homework #2

Principle Components Analysis

EE 8993: Fundamentals of Speech Recognition

May 15, 1998

submitted to:

Dr. Joseph Picone

submitted by:

Suresh Balakrishnama

Institute of Information and Signal Processing,
Department of Electrical & Computer Engineering,
Mississippi State University,
MS 39759, USA.

1. Introduction

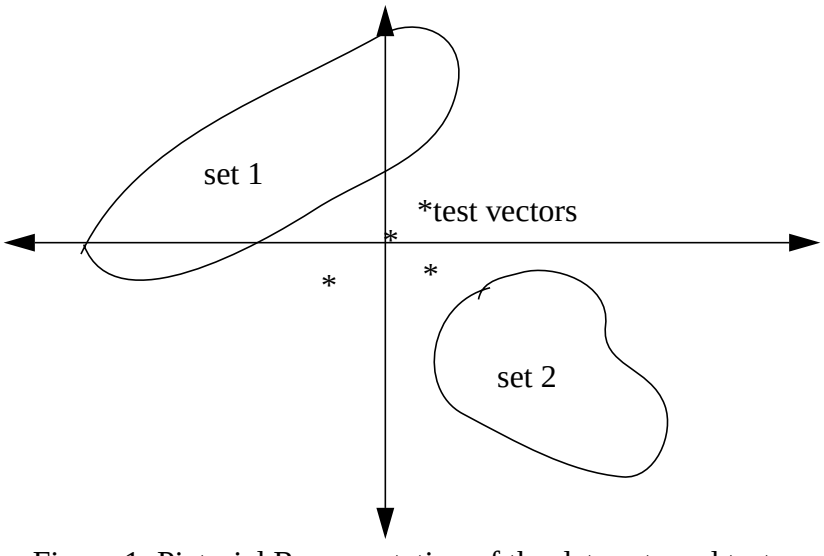
Principle Components Analysis is one method for data compression and dimensionality reduction which can be used for feature extraction and data classification. According to this technique, the first principle component of a sample or test vector is the direction along which there is largest variance over all samples. The approach in this technique is that the direction along which there is maximum variation is most likely to contain the information about the class discrimination. The prime objective is to transform the given data sets into a new space where data discrimination is easier.The Euclidean distance is calculated between the test vectors and the given data sets, based on the minimum distance the test vector is classified to a given set.

2. Problem Description

The data sets, set1 and set2 were given in linear space each consisting of 100 points. The test vectors were also given which are to be classified belonging to either set1 or set2. The condition given was that each data set 1 should have a mean of (-2,2) and data set2 should have a mean of (2, -2). The objective is to transform the data sets into a new space and using the euclidean distance classify the test vectors to one of the two sets.

Data point	Coordinates
x1	-1,-1
x2	0, 0
x3	0.5, 0.5
x4	0.5, -0.5

Table 1: Test vectors to be classified



. Figure 1: Pictorial Representation of the data sets and test vectors

3. Implementation and Results

This classification problem was solved using matlab package. The data sets were created using xmgr, a graphical tool in Unix environment. The steps adopted to solve this problem are as follows:

1) Data set1 and set2 were created in such a way that the mean of set1 was (2,-2) and mean of set 2 was (-2,2). The data obtained from xmgr was adjusted in matlab code to get the required mean. The four test vectors were defined and all these points were plotted in a graph which pictorially demonstrates three points belonging to both sets.

x1=(-1,-1)
x2=(0,0)
x3=(0.5,0.5)
x4=(0.5,-0.5)

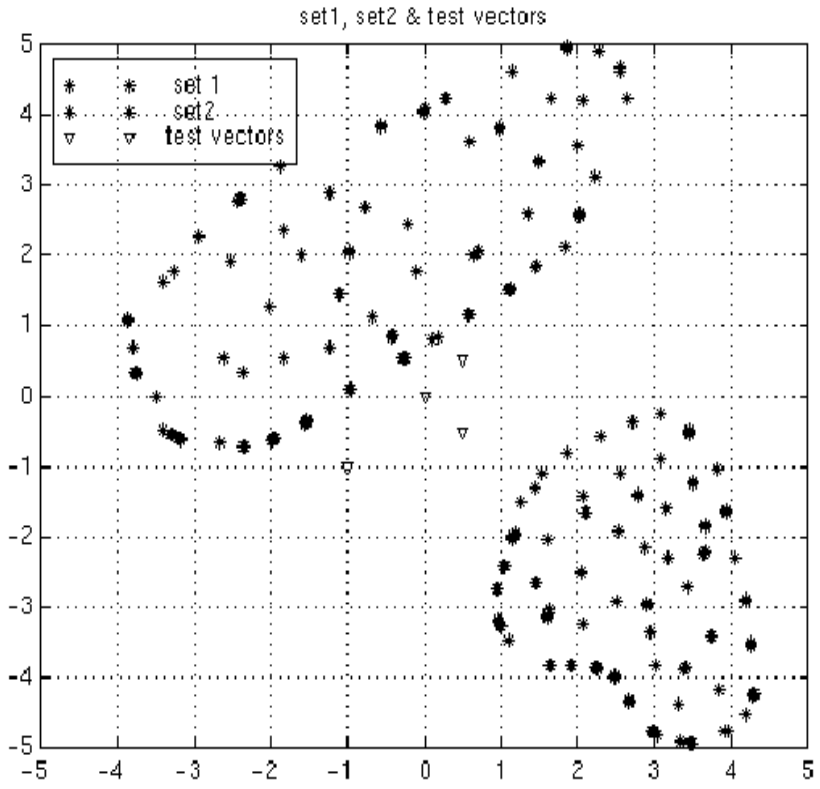


Figure 2:Generated data sets and given test vectors to be classified.

2) The Euclidean distance between the test set and the given data points were calculated using equation (1). The minimum distance determines which data set the data points belong. Euclidean distance can be calculated using equation (1).

$$|x, y| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (1)$$

The results obtained in normal space are as follows:

x	distance(x-u1)	distance(x-u2)	class
-1,-1	3.1623	3.1623	set 2
0, 0	2.8285	2.8284	set 2
0.5, 0.5	2.9155	2.9155	set 2
0.5, -0.5	3.5356	2.1213	set 2

Table 2: Classification of various data points using minimum distance criterion.

The test vectors are classified using the euclidean distance, but the data is not always orthonormal, i.e., $\dot{\Phi}\Phi^T = I$ where Φ is an $n \times n$ matrix, consisting of n eigenvectors as $\Phi = [\Phi_1 \ldots\ldots \Phi_n]$. The Euclidean distance calculated does not classify the data correctly. Hence, some type of transformation is to be applied on the data. A linear operation must be applied to transform such data to orthonormal space so that Euclidean distance can give meaningful results.

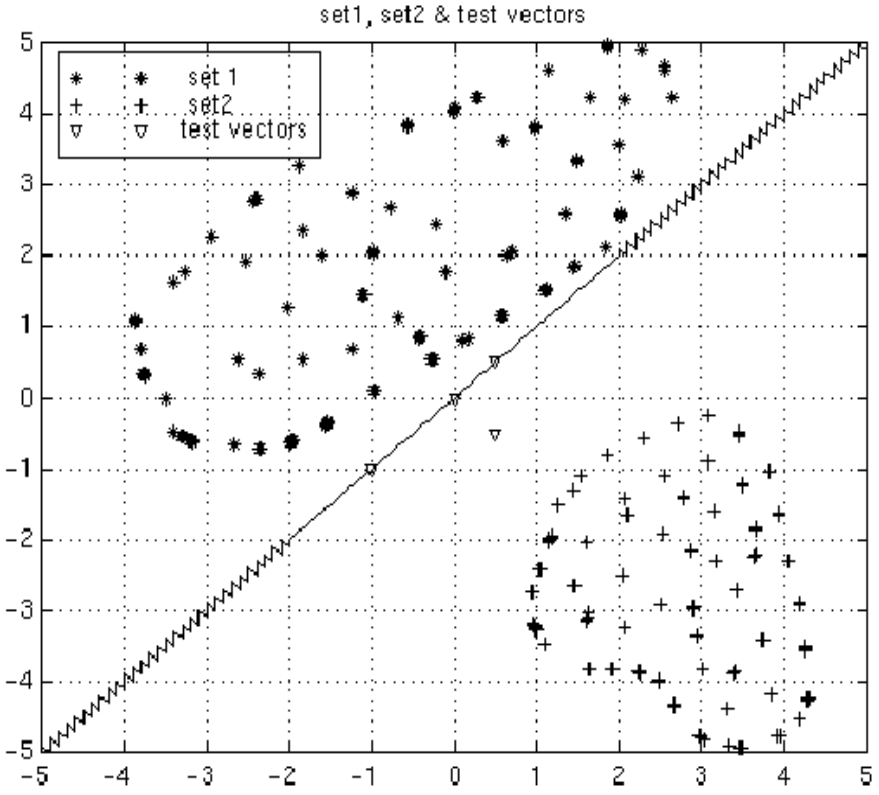


Figure 3:Generated data sets and decision line region in original space

3) Data is transformed using the whitening transformation technique, as given by Equation (2).

$$Y = TX \tag{2}$$

where Y is the transformed data, T is the transformation matrix, and X is the original data. The transformation matrix T can be calculated using Equation (3).

$$T = \Lambda^{-\frac{1}{2}}\Phi^{T} \tag{3}$$

where T is the transformation matrix, Λ is the eigen values and Φ^{T} is the transpose of the eigen vectors. The Λ, Φ^{T}, and T for the two data sets are shown in Table 3. Each of the data sets was transformed using the corresponding transformation matrix T. The transformed data sets in transformed space are shown in figure 4.

	Cov	Λ	Φ^{T}	T
set 1	$\begin{bmatrix} 3.5271 & 2.1607 \\ 2.1607 & 2.8932 \end{bmatrix}$	$\begin{bmatrix} 5.3940 & 0 \\ 0 & 1.0263 \end{bmatrix}$	$\begin{bmatrix} -0.7567 & -0.6538 \\ 0.6538 & -0.7567 \end{bmatrix}$	$\begin{bmatrix} -0.3258 & -0.2815 \\ 0.6454 & -0.7469 \end{bmatrix}$
set 2	$\begin{bmatrix} 1.1298 & -0.2051 \\ -0.2051 & 1.6074 \end{bmatrix}$	$\begin{bmatrix} 1.0538 & 0 \\ 0 & 1.6834 \end{bmatrix}$	$\begin{bmatrix} 0.9377 & 0.3474 \\ -0.3474 & 0.9377 \end{bmatrix}$	$\begin{bmatrix} 0.9135 & 0.3384 \\ -0.2677 & 0.7227 \end{bmatrix}$

Table 3: Eigen values, eigen vectors, and transformation matrices for the two data sets.

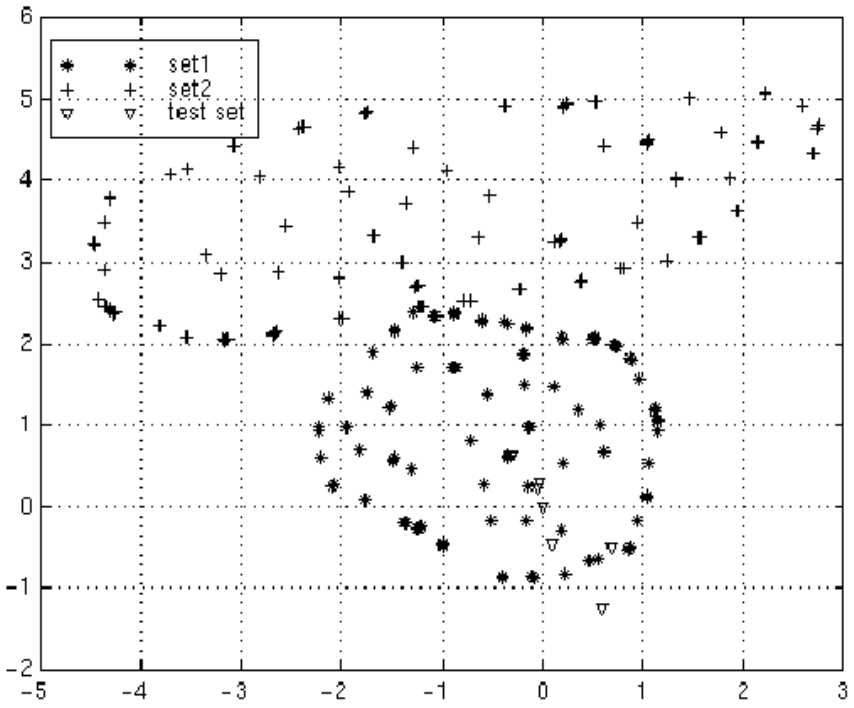


Figure 4: Data sets in the transformed space

4) Having transformed the data sets and the data points, the Euclidean distances were recomputed and the data points were classified based on the new euclidean distance. The results are summarized in Table 4. The covariance of the transformed data sets were also computed. The covariance matrices are identity matrices shown in Table 5.

x	distance(x-u1)	distance(x-u2)	class
-1,-1	2.0800	3.5491	set 1
0, 0	1.8141	3.1185	set 1
0.5, 0.5	1.7486	3.1016	set 1
0.5, -0.5	2.5039	2.5465	set 1

Table 4: Classification of various data points that have been transformed.

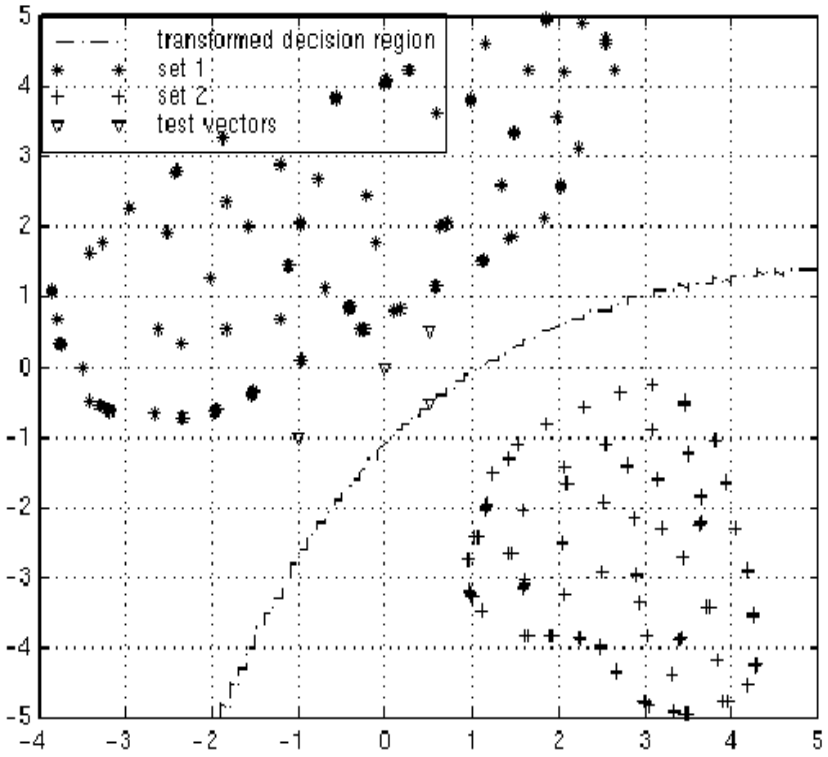


Figure 5: Data sets and decision region in transformed space.

5) The covariance matrices of both set 1 and 2 were computed which were identity matrices. The purpose of whitening transform was to transform the given data into a new space where the variance along all direction are equal which is important for classifying.

	Cov
transformed set 1	$\begin{bmatrix} 1.0000 & 0.0000 \\ 0.0000 & 1.0000 \end{bmatrix}$
transformed set 2	$\begin{bmatrix} 1.0000 & 0.0000 \\ 0.0000 & 1.0000 \end{bmatrix}$

Table 5: Covariance matrices of the transformed data sets

6)Next, the points in each data set were transformed by T as shown by equation (4). The purpose of this transformation is to readjust the original data sets so that their means are centered

around the origin.

$$\bar{y} = T_n(x_o - \mu_n) \text{ for all set 'n' for n=1,2} \quad (4)$$

The results of the transformation of the readjusted data are given in Figure 6.

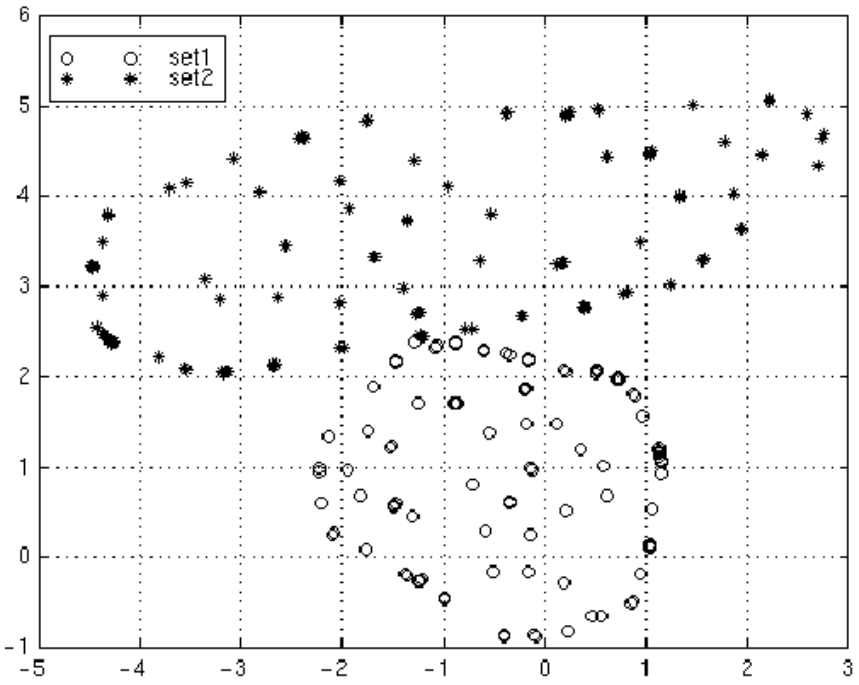


Figure 6: Data sets and decision region in transformed space.

4. References

[1] K. Fukunaga, "Introduction to Statistical Pattern Recognition," Academic Press, San Diego, California, 1990.