EE 8993: Speech Recognition

Homework Assignment #2
Principal Component Analysis

January 23, 1998

submitted to:

Dr. Joseph Picone

Department of Electrical and Computer Engineering
413 Simrall, Hardy Rd.
Mississippi State University
Box 9571
MS State, MS 39762

submitted by:

Julie Ngan

Department of Electrical and Computer Engineering
Mississippi State University
Box 9571
Mississippi State, Mississippi 39762
Tel: 601-325-8335
Fax: 601-325-3149
email: ngan@isip.msstate.edu

I. Problem Definition

Define two sets such that the distributions approximate an ellipse and a pear shape. The elliptical distribution should stretch from lower-left to upper-right at about a $45°$ angle and should be longer in that direction than it is wide. Set 2 should look like a pear with the stem pointing at $-45°$. Set 1 should have a mean of approximately $m1 = (-2,2)$ and Set 2 should have a mean of approximately $m2 = (2,-2)$. Each set contains 100 points.

Define a test set of four points:

$$x_1 = (-1,-1)$$
$$x_2 = (0,0)$$
$$x_3 = (1/2,1/2)$$
$$x_4 = (1/2,-1/2)$$

Each sample point is then classified as a member of either set using minimum Euclidean distance. Then the two data sets will be analyzed to find the decision regions using Principal Component Analysis, and the sample points will be classified again using the results.

II. Data Set Generation

The two test sets are generated using the point operation function in *xmgr*. Points are randomly drawn to obtain the required shapes and then vertically or horizontally shifted to obtain the required means. A plot of the two data sets is shown in Figure 1.

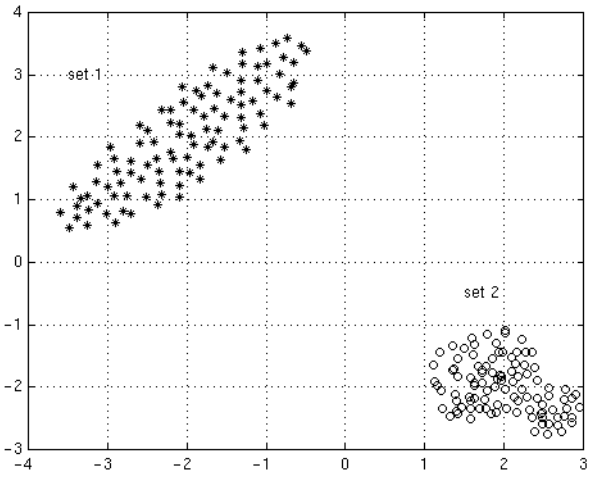

Figure 1 The two generated data sets.

III. Euclidean Distance and Direct Data Classification

The Euclidean distance between each sample vector and each of the two data set means is calculated by Equation 1:

$$d(\bar{x}, \overline{m_n}) = \sqrt{\sum_{k=1}^{N} |x_k - m_{nk}|^2}, \text{ where } n = 1, 2 \quad (1)$$

Each sample points is classified as a member of set 1 or set 2 according to minimum Euclidean distance. The results are shown in Table 1:

Sample point	Euclidean Distance from Data set 1	Euclidean Distance from Data set 2	Classification
x1	3.1623	3.1623	set 2
x2	2.8284	2.8284	set 2
x3	2.9155	2.9155	set 2
x4	3.5355	2.1213	set 2

Table 1 Results of classifying the sample points.

IV. Direct Decision Region

Because the means are located on the line where $x = -y$, the line $x = y$ will be equidistant from the means at all time. As a result, the decision region is no more than the $x = y$ line. Points on the left-hand side of the line belong to set 1, whereas points on the right-hand side of the line belong to set 2. Figure 2 shows a plot of the two data sets with the decision line.

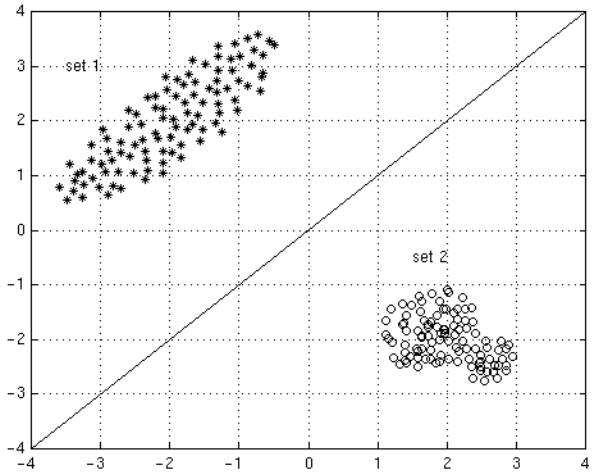

Figure 2 Plots of the two data sets with decision region.

V. Prewhitening Transformation

The Euclidean distance fits with our physical notion of distance. However, when different dimensions of the vector are not orthonormal, i.e. they are not equally important in our data sets. Using Euclidean distance alone will yield undesirable results unless a linear operation is applied which transforms the vector representations to ones based on orthonormal vectors [1].

This is done by decomposing the covariance matrix $\underline{C}$ into its eigenvectors and eigenvalues:

$$\underline{C} = \underline{\Phi}\underline{\Lambda}\underline{\Phi}^{T} \tag{2}$$

where $\underline{\Phi}$ denotes a matrix of eigenvectors of $\underline{C}$ and $\underline{\Lambda}$ denotes a diagonal matrix whose elements are the eigenvalues of $\underline{C}$. Then the transformation matrix can be written as:

$$\underline{T} = \underline{\Lambda}^{-1/2}\underline{\Phi}\bar{x} \tag{3}$$

The values calculated for the two data sets are shown in Table 2:

	Data set 1	Data set 2
Covariance Matrix	$\begin{bmatrix} 0.6617 & 0.5572 \\ 0.5572 & 0.6633 \end{bmatrix}$	$\begin{bmatrix} 0.2343 & -0.0690 \\ -0.0690 & 0.1713 \end{bmatrix}$
$\underline{\Lambda}$	$\begin{bmatrix} 3.0818 & 0 \\ 0 & 0.9055 \end{bmatrix}$	$\begin{bmatrix} 1.8944 & 0 \\ 0 & 2.8065 \end{bmatrix}$
$\underline{\Phi}$	$\begin{bmatrix} 0.7076 & 0.7066 \\ -0.7066 & 0.7076 \end{bmatrix}$	$\begin{bmatrix} -0.8413 & -0.5405 \\ 0.5405 & -0.8413 \end{bmatrix}$
$\underline{T}$	$\begin{bmatrix} 2.1807 & -2.1776 \\ 0.6398 & 0.6407 \end{bmatrix}$	$\begin{bmatrix} -1.5939 & 1.0240 \\ -1.5169 & -2.3612 \end{bmatrix}$

Table 2 Values calculated for the two data sets.

VI. Transformed Data Sets

The two data sets are transformed into the new spaces using the respective transform matrices. The results are plotted in Figure 3. The means of the two new data sets are then determined as $(-8.7165, 0.0019)$ and $(-5.2356, 1, 6885)$ respectively. The four sample points are transformed into the two transformed spaces as shown in Table 3. The new Euclidean distances between the new means and the new sample points are calculated and the sample points are re-classified as shown in Table 4.

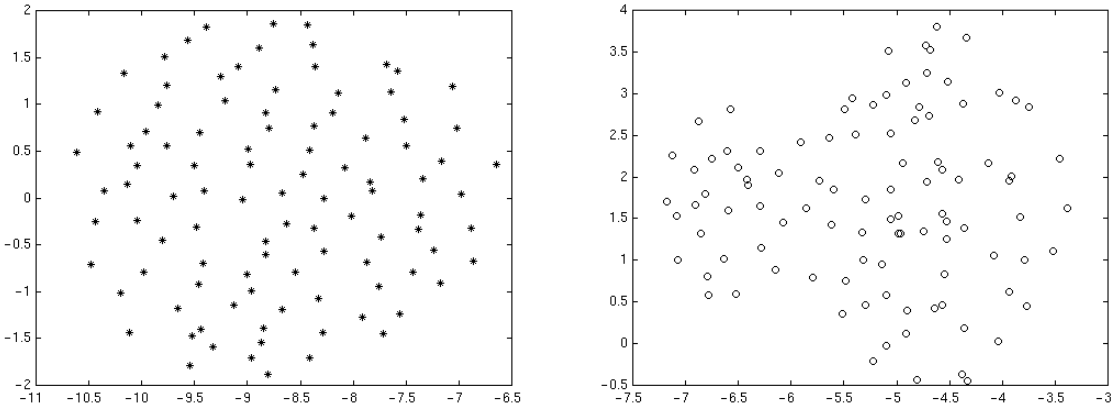

Figure 3 The data sets in transformed spaces, set 1 on left, set 2 on right.

Sample point	New values in transformed space 1	New values in transformed space 2
x1	(-0.0032, -1.2805)	(0.5699, 3.8781)
x2	(0, 0)	(0, 0)
x3	(0.0016, 0.6403)	(-0.2849, -1.9391)
x4	(2.1791, -0.0005)	(-1.3089, 0.4221)

Table 3 The transformed sample points.

Sample point	Euclidean distance from data set 1	Euclidean distance from data set 2	Classification
x1	8.8072	6.2047	set 2
x2	8.7165	5.5012	set 2
x3	8.7415	6.1375	set 2
x4	10.8957	4.1259	set 2

Table 4 The Euclidean distance from the transformed sample points to the transformed means.

VII. New Decision Region

The new decision region can be found by locating points where the Euclidean distance between the transformed points and the transformed means are equidistant. These points are plotted along with the untransformed data sets and the decision region is shown to be a parabola, as shown in Figure 4.

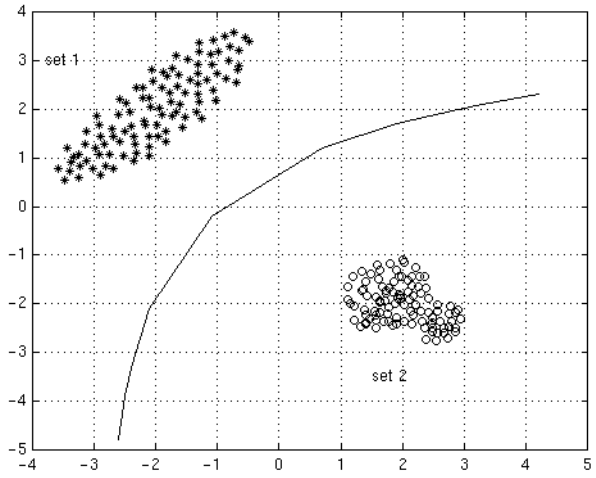

Figure 4 The new decision region.

VIII. Demonstration of Orthonormality

The orthonormality of data sets is shown by computing the covariance matrices of the transformed data to show identity matrices of $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and the plotting of Equation 4 on each set respectively produce a circle, as shown in Figure 5 and Figure 6.

$$\bar{y} = \underline{T}(\bar{x} - \bar{\mu}) \tag{4}$$

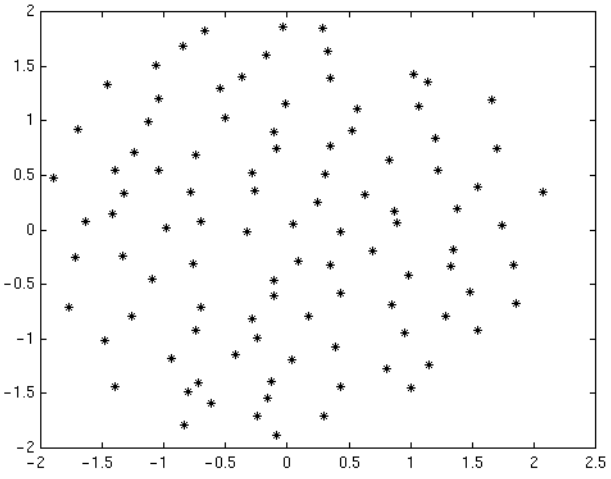

Figure 5 Orthonormalized data set 1.

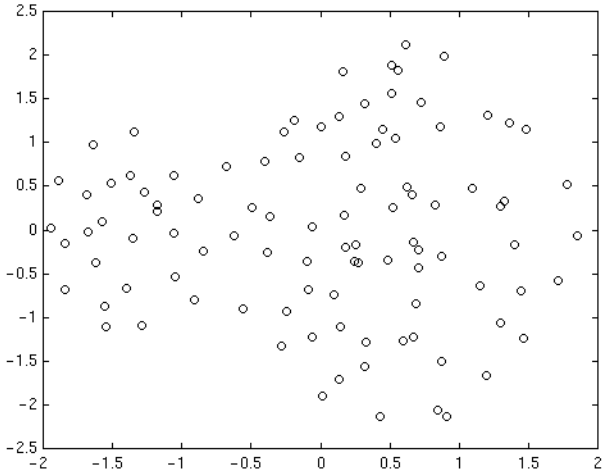

Figure 6 Orthonormalized data set 2.

IX. REFERENCES

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