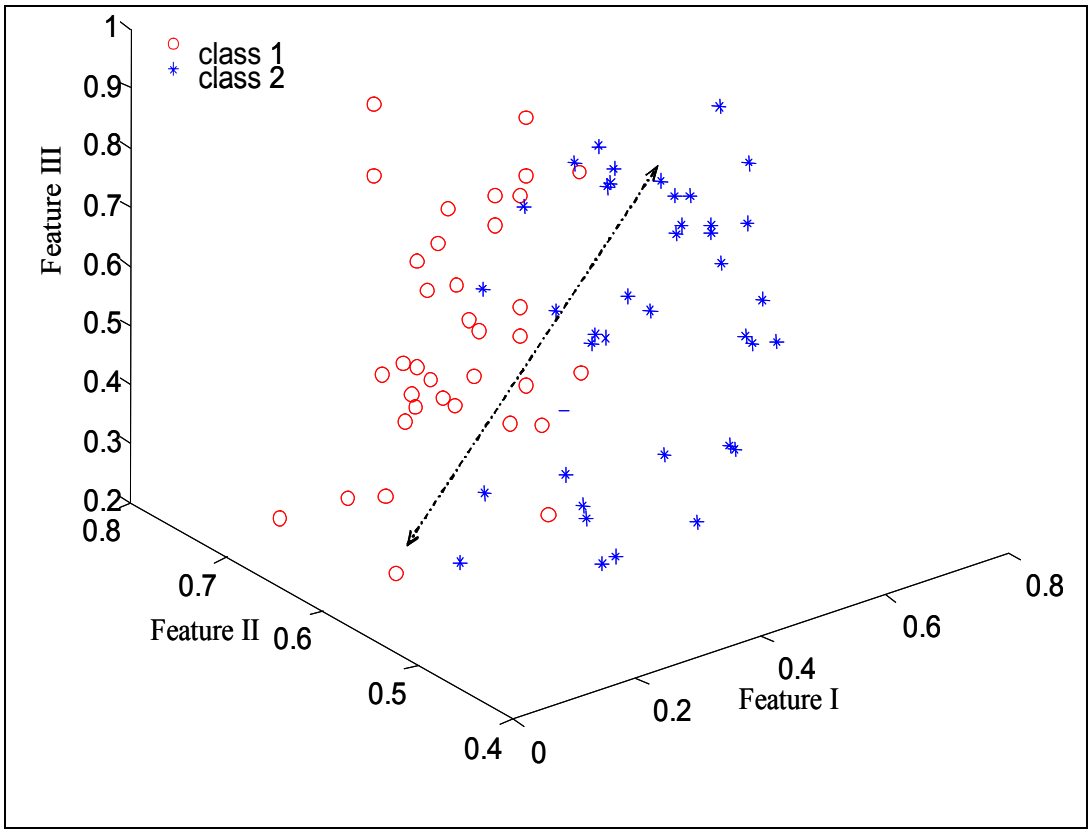

Figure 4.9 An Example Two-Class Data-distribution Showing Maximum Variance in a Direction that does not Favor Class Discrimination.

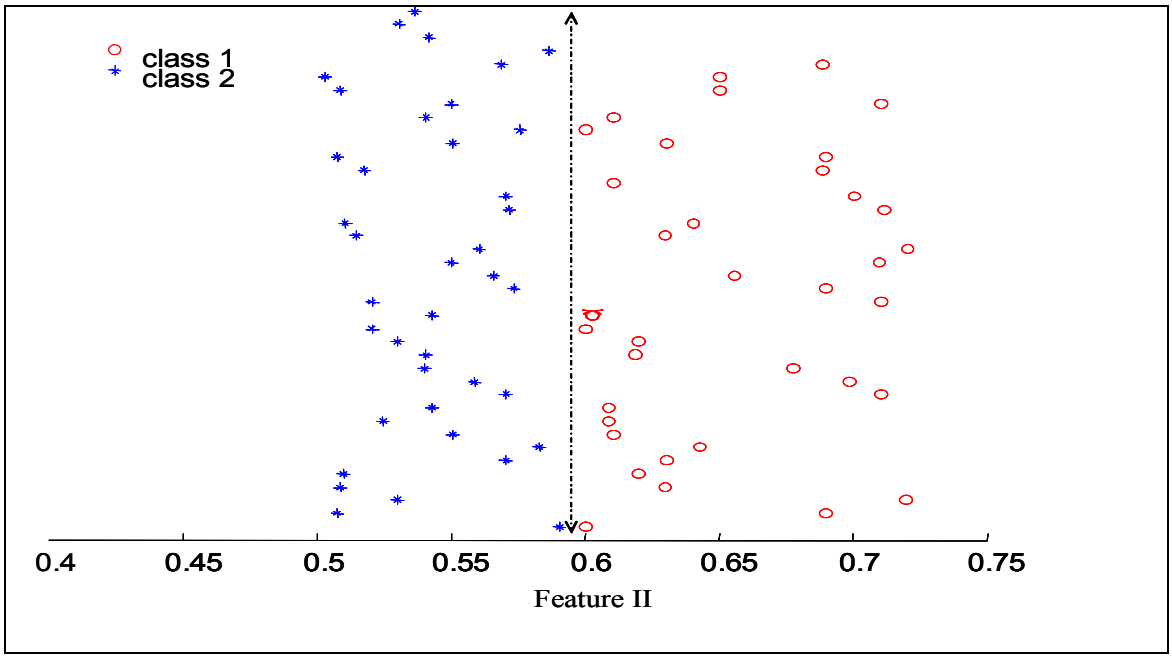

Figure 4.10 Discrimination Capability of Feature II.

Assume that the 3-dimensional data-distribution is subjected to PCA in an effort to find the lower-dimensional structure with which classification can be done with better computational efficiency. As it can be seen from Figure 4.11, the maximum variance in the data-distribution is oriented in a direction along with largest principal component. The data is projected onto the individual principal components in order to determine whether any of the principal components retains the discrimination capability of Feature II from the original feature space. It is shown in Figure 4.12 that none of the transformed principal components are able to retain this classification accuracy. In order to compute the classification accuracies of the data projected onto the principal components, a maximum likelihood classifier is used. Maximum likelihood classifier is a parametric classifier wherein the data-distribution parameters computed for deciding the classification boundaries are the mean and covariance of the data-distribution [36]. A leave-one-out test strategy is used for the accuracy estimation. In leave-one-out testing,

each sample is tested against the entire data-distribution in each iterative step. Further description of leave-one-out test strategy can be found in Chapter IV. The largest PCA component by itself results in a maximum likelihood classification accuracy of only 62%. The second and third PCA components result in 65% and 87% classification accuracy, respectively. This also shows that the popular practice of retaining only the higher-order PCA components in a classification application may result in poor classifier performance. Figure 4.12 confirms that none of the transformed PCA features retain the class discrimination information that was readily available in Feature II.

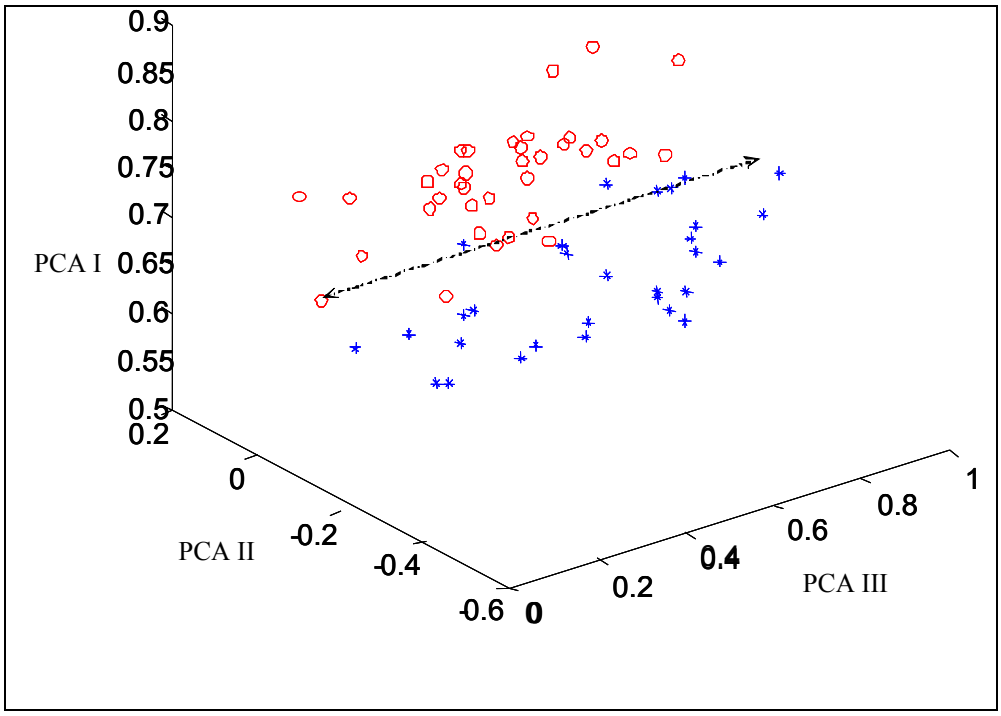

Figure 4.11 Two Class Distribution Projected onto PCA Space.

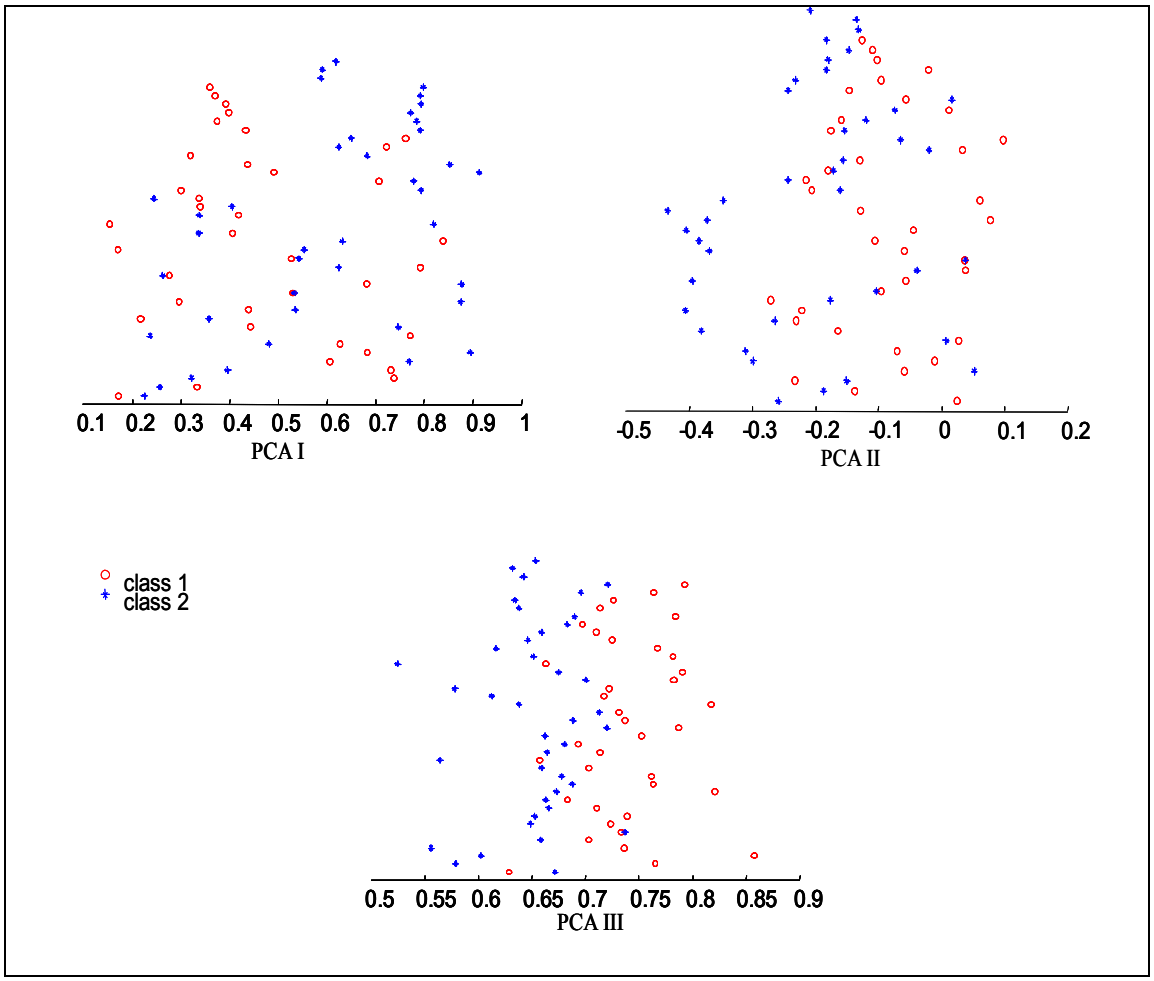

Figure 4.12 Example Data Projected onto Individual PCA components.

The inability of PCA to find the optimal minimum dimension for classification performance becomes more prominent as the number of classes increases. Consider another example wherein an artificially generated two-class data-distribution is plotted on a 3-dimensional feature space as shown in Figure 4.13. Since the maximum variation in the data (see dotted lines) is oriented in a direction that assists class separation, PCA can efficiently extract the single dimension that permits the most accurate classification, while in the original space, none of the individual features are able to separate the classes with high accuracy. This is further verified by performing a maximum likelihood

classification using a leave-one-out testing strategy. Feature I, Feature II, and Feature III obtained a classification accuracy of 83%, 75%, and 71 %, respectively.

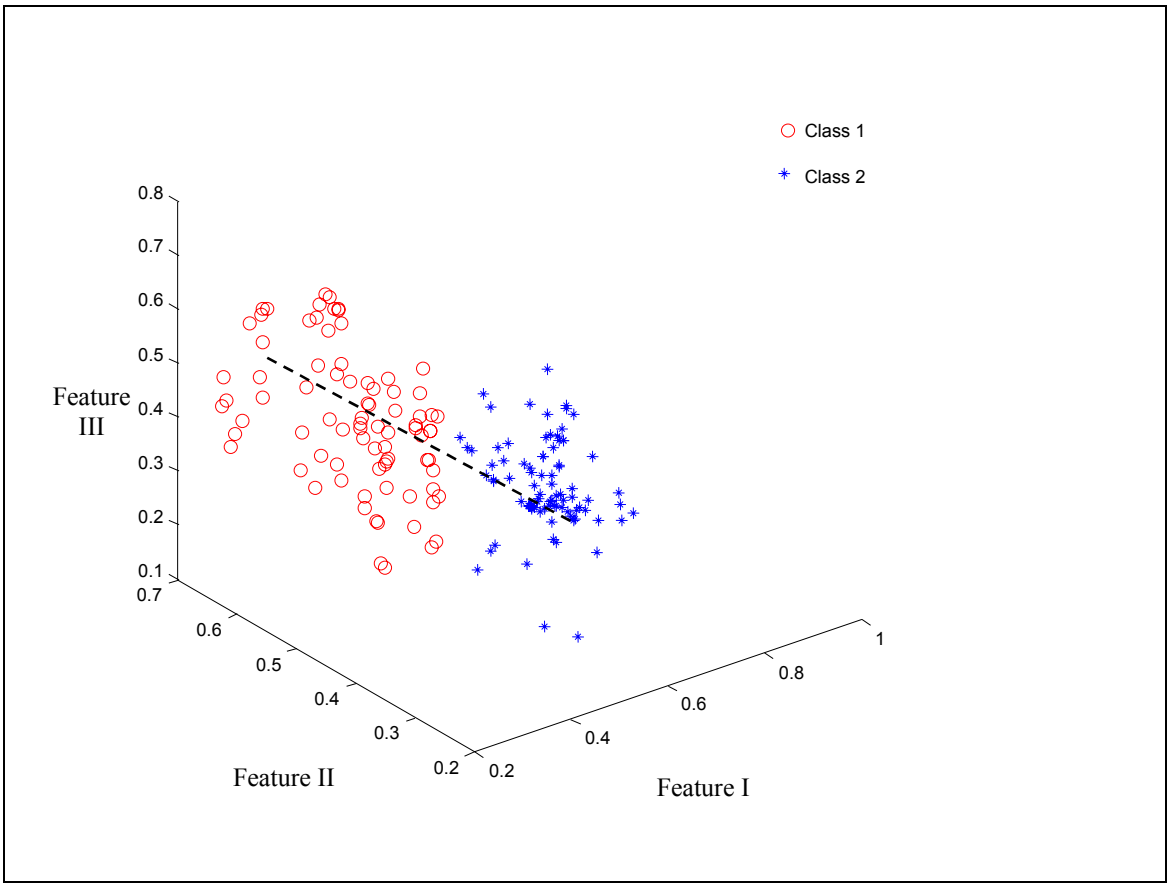

Figure 4.13 An Example Two-Class Distribution.

Suppose the data-distribution is now subjected to PCA. Figure 4.14 shows the data projected onto PCA space. The data projected onto the individual principal components is classified using a maximum likelihood classifier. The PCA I, which is the largest principal component, results in a 97% classification accuracy. In this case, the maximum variance is oriented in a direction that will enhance class separation when subjected to PCA. Hence, PCA features yielded higher classification accuracy in this case.

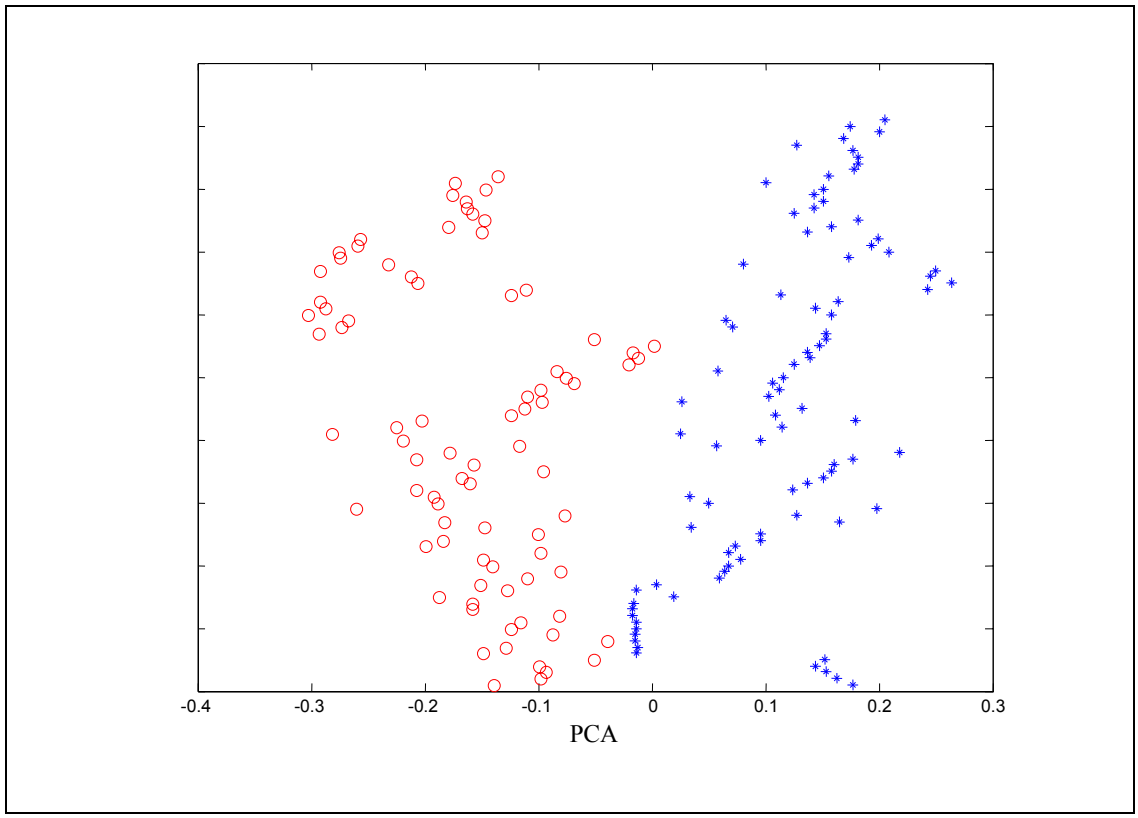

Figure 4.14 Two Class-Distribution projected onto the First Principal Component.

Now consider the case wherein a new class (class 3) is introduced to the existing data-distribution. From Figure 4.15, it is clear that the addition of the new class has changed the direction of maximum variance (see dotted lines) of the total distribution. The three-class distribution shown in Figure 4.15 is subjected to PCA analysis in an effort to find out the lower-dimension with which the classes can be discriminated. Due to the change in the direction of maximum variance, the principal components are now oriented in a new direction as compared to the previous two-class example. Figure 4.16 shows the data-distribution projected onto the first principal component.

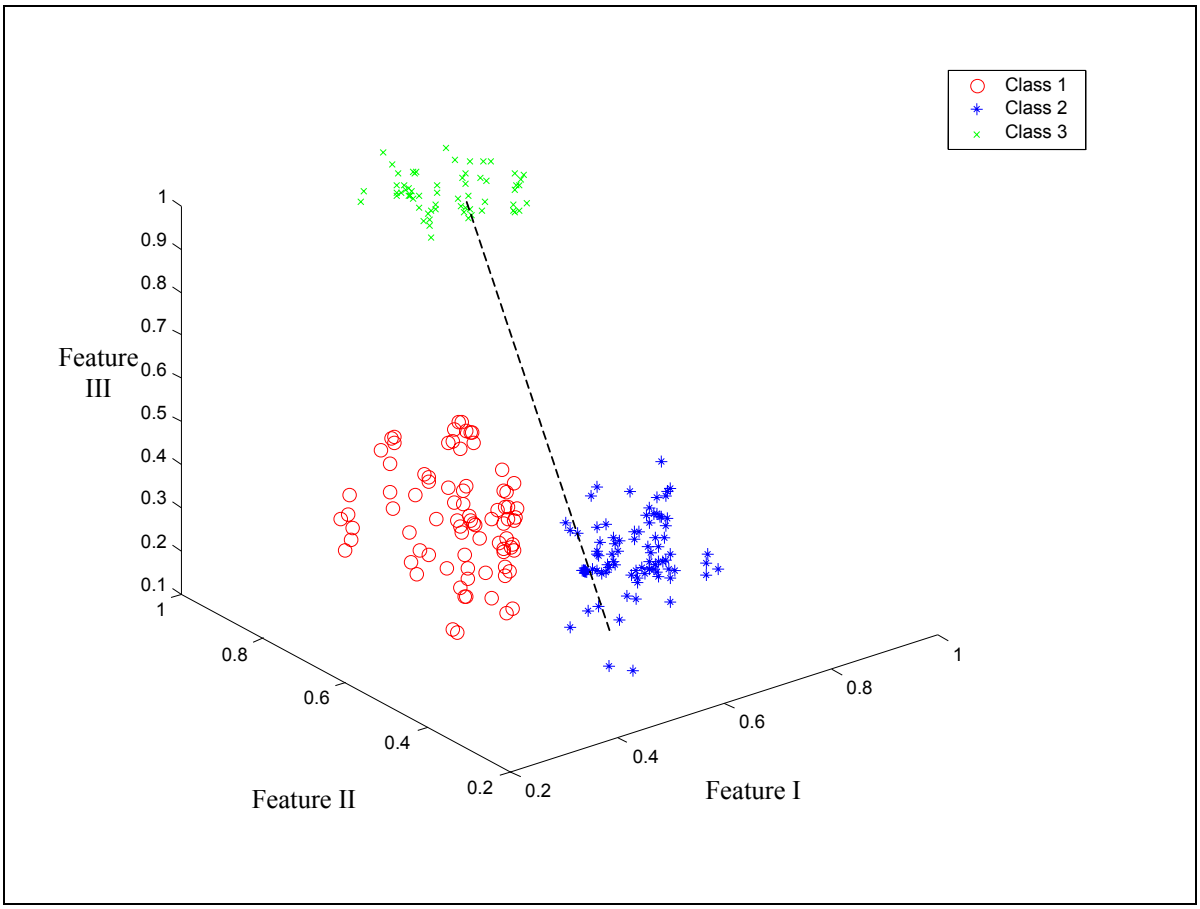

Figure 4.15 An Example Three-Class Distribution.

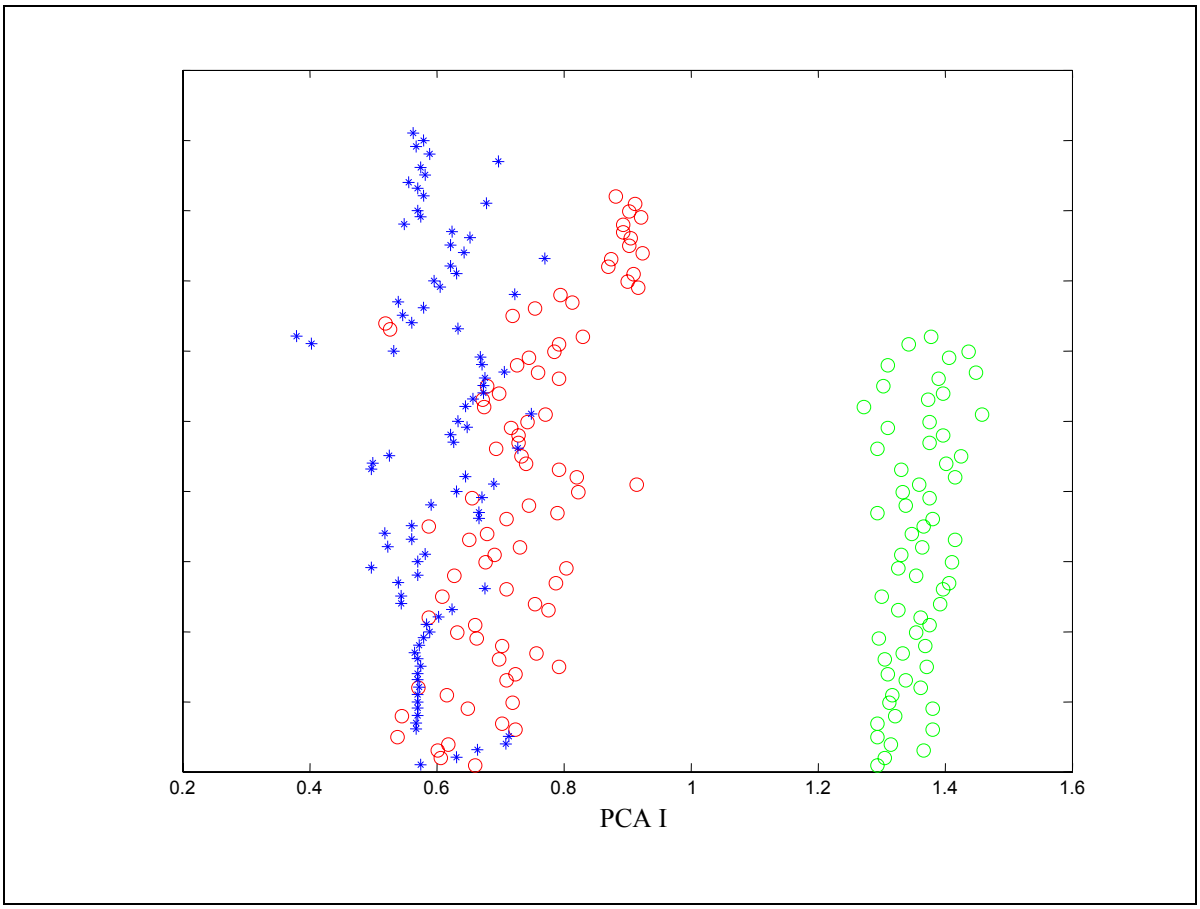

Figure 4.16 Three-Class Distribution Projected onto the First Principal Component.

The data is projected onto the individual principal components and is classified using maximum likelihood classifier with a leave-one out test strategy. The three principal components PCA I, PCA II, and PCA III yielded overall classification accuracies of 85%, 70%, and 48% respectively. The individual class accuracies for the data projected onto the largest principal component (PCA I) that is oriented in the direction of maximum variance are 69%, 89% and 100% for Class 1, Class 2, and Class 3, respectively. In an effort to increase the overall classification accuracy, the second largest principal component is also included as part of the reduced feature set resulting in an improved overall classification accuracy of 94%. In the reduced two-dimensional PCA

space composed of PCA I and PCA II, Class 3 is classified with 100% accuracy and Class 1 and Class 2 are classified with 95% and 89% accuracies, respectively. In this example, PCA I was able to separate Class 3 from Class 1 and Class 2. The obvious choice of the next principal component would be the one, which can separate Class 1 and Class 2. As seen earlier, in the case of a two-class distribution, PCA was able to extract this lower-dimension. In the current scenario, PCA is incapable of extracting this lower-dimension that can perfectly separate Class 1 from Class 2 because of the orthogonal constraints in deciding the orientation of the second principal axis.

In the above three-class case, a greedy-search analysis is conducted on the original space to find the best-feature subset; this approach may prove to be more effective than PCA. It is known that this kind of best-feature-subset search, greedy search, is not computationally efficient when a large number of features are involved. In this case, the greedy search results in a feature subset formed of Feature I and Feature II. The feature subset formed by Feature I and Feature II results in 100% classification accuracy for all the three classes. This shows the limitation of a reduced PCA feature set in classifying the underlying classes with best-possible classification accuracy. In short, in a multi-class data-distribution, PCA may not be able to extract an optimal lower-dimension with which all the classes can be discriminated with the best possible classification accuracy. The introduction of different classes tends to spread the entire data-distribution. In cases wherein classes are more dissimilar, the spreading is more prominent. When the overall data-distribution tends to become more spherical, rotational transforms like PCA will be ineffective against such distributions.

4.4 PCA ANALYSIS ON SYNTHETIC DATASETS

***Synthetic Dataset I*:** From the data-distribution shown in Figure 4.17, it is evident that the maximum variance present in the data is oriented in a direction that is not congruent with the direction in which classes are separated. PCA applied to this data-distribution maximizes the variance present in the data. The principal components obtained from PCA are not appropriate features for the class discrimination as they represent this variance. This fact can be observed in Figure 4.17. It is clear from the data projection on the PCA space that neither of the principal components are better features for class discrimination than Feature I in the original feature space. It is also observed that Feature I in the original feature space, which can be considered as a potential classification feature, is lost in the PCA domain. This illustrates that PCA is incapable of retaining those features that separates the underlying classes in the original feature domain. The incapability of both the transformed features in PCA domain (PCA I and PCA II) to distinguish the underlying classes illustrates that even an exhaustive search algorithm in the case of higher-dimensional PCA feature space may not prove effective.

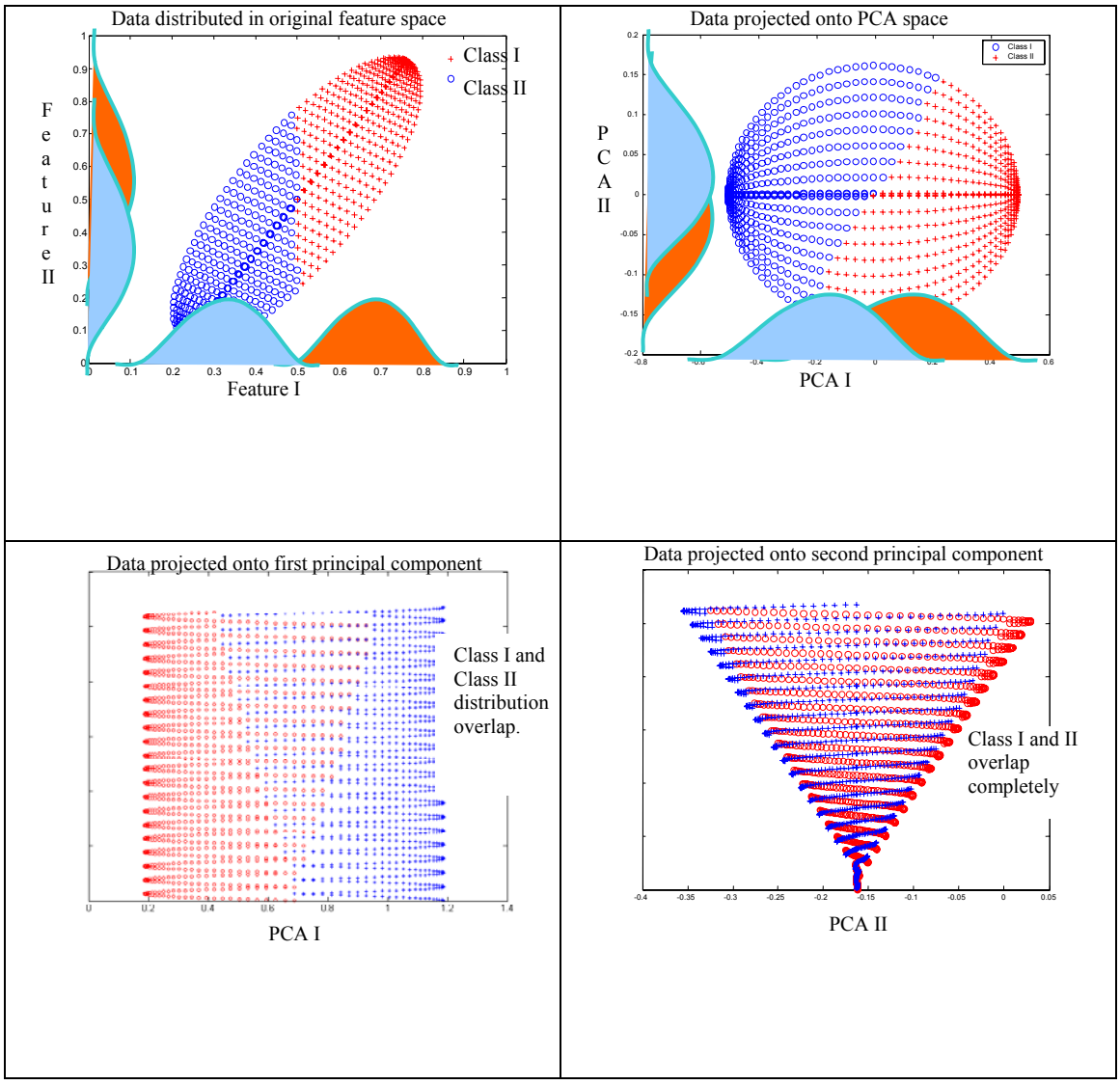

Figure 4.17 PCA Applied to *Synthetic Dataset I*.

Synthetic Dataset II: It is observed from the Figure 4.18 that within-class variance (S_W) present in the data dominates the between-class variance (S_B). PCA applied to such a data-distribution results in higher-order principal components that amplify the within-class variance present in the data. These higher-order principal components when used as features will not aid in classification. Thus, by ignoring the lower-order principal components in order to achieve dimensionality-reduction, some of

the potential classification features that would have contributed to higher classification accuracy are lost. It is seen from the figure that the lower-order principal component is a better feature for class discrimination when compared to the higher-order component. This analysis demonstrates that the popular practice of retaining higher-order principal components and ignoring the lower-order components for dimensionality-reduction purposes may not be the best approach when the goal is target-detection or image classification.

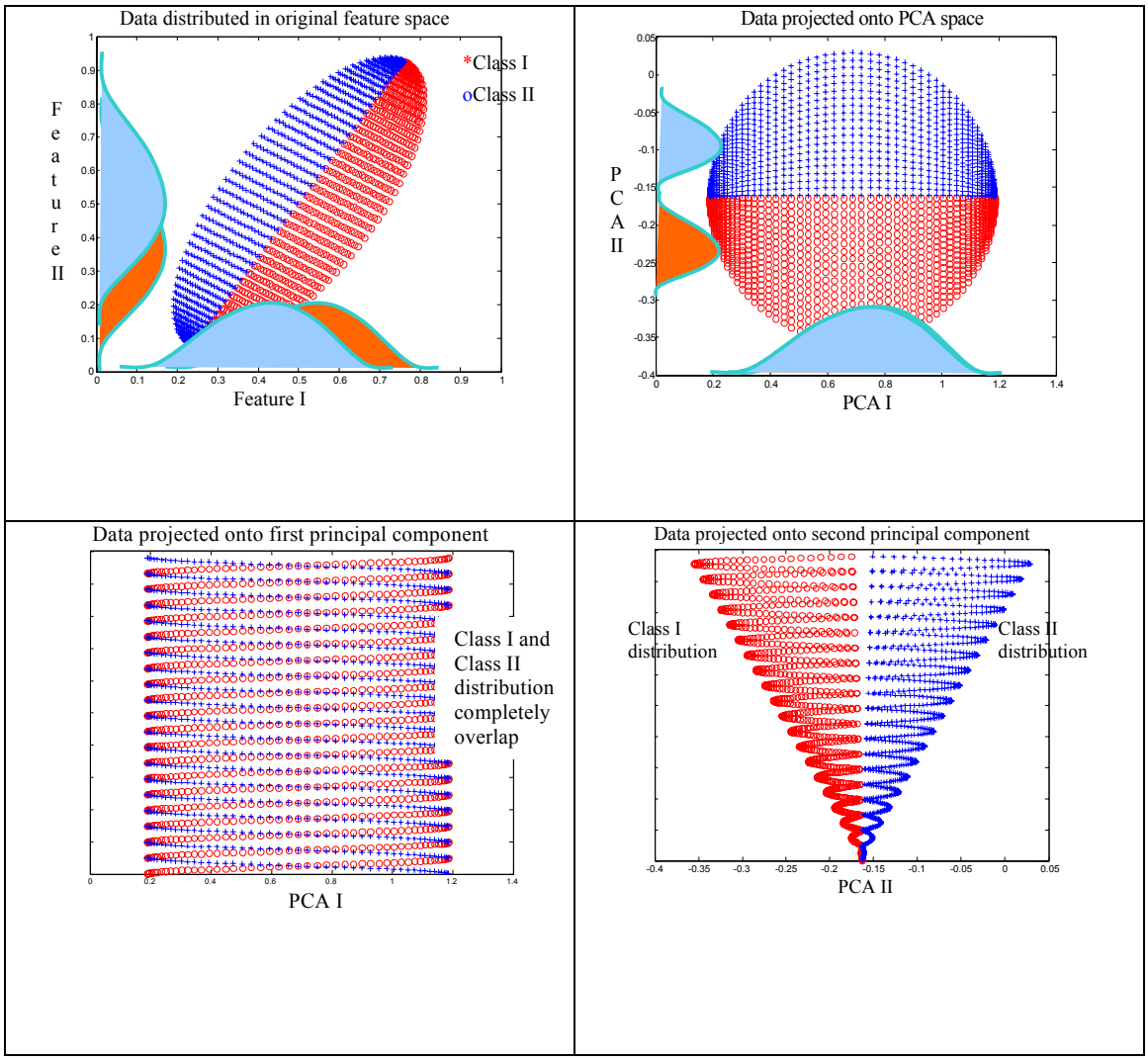

Figure 4.18 PCA Applied to *Synthetic Dataset II*.

4.5 PIXEL UNMIXING USING PCA FEATURES

Previous work on pixel unmixing in [30] investigates how dimensionality-reduction using feature-extraction affects hyperspectral unmixing. Bruce and Li have reported that feature-extraction methods that can reduce within-class variance and increase between-class variance can improve the end-member separability. The analysis found in previous sections of this chapter demonstrates that, for certain data-distributions, PCA features will result in maximizing the within-class variance of the data. Hence, PCA may not be an appropriate feature-extraction method for pixel unmixing.

CHAPTER V

EXPERIMENTAL ANALYSIS – HYPERSPECTRAL DATA REDUCTION

In the initial sections of this chapter, a detailed description on different feature-extraction techniques that are implemented for the experimental analysis is given. The classification accuracy of features extracted using these techniques are compared to the classification accuracy obtained using PCA features. In the later sections, classification performance of features extracted from different spectral regions is compared to PCA features.

5.1 METHODOLOGY

The metrics and discrimination functions used in the feature-extraction algorithms are explained below.

5.1.1 Receiver-Operating-Characteristic (ROC) Curves

Receiver-operating-characteristic (ROC) curves are used to evaluate the class-discrimination capability of features [35]. The area under the ROC curve represents the amount of overlap between the two classes for the feature under investigation. A ROC curve area of 0.5 denotes full overlap between the two features, which is the worst condition for class discrimination, and area of 1.0 denotes no overlap between the features, which is the ideal condition for class discrimination. Thus based on the

ROC values, individual features or combined-feature subsets can be ordered according to their class-separation capability.

5.1.2 Linear Discriminant Analysis (LDA)

LDA is a transformation method employed here to linearly combine features such that their between-class variance is maximized and their within-class variance is minimized. ROC curves can be used for only a single-dimensional feature set. Hence, for evaluating the class-discrimination capability of multidimensional feature sets, LDA is used to linearly combine the multidimensional feature set into a single-dimensional feature. The weights for the linear combination are formed from the C-1 eigenvectors (where C represents the number of classes) obtained from the eigen decomposition of the between-class covariance matrix and within-class covariance matrix ratio as described in section 2.4.

5.1.3 Battacharya Distance

The Battacharya distance (BD) is a statistical distance measure that is used to assess the discrimination capability of multidimensional features. It is reported in [4] that BD is good predictor of classification accuracy. The BD is defined as

$$BD = \frac{1}{8}(\vec{m}_1 - \vec{m}_2)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (\vec{m}_1 - \vec{m}_2) + \frac{1}{2}\frac{\ln\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}, \qquad (11)$$

where $\vec{m}_i$ and Σ_i are the mean vector and covariance matrix for class ω_i respectively.

5.1.4 Maximum-Likelihood Classifier

Maximum-likelihood is a conceptually simple yet powerful parametric classifier. The statistical parameters of each class, namely sample mean and covariance, are

determined using training data. The maximum-likelihood classifier assumes a Gaussian distribution. Hence, the probability density function for a sample vector $\vec{x}$ is given by

$$P(\vec{x}/\omega_i) = \frac{1}{\sqrt{(2\pi)^J |\Sigma_i|}} e^{-1/2(\vec{x}-\vec{m}_i)^T \Sigma_i^{-1}(\vec{x}-\vec{m}_i)}, \quad (12)$$

where $\vec{m}_i$ and Σ_i are the mean vector and covariance matrix of class ω_i respectively, J is the dimensionality of the feature space, and $\vec{x}$ is the test vector.

The discrimination function for the test vector $\vec{x}$ is given by

$$g_i(\vec{x}) = \ln\left(P(\vec{x}/\omega_i)\right), \quad (13)$$

$$g_i(\vec{x}) = -\frac{1}{2} J \ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| + (\vec{x}-\vec{m}_i)^T \Sigma_i^{-1}(\vec{x}-\vec{m}_i), \quad (14)$$

The first part of equation (14), $-\frac{1}{2} J \ln(2\pi)$, is a constant. The second part of equation (14), $\frac{1}{2}\ln|\Sigma_i|$, represents the size and shape of the cluster. The third part of equation (14), $(\vec{x}-\vec{m}_i)^T \Sigma_i^{-1}(\vec{x}-\vec{m}_i)$, represents the statistical distance between the sample vector and the mean of the cluster. The sample vector $\vec{x}$ is assigned to the class ω_i for which (13) is maximized.

5.1.5 Leave-One-Out Testing Method

The reason behind adopting the leave-one-out testing strategy is the limitation of training data when compared to the number of features involved. In the leave-one-out testing method, all of the training data except one is used for estimating the classification parameters for the maximum-likelihood classifier. The left-out data is classified based on the classification parameters. This process gives maximum advantage with limited training data. This is an iterative process that is completed when all the individual

samples in the data have been removed and then classified with the newly trained classifier. The accuracy is calculated as the percentage of test data that is correctly classified.

5.2 FEATURE-EXTRACTION METHODS

In this section, different feature-extraction methods that are used with experimental hyperspectral data are described.

5.2.1 Unsupervised KLT or PCA

PCA is the basic form of Karhunen-Loève transform (KLT), wherein class-specific distribution statistics are not used in the construction of the transformation matrix. The eigenvectors obtained from the eigen decomposition of the total sample covariance matrix forms the transformation matrix. The eigenvectors are arranged in the order of decreasing eigenvalue magnitudes. In fact, PCA can be considered as the unsupervised form of KLT (KL1). The unsupervised KLT matrix A is given by (4). The data is mapped onto the PCA domain by subtracting the data from its mean value and then multiplying it with the transformation matrix *A*. This is given by

$$\vec{Y}_i = A(\vec{x}_i - \vec{m}), \tag{15}$$

where $\vec{Y}_i$ is the transformed signal, $\vec{x}_i$ is ith original signal, and $\vec{m}$ is the mean of the original-signal distribution. One of the main drawbacks of PCA is that it considers the statistical distribution parameters such as mean and covariance of the total data-distribution in its analysis rather than class-specific data-distributions.

5.2.2 Transformation Based on Within-class information (KL2)

KL2 is an alternative approach to extracting discriminating features using Karhunen-Loève transforms [31]. The major difference between PCA and other variants of the KLT is that the class specific statistics are used in the construction of the KLT matrix. In this case, the average within-class covariance matrix is used in the estimation of the transformation matrix. The eigenvectors obtained from the eigen decomposition of the average within-class covariance matrix forms the transformation matrix. As reported by Chien and Fu, in this case, the eigenvectors are arranged in the order of increasing magnitude [32]. As suggested by Tou and Hedron, the transformation matrix is formed from eigenvectors corresponding to the lowest eigenvalues [33]. The rationale behind selecting smallest eigenvalues is that the uncertainty caused by the within-class variance is lowered in the features extracted as a result of this transform. It is reported in [31] that this method will not guarantee features having higher discrimination capability when used for multi-class feature analysis. The average S_W is given by $\tilde{\Sigma}$

$$\tilde{\Sigma} = \sum_{i=1}^{C} P(\omega_i)(\vec{x}_i - \vec{m}_i)(\vec{x}_i - \vec{m}_i)^T \,, \tag{16}$$

where C denotes the total number of classes, ω_i denotes the ith class, $\vec{x}_i$ denotes the sample from ith class, and $\vec{m}_i$ denotes the mean of the ith class.

The eigen decomposition of $\tilde{\Sigma}$ is given by

$$\tilde{\Sigma}\vec{a}_i = \lambda_i \vec{a}_i \,, \tag{17}$$

where $\vec{a}_i$ denotes the i[th] eigen vector, λ_i denotes the corresponding eigenvalue, and *i* varies from 1 to *k* where *k*, is the original dimension of data space. In this case, the KLT matrix, A, is given by

$$A = [\vec{a}_1 \mid \vec{a}_2 \mid ..... \mid \vec{a}_J], \tag{18}$$

where $\vec{a}_1, \vec{a}_2, .., \vec{a}_J$ are the eigenvectors associated with the *J* smallest eigenvalues obtained from the eigen decomposition of the average within-class covariance matrix $\tilde{\Sigma}$. Since the prior probabilities of the classes cannot be estimated for (16), all the classes are given equal probability.

5.2.3 Ordering Features Based on Entropy (KL3)

In the KL3 method, the transformation matrix is constructed from the average within-class covariance matrix $\tilde{\Sigma}$ (16). This method differs from the previous KLT variants in the way in which the features are arranged. The features are arranged based on decreasing entropy. The objective behind this kind of feature-ordering strategy is the same as in the previous method, namely to extract features with less uncertainty. The variance of the feature *j* for class ω_i weighted by the prior probability of class ω_i is given by

$$\lambda_{ij} = P(\omega_i)\vec{a}_j^T \Sigma_i \vec{a}_j, \tag{19}$$

where Σ_i is the covariance matrix for class ω_i.

The entropy for the j[th] feature in the KLT space is given by

$$H_j = -\sum_{i=1}^{C} \frac{\lambda_{ij}}{\lambda_j} \log\left(\frac{\lambda_{ij}}{\lambda_j}\right) \tag{20}$$

where λ_j is defined as

$$\lambda_j = \sum_{i=1}^{C} \lambda_{ij} \tag{21}$$

In this case the KLT matrix, A, is given by

$$A = [\vec{a}_1 \mid \vec{a}_2 \mid ..... \mid \vec{a}_J], \tag{22}$$

where $\vec{a}_1, \vec{a}_2 ... \vec{a}_J$ are the eigenvectors associated with J lowest H_j values computed as in (20).

5.2.4 Greedy Search on DWT Coefficients using ROC analysis (DWT+ROC+ LDA)

This is an algorithm to find the best subset of wavelet coefficients by combining them into one feature (since there are only two classes) using LDA and by using ROC [35] as a metric to measure class-discrimination capability. A detailed description on this algorithm can be found in [7]. The algorithm is briefly explained below:

Step 1: A 6-level wavelet decomposition is performed on the 100-band hyperspectral signal. The wavelet coefficient having the highest class separation is chosen based on the maximum ROC value. This is termed as the best wavelet coefficient.

Step 2: The next-best wavelet coefficient is determined by combining the remaining wavelet coefficients independently with the previous best wavelet coefficient/coefficients using LDA. The class separation of the combined wavelet coefficients is measured using their ROC value. The second-best wavelet coefficient is the coefficient which when combined with the first-best wavelet coefficient, results in the greatest increase in the ROC value obtained in Step 1.

Step 3: The next-best wavelet coefficients are determined by repeating Step 2. The iteration can be stopped when there is no significant increment in ROC value or when a predetermined level of dimensionality-reduction is achieved.

5.2.5 Greedy Search on Spectral Bands using ROC analysis (ORG+ROC+ LDA)

This feature-extraction method is similar to the algorithm described in Section 5.2.4 with the exception that the greedy search is performed on the original spectral domain. The feature set is formed from applying LDA to the spectral-band combinations. The subset of spectral bands that yields the highest-class separation is chosen based on the maximum ROC value.

5.2.6 Greedy Search on Spectral Bands using BD metric (ORG+ROC+LDA)

This method is similar to the one described in Section 5.2.5. The only difference is that the BD is used as the class-separation metric instead of ROC.

5.3 TWO-CLASS ANALYSIS

The hyperspectral data collected for this experimental analysis consists of two vegetative classes, namely the herbaceous class and the woody class. The data is collected using a hand-held sensor, specifically an Analytic Spectral Devices (ASD) spectroradiometer [34] .The spectral range of an ASD is 350nm to 2500nm. The data consists of reflectance values calculated at 2151 spectral bands for the entire spectral range of the sensor. The rationale behind choosing the above classes for investigating PCA analysis is because the classes consist of different vegetative species and hence have larger within-class variance. It is predicted that the maximum variance exhibited by the data is not in a direction that will assist class discrimination. Hence, as we have seen in the theoretical assessment of PCA, the transformation may tend to project the data in a direction that may not facilitate extraction of better classification features.

The test data set consists of 211 samples with 146 samples collected for the herbaceous class and 65 for the woody class. Due to limitations in training data, and in

order for the covariance matrix generated from this limited training set to be a better estimate of the underlying class distribution, the initial spectral feature set of 2151 bands is reduced to 100 bands by performing a moving average on the original spectral signal. Different feature-extraction methods are applied on this 100-dimensional space and comparison of the results is performed. The mean spectrum for each of the two vegetative classes is shown in Figure 5.1.

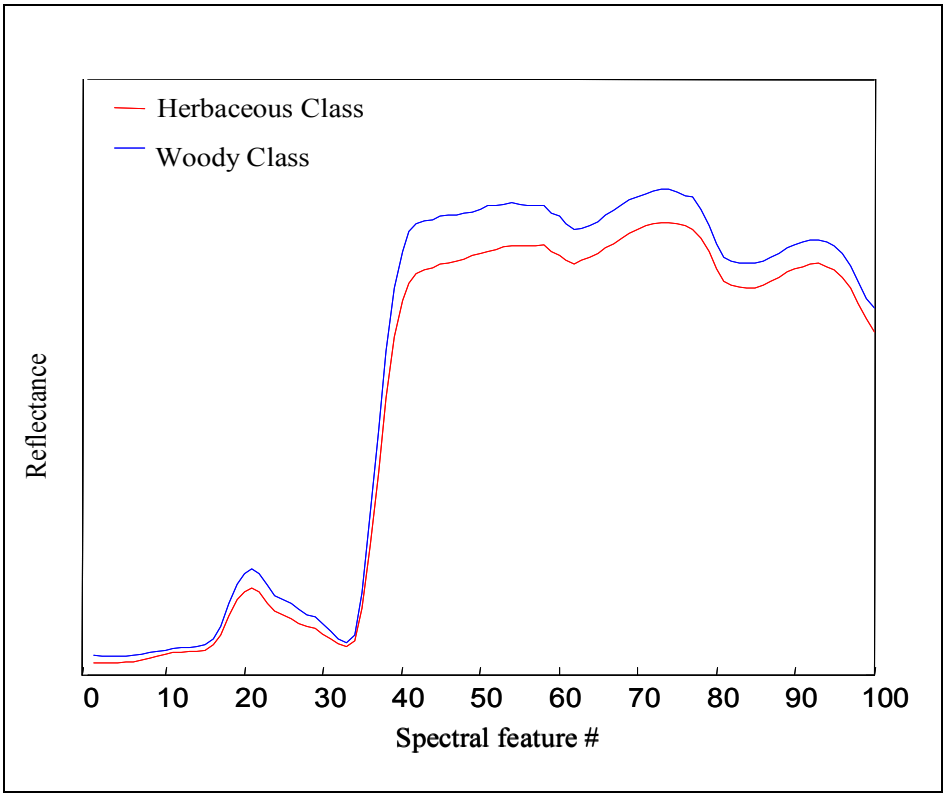

Figure 5.1 Mean Spectral Signatures for Herbaceous and Woody Classes.

5.3.1 Comparison of Feature-extraction Methods

The hyperspectral signatures are subjected to different feature-extraction methods. The classifications are conducted using the extracted features. The results for various feature methods are compared. The features are classified using a maximum-likelihood

classifier. The testing is done using a leave-one-out test strategy in order to compensate for the limited training data. The different feature-extraction methods include PCA, variants of KLT described in Chapter III, greedy search using DWT coefficients [7], and greedy search using original spectral features [6]. The greedy-search algorithm used in this study employs a forward feature-selection strategy to combine the best features.

5.3.2 Results and Discussion

The maximum-likelihood classification results obtained for the above discussed feature-extraction algorithms are shown in Figure 5.2. It can be seen from Figure 5.2 that feature-extraction methods such as greedy search on DWT coefficients (DWT+ROC+LDA), KL2 and KL3 outperform PCA significantly when the number of features are less than 10. The poor classification accuracy of PCA is due to the fact that the top components obtained using PCA represents the within-class variance present in the data rather than the between-class variance. Note that the greedy-search method for best-band combination on the original spectral space (ORG+ROC+LDA) yielded the best classification accuracies as the number of features is increased to thirty.

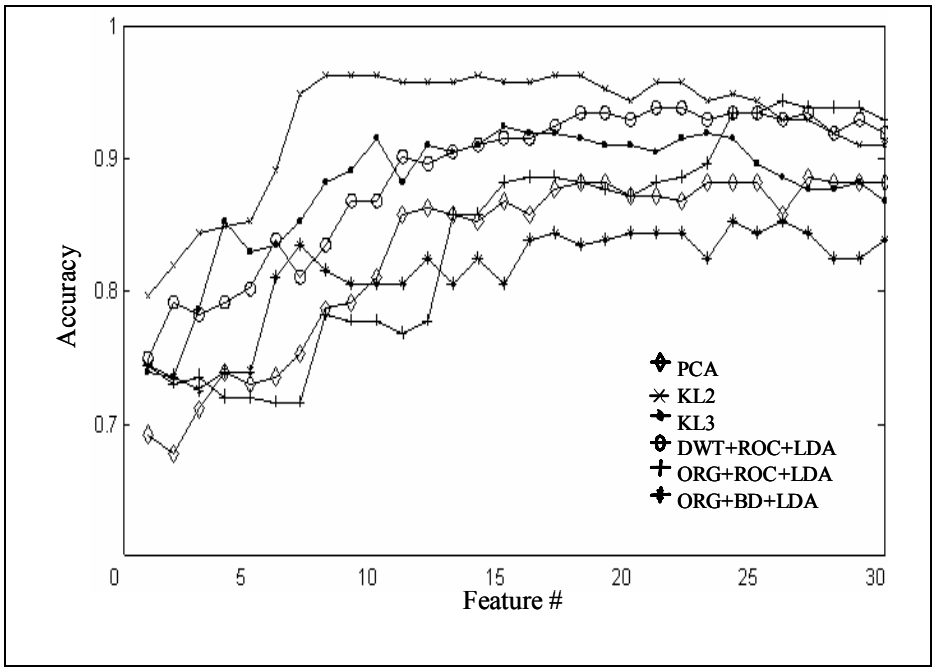

Figure 5.2 Comparison of Classification Accuracies for Different Feature-extraction Methods.

5.4 MULTI-CLASS ANALYSIS

An example database consisting of seven different classes of hyperspectral signatures ranging from 350nm to 2500nm are used here. The data is collected using a hand-held sensor, specifically an Analytic Spectral Devices (ASD) spectroradiometer [34]. The data consists of leaf-reflectance values calculated at 2151 spectral bands. Some of the spectral-band data between 350nm and 355nm had to be discarded as they were corrupted by noise. The original spectral data is first reduced by averaging the bands according to the Hyperion spectral-resolution specification [1]. This is done in order to account for the limited number of training samples. Also by simulating the Hyperion data, the analysis can be used to predict how PCA would perform on satellite hyperspectral image. The resulting 196 band hyperspectral data is used in the experimental analysis. The hyperspectral data consists of one agricultural crop and six

weed species. The mean spectrum of each of the seven classes is shown in Figure 5.3. The experimental data information for the seven classes is given in Table 3.

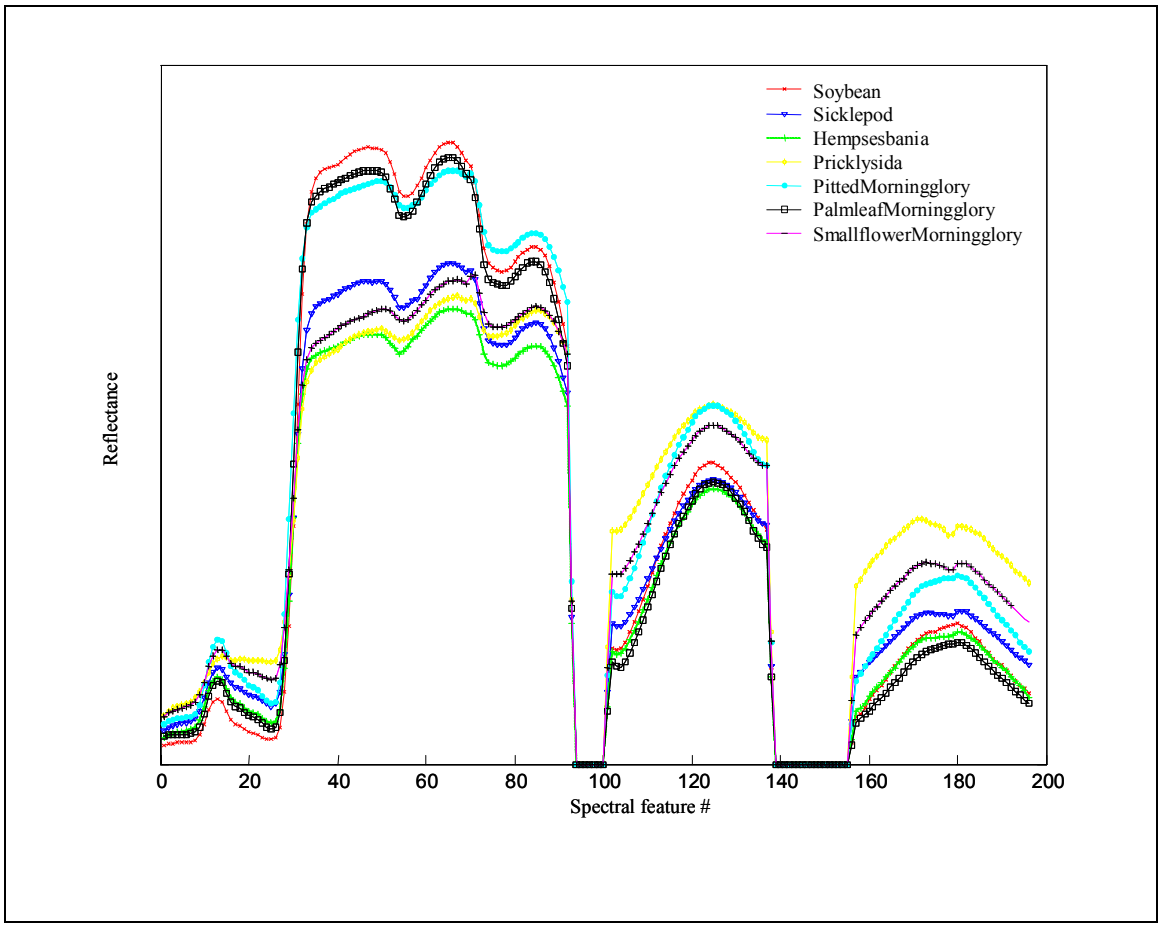

Figure 5.3 Mean Spectral Signature for the Seven Classes.

Table 3

SEVEN CLASS EXPERIMENTAL DATA FOR COMPARISON ANALYSIS

Class	Number of Training Samples	Number of Testing Samples
Soybean	31	93
Sicklepod	24	72
Hempsesbania	46	141
Pricklysida	24	72
PittedMorningglory	23	72
PalmleafMorningglory	24	72
SmallflowerMorningglory	22	66
Total	194	588

5.4.1 Comparison of Feature-extraction Methods

The seven-class hyperspectral dataset for the multi-class experiment is analyzed using different feature-extraction methods. The classification is performed using the extracted features. The results of the various feature-extraction methods are compared. The features are classified using a maximum-likelihood classifier. The different feature-extraction methods include PCA, KL3, and greedy search implemented on the original spectral band domain using BD as the discrimination metric (ORG+BD+LDA). KL2 and greedy search using DWT coefficients are not compared here, as their extracted features in the dimensional range of 5 to 20 resulted in singular matrices for maximum-likelihood computations.

5.4.2 Results and Discussion

The comparison of classification accuracies obtained using different feature-extraction methods is shown in Figure 5.4. The classification results show that PCA performed poorly when compared to the greedy-search implemented on the original spectral band (ORG+BD+LDA). The classification accuracy of the KL3 feature-

extraction method is poor compared to PCA and ORG+BD+LDA. It can be observed that, as the number of features is increased (greater than 11), the classification accuracy of KL3 decreases. This can be attributed to the unreliable estimation of second-order statistics such as average within-class covariance using limited training data. The poor performance of PCA features when compared to features extracted using greedy search supports the argument that features extracted using PCA may not be appropriate for classification of hyperspectral data.

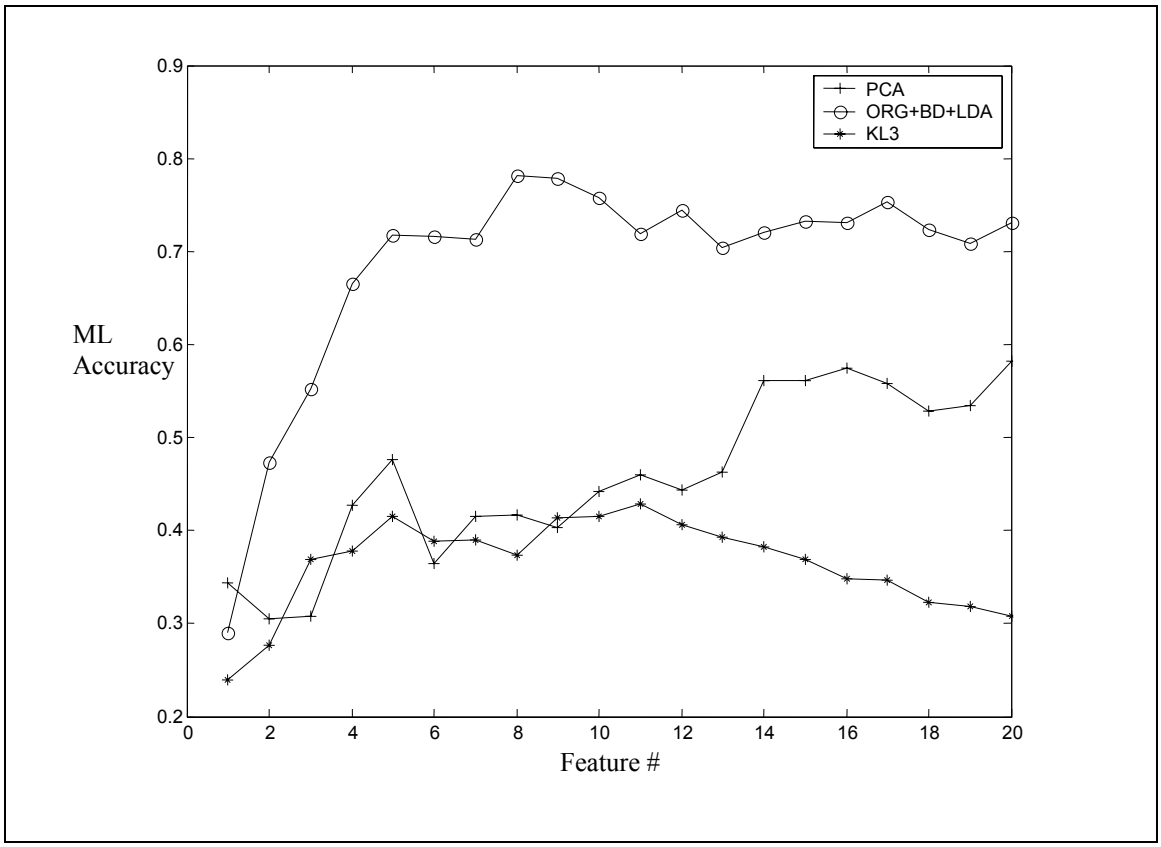

Figure 5.4 Comparison of Classification Accuracies.

5.5 Comparison of PCA Features with Features Extracted from different Spectral Regions

The large variance exhibited by certain spectral regions of the hyperspectral data influence PCA, resulting in higher-order principal components that are oriented in a direction which may not be appropriate for class separation. In order to understand the impact of this effect on the extraction of classification features, a study is done on the performance of classification features extracted from different spectral regions of the hyperspectral data. In Figure 5.5, it can be observed that experimental hyperspectral data exhibits large variance in the near infrared and far infrared regions compared to the visible part of the spectrum.

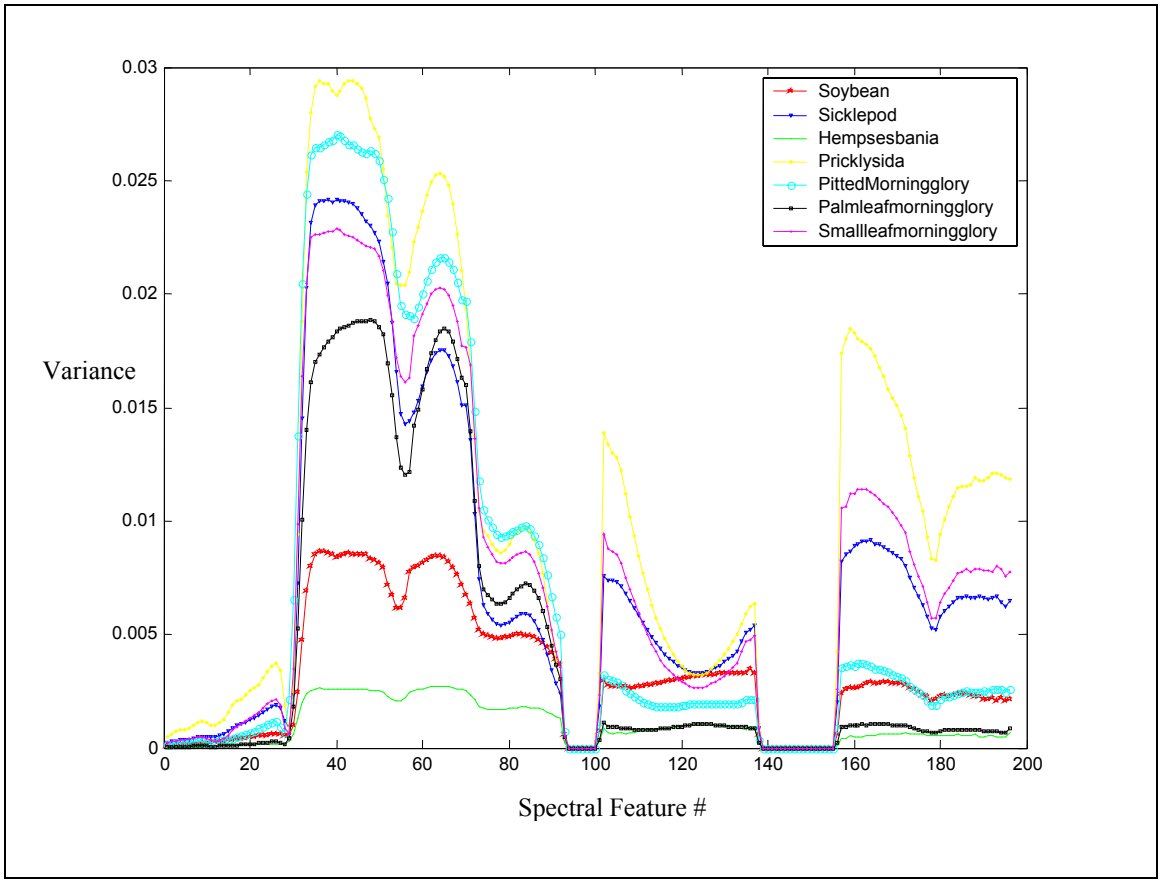

Figure 5.5 Spectral Variances for the Seven Classes.

In order to extract features from different spectral regions, the spectral data is partitioned into separate groups based on inter-band correlation and statistical separation between the underlying classes. The inter-band correlation for bands *i* and *j* is given by *q*

$$q_{ij} = \frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \tag{23}$$

where Q is the interband covariance matrix. The hyperspectral group correlation *C* is defined as the minimum correlation for any band pair within the group,

$$C_n = \min(q_{ij}), \tag{24}$$

where *n* is the group number.

The statistical separation between the underlying classes is measured using BD. The groups are partitioned based on maximizing a function which is the product of the hyperspectral group correlation and minimum BD of all the class pairs for that hyperspectral group. The function for group *n* is given by D_n

$$D_n = C_n \min(BD_n) \tag{25}$$

A detailed description on the group-partitioning algorithm can be found in [14]. The group-partitioning algorithm results in 14 groups, with maximum number of features in each group limited to 20. Four groups that demonstrated relatively better classification performance compared to other groups is shown in Figure 5.6. These four groups are referred to as group 1 group 2, group 3, and group 4. The features extracted from each of the four groups are classified using a maximum-likelihood classifier.

PCA transformation results in projecting the original data onto a space that is oriented in a direction that is aligned with the maximum variance present in the data. Hence, it is possible that the higher-order principal components generated by the linear

combination of spectral bands may have less contribution from the spectral bands pertaining to the visible region as they have less variance compared to other parts of the spectrum. The weights for the linear combination are merely determined by the variance present in the spectral band and not their discrimination ability. From a classification perspective, this is not promising. The classification performance of features extracted using PCA and features extracted from the other part of the spectrum is compared.

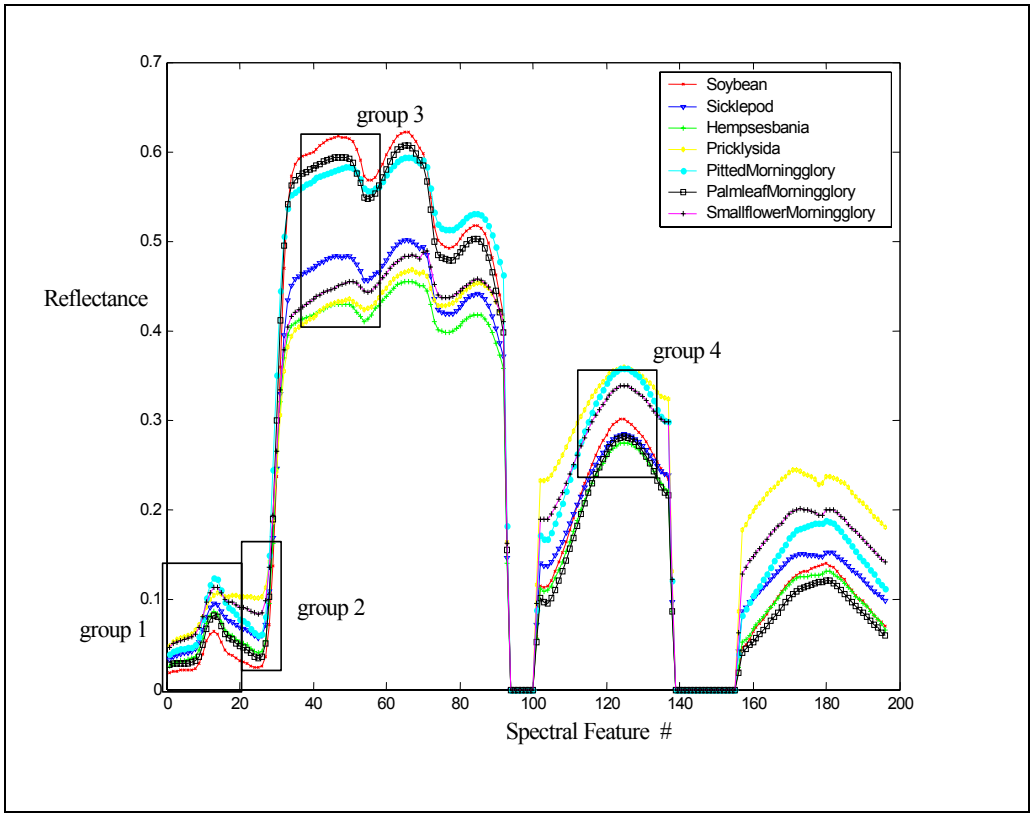

Figure 5.6 Best Groups (overlaid rectangles).

5.5.1 Results and Discussion

The classification performance of features extracted using PCA is compared to features extracted from different spectral regions of the hyperspectral data. Nineteen features are extracted from the visible part of the spectrum (group1) and twenty features

are extracted from the whole spectrum using PCA. It can be observed from Figure 5.7 that features extracted from the visible part of the spectrum results in around 76% classification accuracy where as features extracted using PCA results in a classification accuracy of 58%. This supports the fact that the contribution from the visible part of the spectrum in the generation of higher-order principal components is less compared to that of other areas of the spectrum. This is one of the reasons for the difference in classification accuracies obtained using features extracted from PCA and group1. Also it is to be noted that a study on the properties of distributions for lower-dimensional projections from higher-dimensional data by Hall and Li shows that, for high-dimensional data set the lower-dimensional linear projections are nearly normal [37]. The poor classification accuracy result demonstrated by PCA agrees with this study. The poor classification accuracy of the higher-order principal components supports the fact that these components have less contribution from the more discriminating spectral features of the visible spectrum. In order to analyze the contribution of different spectral regions in the generation of higher-order principal components, the eigenvectors associated with the 5 largest eigenvalues are plotted in Figure 5.7. It can be seen from Figure 5.8 that the weights associated with the visible part of the spectrum for all the 5 principal components is smaller than the weights compared to other part of the spectrum, although features from the visible part of the spectrum resulted in a better class discrimination. Hence, it can be concluded that lower-dimensional structure established by PCA may not be appropriate to discriminate between the underlying classes for data-distributions, in the case that the maximum variance in the data is oriented in a direction that is not suitable for class separation.

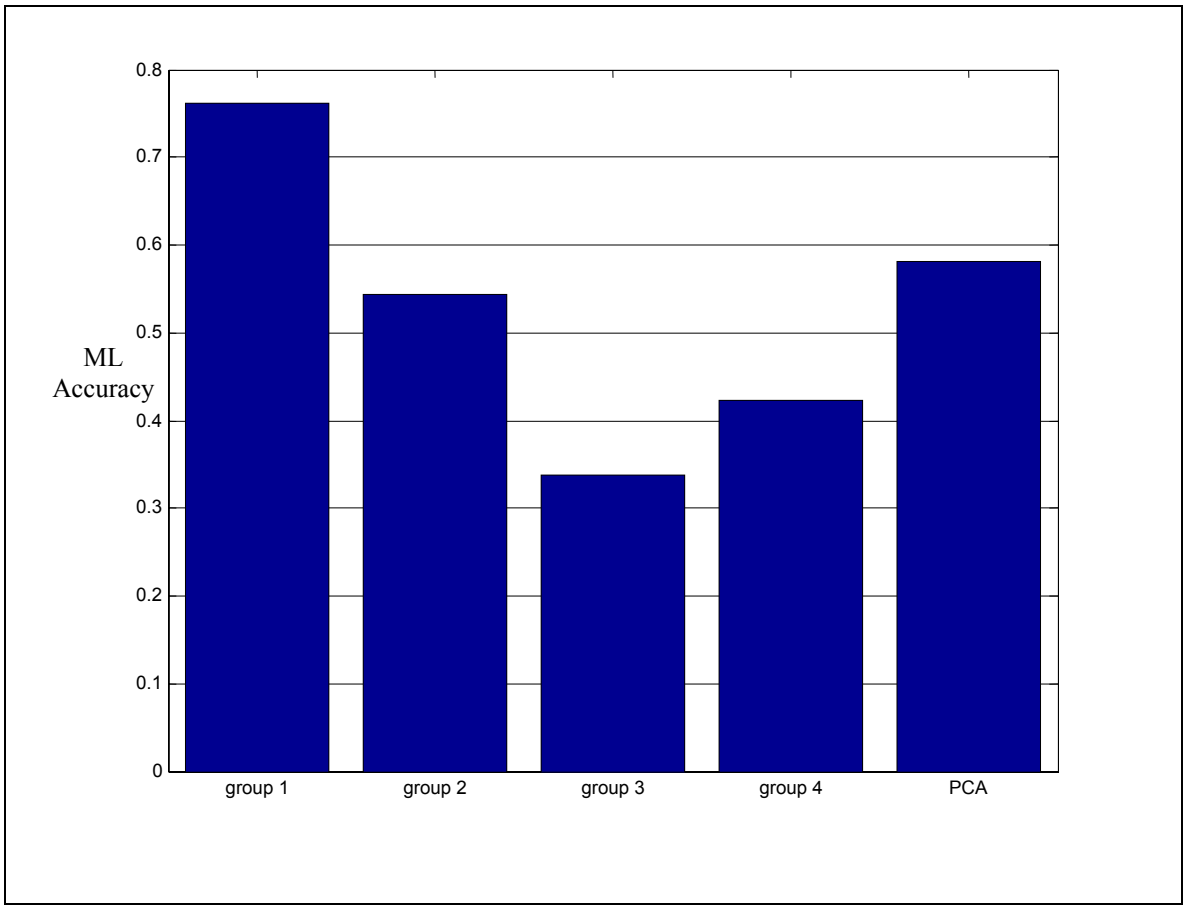

Figure 5.7 Group Classification Accuracies Compared to PCA

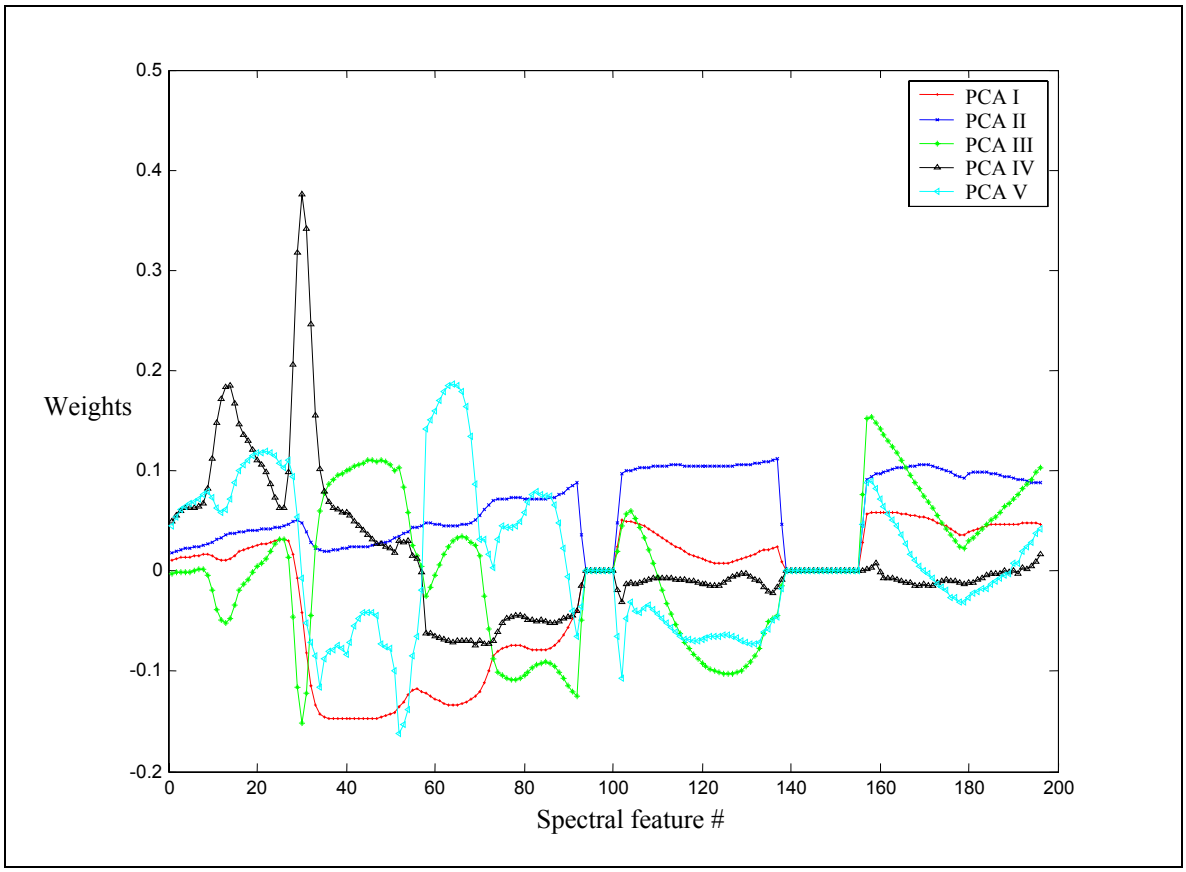

Figure 5.8 Linear Spectral Weights Associated with the 5 Largest Principal Components.

CHAPTER VI

CONCLUSIONS

Advancements made in sensor technology in recent years has paved the way for the research community to conduct intensive and detailed examination of an object or target using remote-sensing. Hyperspectral sensors are an example of this advancement. However, due to the large dimensionality of hyperspectral data, computationally efficient analysis of data is not possible without the use of dimensionality-reduction methods. The limitation regarding the training-data size is another constraint that imposes the dimensionality-reduction requirement. There have been numerous studies done in recent years on developing efficient dimensionality-reduction techniques. Many of the dimensionality-reduction techniques applied for classification were not intentionally developed for classification applications. PCA is one such dimensionality-reduction method that gained popularity because of its ease of use, availability through popular remote-sensing software packages and optimal nature in a mean square error sense. Although there exist numerous dimensionality-reduction techniques, suitability of those techniques from a classification application perspective is seldom considered. Hence, this thesis was focused on analyzing PCA as a feature-extraction method and its usefulness for a classification application. Theoretical and experimental analyses showed that features obtained using PCA may not be suitable for discriminating between underlying

classes. It is also seen that the popular practice of ignoring the lower-order PC components in order to achieve dimensionality-reduction may in fact result in losing some of the discriminatory information present in the data. For data-distributions wherein the within-class variance dominates the between-class variance, the largest principal components will be oriented in a direction that represents these within-class variances and hence may not be useful from a classification standpoint. The two-class and multi-class experimental results demonstrate that PCA is not the best feature-extraction method for hyperspectral data. The comparison of features obtained from different individual spectral regions with PCA features show that features obtained from certain spectral regions performed better than PCA features in classifying the underlying classes. This is ascribed to the fact that, for certain data-distributions, higher-order principal components are not be oriented in a direction that best discriminates the classes. It is also seen that the formation of higher-order principal components is influenced by the large variations present in certain spectral regions of the data, although these spectral regions may not contribute to the discrimination of the classes. The experimental and theoretical analyses presented in this thesis demonstrate that PCA may not be the appropriate feature-extraction technique when the objective is classification or target-recognition.

In order to minimize the influence of large spectral variations in the extraction of PCA features, PCA can be applied to different spectral groups formed by combining the adjacent bands in the original hyperspectral space. The grouping may be performed based on class-separation criteria. The features obtained from these spectral subgroups can be further fused together at a feature level, or the classification results of these spectral

subgroups can be fused together at a decision level. This is a potential for future research area.

REFERENCES

[1] S. Craman, "Hyperion Grating Imaging Spectrometer," TRW Space &Electronics, http://eo1.gsfc.nasa.gov/overview/workshop/06.pdf .

[2] AVIRIS Image Cube, http://aviris.jpl.nasa.gov/html/aviris.cube.html .

[3] A.M. Martinez and R. Benavente, "The AR Face Database," *CVC Technical Report #24*, June 1998.

[4] D. A. Landgrebe, "Information Extraction Principal and Methods for Multispectral and Hyperspectral Image Data," *Information Processing for Remote Sensing*, edited by C.H. Chen, World Scientific Publishing 2000.

[5] C. Lee and D. A. Landgrebe, "Analyzing High-dimensional Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 31, no. 4, pp. 792-800, July 1993.

[6] A. Mathur, L.M. Bruce and J. Byrd, "Discrimination of Subtly Different Vegetative Species via Hyperspectral Data," in *IEEE Proc. Geoscience and Remote Sensing Symposium*, vol. 2, pp. 805-807, June 2002.

[7] L.M. Bruce, C.H. Koger and J. Li, "Dimensionality-reduction of Hyperspectral Data using Discrete Wavelet Transform Feature-extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 10 pp. 2318-2338, October 2002.

[8] C. Lee and D. A. Landgrebe, "Feature-extraction Based on Decision Boundaries," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 15, no. 4, pp. 388-400, April 1993.

[9] S. Kumar, J. Ghosh and M.M Crawford, "Best Bases Feature-extraction Algorithms for Classification of Hyperspectral Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1368-1379, July 2001.

[10] L.O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 2653-2664, November 1999.

[11] N. Saito and R. R. Coifman, "Local discriminant bases", in *Mathematical Imaging: Wavelet Applications in Signal and Image Processing II*," Proc. SPIE, vol. 2303, pp. 2–14, July 1994.

[12] L.O. Jimenez, A. Morales-Morell, and A. Creus, "Classification of Hyperdimensional Data Based on Feature and Decision Fusion Approaches using Projection Pursuit, Majority Voting, and Neural Networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1360-1366, May 1999.

[13] J. A. Benediktsson and I. Kanellopoulos, "Classification of Multisource and Hyperspectral Data Based on Decision Fusion," *IEEE Transactions on. Geoscience and Remote Sensing*, vol. 37, pp. 1367-1377, May 1999.

[14] A. Cheriyadat, L. M. Bruce and A. Mathur, "Decision Level Fusion using Best-Bases for Hyperspectral Classification," in *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, Washington, DC, October 2003 (to appear).

[15] S. Kaewpijit, J. Le Mogine and T. El-Ghazawi, "A Wavelet-Based PCA Reduction for Hyperspectral Imagery," in *IEEE Proc*. *Geoscience and Remote Sensing Symposium*, vol. 5, pp. 2581-2583, June 2002.

[16] J. Yang, H. Yu, W. Kunz, "An Efficient LDA Algorithm for Face Recognition," in *International Conference on Automation, Robotics, and Computer Vision (ICARCV'2000)*, Singapore, December 2000.

[17] X. Jia and J. A. Richards, "Segmented Principal Components Transformation for Efficient Hyperspectral Remote Sensing Image Display and Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 1 pp. 538-542, January 1999.

[18] *ENVI User's guide*, ENVI version 3.6, 2nd edition, December. 2002.

[19] *ERDAS Field Guide,* ERDAS IMAGINE version 8.5, 5th edition, 1999.

[20] S. C. Liew, C. W. Chang and K. H. Lim, "Hyperspectral Land Cover Classification of EO-1 Hyperion Data By Principal Component Analysis and Pixel Unmixing," in *IEEE Proc*. *Geoscience and Remote Sensing Symposium*, vol. 6, pp. 24-28, June 2002.

[21] J. M. Preston, A. C. Christney, S. F. Bloomer and I. L. Beaudet, "Seabed Classification of Multibeam Sonar Images," in *IEEE Conference and Exhibition OCEANS 2001*, vol. 4, pp. 5-8, November 2001.

[22] K. Wikantika, S. S. Wihartini, R. Tateishi and A. B. Harto, "Spectral and Textural Information of Multisensor Data for Land Use Classification in

Metropolitan Area," in *IEEE Proc. Geoscience and Remote Sensing Symposium*, vol. 7, pp. 24-28, July 2000.

[23] Chih-Cheng Hung, A. Fashi, W. Tadesse and T. Coleman, "A Comparative Study of Remotely Sensed Data Classification Using Principal Components Analysis and Divergence," in *IEEE Conference on Systems, Man and Cybernetics*, vol. 3, pp. 2444-2449, October 1997.

[24] S. C. Bajic, "Accuracy of a Supervised Classification of the Artificial Objects in Thermal Hyperspectral Images," in *IEEE Proc. Image Analysis and Processing*, pp. 798-803, September 1999.

[25] R. Gonzalez and R. Woods, *Digital Image Processing*, 2nd edition, Prentice Hall Upper Saddle River New Jersey, 2001.

[26] L. O. Jimenez 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, February 1998.

[27] G.F. Hughes, "On the Mean Accuracy of Statistical Pattern Recognizers," *IEEE Transactions on Information Theory*, vol. 14, no 1, January 1968.

[28] ISO 10918-1 (JPEG), "Digital Compression and Coding of Continuous-tone Still Images", 1991.

[29] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A Transformation for Ordering Multispectral Data in Terms of Image Quality with Implications for Noise Removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 26, no. 1 pp. 65-74, January 1988.

[30] J. Li, L.M. Bruce and A. Mathur, "Wavelet Transform For Dimensionality-reduction in Hyperspectral Linear Unmixing," in *IEEE Proc. Geoscience and Remote Sensing Symposium*, vol. 6, pp. 3513-3515, June 2002.

[31] J. Kittler and P. C. Young, "A New Approach to Feature Selection Based on the Karhunen-Loeve Expansion," *Pattern Recognition*, vol. 5, pp. 335–352, May 1973.

[32] Y. T. Chien and K. S. Fu, "On the Generalized Karhunen-Loeve Expansion," *IEEE Trans Information Theory*, vol.15, p. 518, 1967.

[33] J. T. Tou and R. P Heydorn, *Computer and Information Sciences*, vol. II, Academic Press, New York, 1967.

[34] ASD FieldSpec Pro FR portable spectroradiometer, Analytical Spectral Device Inc. 5335 Sterling Drive, Boulder, CO 80301-2344 USA.

[35] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Diagnostic Radiology*, vol. 143, pp 29-36, 1982.

[36] A. Webb, *Statistical Pattern Recognition*, ISBN-0304741643, Oxford University Press, 1999.

[37] P. Hall and K. Li, "On Almost Linearity of Low-dimensional Projections From High-dimensional Data," *The Annals of Statistics*, vol. 21, no. 2, pp. 867-889, 1993.