PRINCIPAL COMPONENT ANALYSIS

EE 8993: Speech Processing

Homework #2

Audrey Le

Professor: Dr. Joseph Picone

1. INTRODUCTION

One method to determine which class a test point belongs to is to find the Euclidean distance between a test point and each of the classes and assign a test point to the class that has the minimum distance. A Euclidean distance is used because it is simple. However, a Euclidean distance does not give meaningful results with data that is not orthonormal data. Instead a linear transformation is used to transform the data so that a Euclidean distance can be used. In this project, a Euclidean distance in the original space and that in a transformed space using principal component analysis are compared.

2. PROBLEM DESCRIPTION

Given four test points (see Table 1), determine which data set each of the test points belongs. The two data sets should have shapes similar to the ones shown in Figure 1. Each data set contains 100 points and should have the mean centered around (-2, 2) and (2, -2). Use principal component analysis and Euclidean distance to classify the data points in transformed spaces and compare the classification results with the results obtained from original space.

Data point	Coordinates
a	-1,-1
b	0, 0
c	0.5, 0.5
d	0.5, -0.5

Table 1: Data points to be classified.

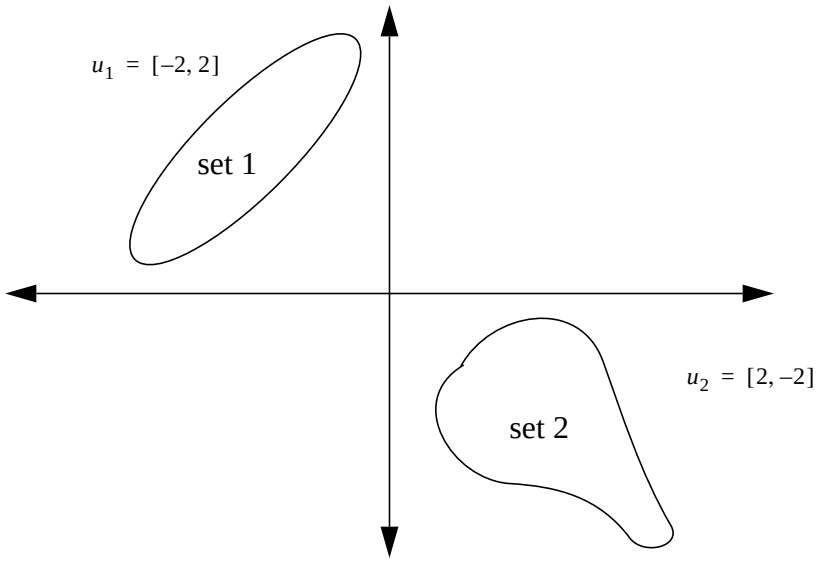


Figure 1: General shape of the data sets.

3. IMPLEMENTATION AND RESULTS

First, the data sets were generated using xmgr, a Unix graphing tool. Data points were generated by using the point-click mouse operation to get the general shapes as indicated in Figure 1. A perl script was used to adjust the means of the two data sets to have the required means. The resulting data sets are given in Figure 2. The test points to be classified are also shown in Figure 2.

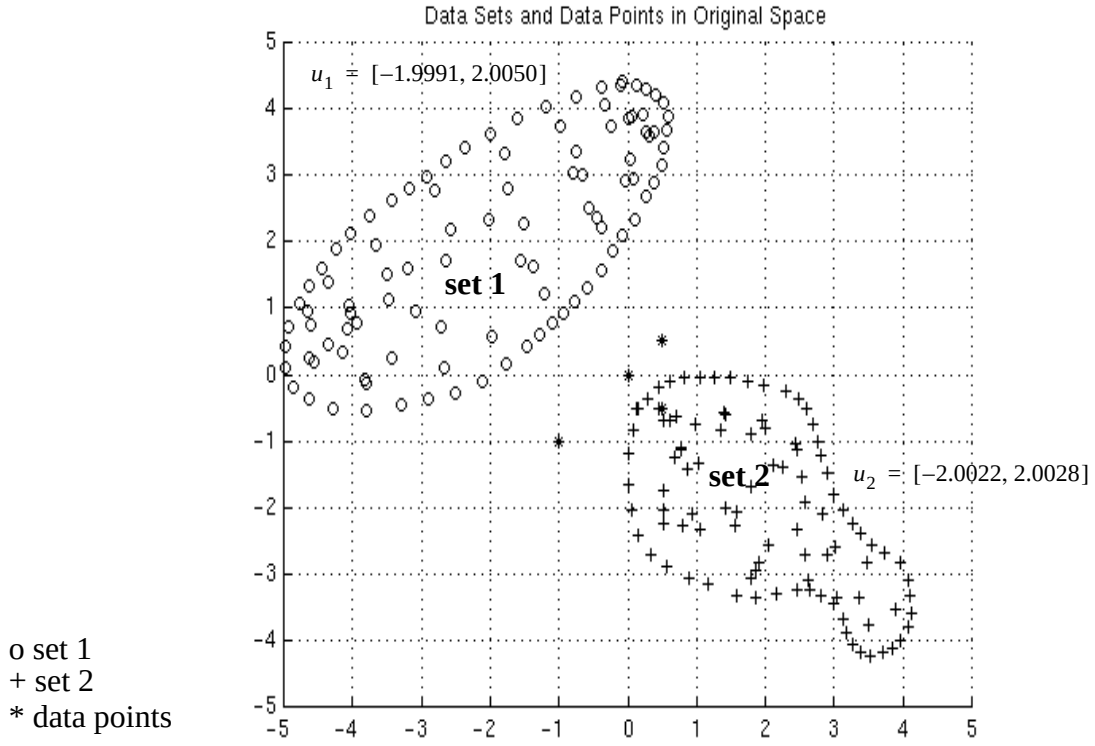


Figure 2: Generated data sets and given data points to be classified.

Next, the Euclidean distance between each of the data points and each of the means of the data sets was calculated using Equation (1). The minimum distance determines which data set the data points belong. The results are summarized in Table 2. The decision regions are given in Figure 3.

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

x	Distance(x-u1)	Distance(x-u2)	Class
-1,-1	3.1667	3.1667	set 2
0, 0	2.8313	2.8320	set 1
0.5, 0.5	2.9173	2.9190	set 1
0.5, -0.5	3.5384	2.1249	set 2

Table 2: Classification of the test points using minimum distance criterion.

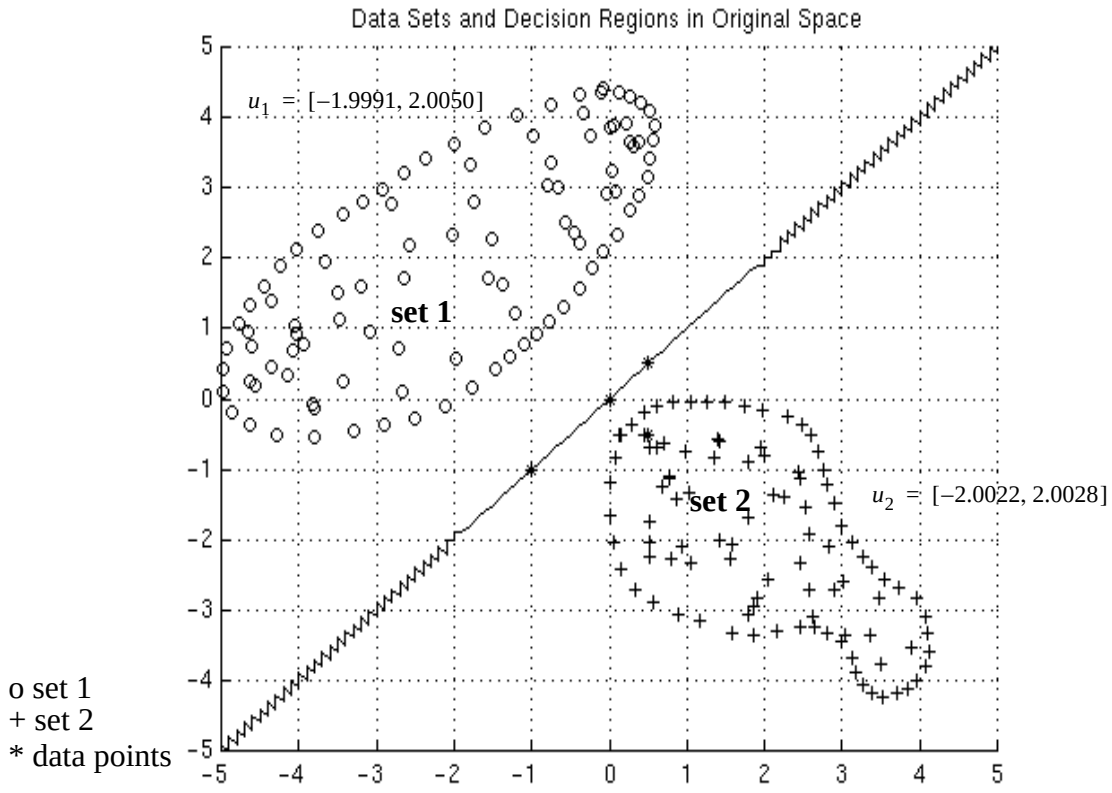


Figure 3: Decision regions in original space.

The Euclidean distance gives the physical notion of distance. However, it works only on orthonormal data [1]. Sometimes the data is not orthonormal. By using the Euclidean distance on such data, we will not obtain meaningful results. A linear operation must be applied to transform such data to orthonormal space so that Euclidean distance can give meaningful results. The transformation is 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. 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 matrix of eigenvalues, and Φ^T is the transpose of the eigenvectors.

Returning to our classification problem we can see that the Euclidean distance in the original space can not differentiate which data set the test points belong. We do not know if Euclidean distance is a good measure or if the data sets are not orthnormal. To determine this we transformed the data sets. The Λ, Φ^T, and T for the two data sets are given in Table 3. Each of the

data sets was transformed using the corresponding transformation matrix T. The transformed data sets are given in Figure 4 and Figure 5 for data set 1 and data set 2 respectively.

Data set	Cov	Λ	Φ^T	T
set 1	$\begin{bmatrix} 3.2637 & 1.8923 \\ 1.8923 & 2.1805 \end{bmatrix}$	$\begin{bmatrix} 0.4617 & 0 \\ 0 & 1.1517 \end{bmatrix}$	$\begin{bmatrix} -0.7985 & -0.6020 \\ 0.6020 & -0.7985 \end{bmatrix}$	$\begin{bmatrix} -0.3687 & -0.2780 \\ 0.6934 & -0.9197 \end{bmatrix}$
set 2	$\begin{bmatrix} 1.4500 & -0.8691 \\ -0.8691 & 1.5204 \end{bmatrix}$	$\begin{bmatrix} 1.2748 & 0 \\ 0 & 0.6516 \end{bmatrix}$	$\begin{bmatrix} 0.7213 & 0.6926 \\ -0.6926 & 0.7213 \end{bmatrix}$	$\begin{bmatrix} 0.9195 & 0.8830 \\ -0.4514 & 0.4700 \end{bmatrix}$

Table 3: Eigenvalues, eigenvectors, and transformation matrices for the two data sets.

o set 1

* data points using t1

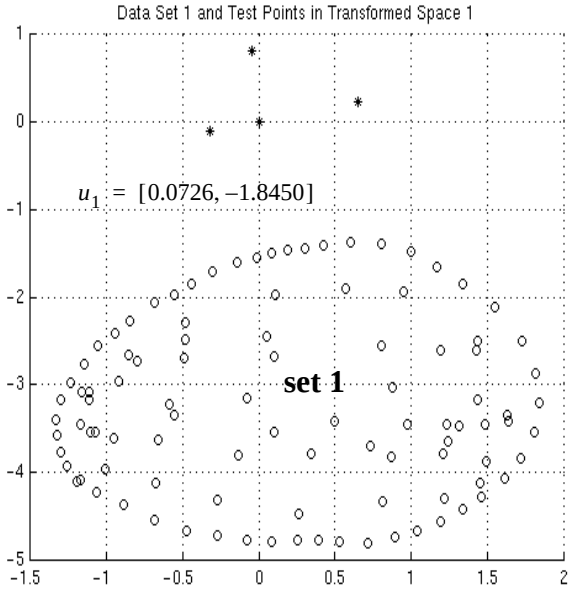


Figure 4: Data set 1 and test points in transformed space 1.

y	Distance(y-u1)	Distance(y-u2)	Class
-1,-1	3.4877	2.6175	set 2
0, 0	3.2350	1.8465	set 2
0.5, 0.5	3.1572	2.0311	set 2
0.5, -0.5	4.0428	1.3854	set 2

Table 4: Classification of the test points that have been transformed.

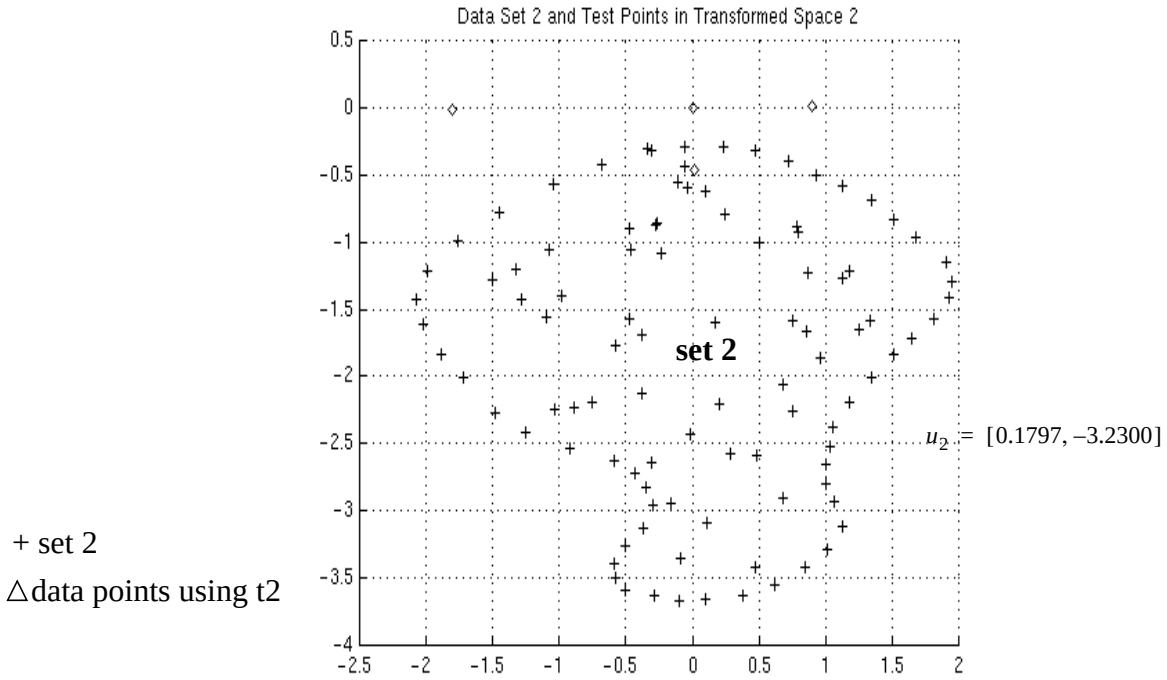


Figure 5: Data set 2 and test points in transformed space 2.

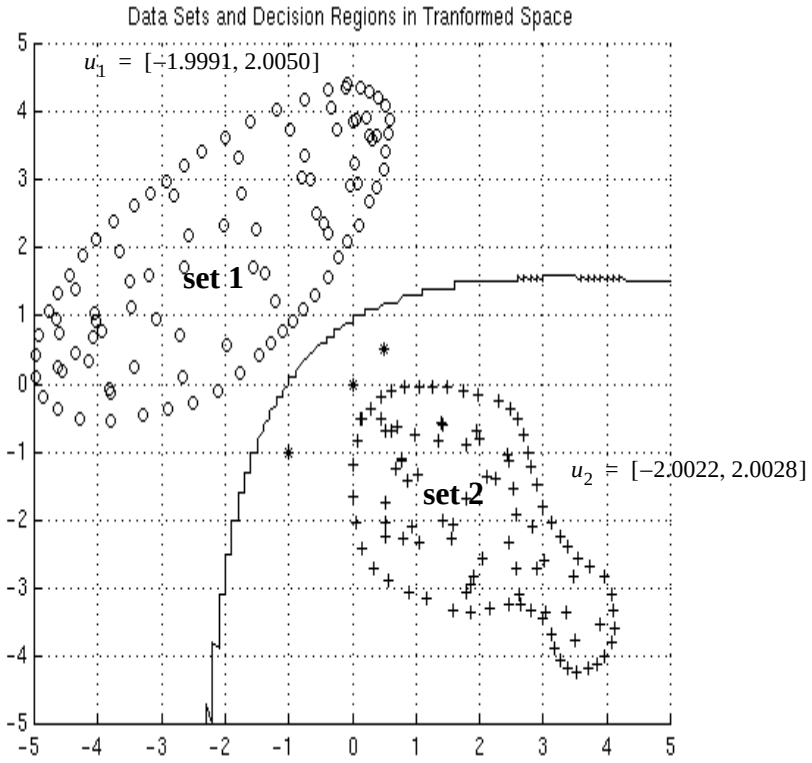
Transformed data set	Cov
set 1	$\begin{bmatrix} 1.0000 & 0.0000 \\ 0.0000 & 1.0000 \end{bmatrix}$
set 2	$\begin{bmatrix} 1.0000 & 0.0000 \\ 0.0000 & 1.0000 \end{bmatrix}$

Table 5: Covariance matrices of the transformed data sets.

Having transformed the data sets and the test points, the Euclidean distances were recomputed and the test points were reassigned according to the new distance. The results are summarized in Table 4. The decision regions are given in Figure 6. The covariance matrices of the transformed data sets were found to be identity matrices as expected and are shown in Table 5.

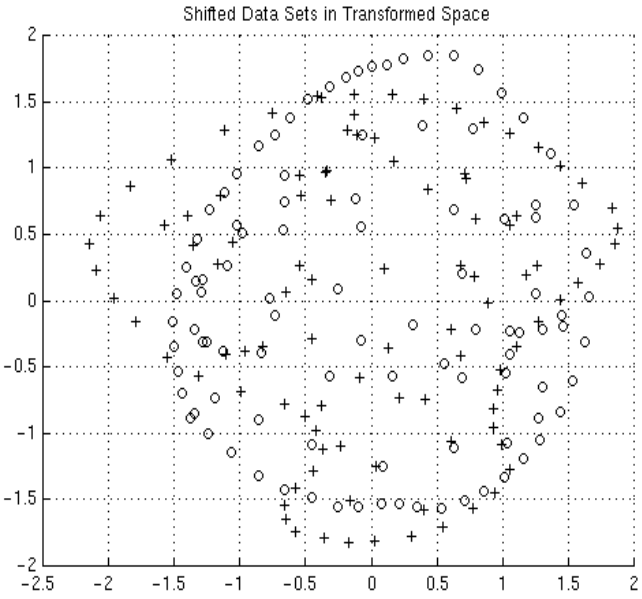
Next, the original data sets were readjusted so that their means are centered around the origin. The readjusted data sets were transformed using the transformation matrices obtained above. This step is expressed in Equation (4). The results of the transformation of the readjusted data are given in Figure 7. Transformed data points are also shown in Figure 7.

$$Y = T(X - u) \tag{4}$$



o set 1
+ set 2
* data points

Figure 6: Decision regions in transformed space.



o set 1
+ set 2

Figure 7: Shifted and transformed data sets.

The main point we can conclude from the results that we have obtained is that principal component analysis takes the variance of the data into account in classification while straight Euclidean

distance does not.

4. REFERENCES

[1] J. R. Deller Jr., J. G. Proakis, and J. H. L. Hansen. *Discrete-Time Processing of Speech Sig nals*. Macmillan Publishing Company, 1993.