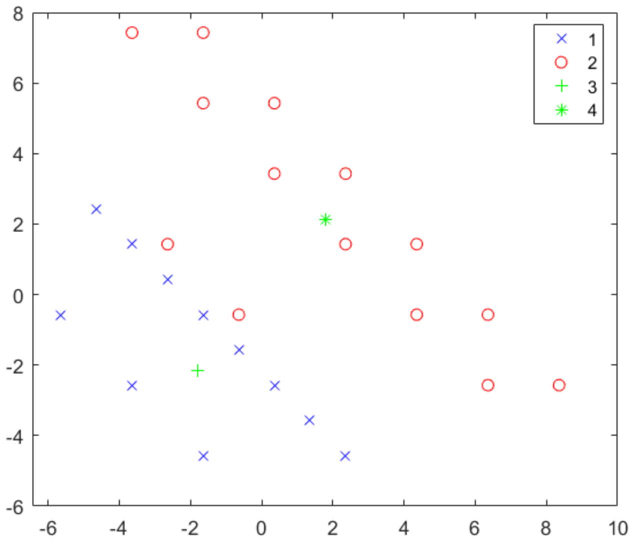
Solution to selected exercises on feature selection/PCA

Exercise 1: Feature selection, dimensionality reduction and classification

A two-class dataset is sketched below.



The mean values for the two classes are indicated in green.
In this exercise, the following values are given:

$$\text{The mean for the entire data set is zero.} \quad \mu 1 = \begin{bmatrix} -1.8 \\ -2.1 \end{bmatrix} \quad \mu 2 = \begin{bmatrix} 1.8 \\ 2.1 \end{bmatrix}$$

$$\text{Covariance} \quad R = \begin{bmatrix} 12.4 & -2.4 \\ -2.4 & 13.1 \end{bmatrix}$$

$$\text{Eigenvalues of R}: \lambda_1 = 10.4 \quad \lambda_2 = 15.2$$

$$\text{Eigenvectors of R} \quad v_1 : \begin{bmatrix} -0.76 \\ -0.65 \end{bmatrix} \quad v_2 : \begin{bmatrix} -0.65 \\ 0.76 \end{bmatrix}$$

a) In Figure 1 the appendix, indicate the decision boundary for this problem if we use the two original features and a Gaussian classifier with equal diagonal covariance matrix.

Answer: See appendix

b) Based on this decision boundary, compute the confusion matrix for the training data.

Answer: Confusion matrix:

$$\begin{array}{cc} & \textit{True class} \\ \textit{Estimated class} & \begin{bmatrix} 11 & 0 \\ 2 & 12 \end{bmatrix} \end{array}$$

c) Principal component analysis projects the data based on a criterion function. What is this criterion function?

Answer: Maximum variance (equal to minimum signal representation)

d) How do we find the first principal component of the data?

Answer: Find the eigenvector of the correlation matrix R corresponding to the largest eigenvalue

e) Indicate in Figure 2 in the appendix the direction of the first principal component of the data.

Answer: See appendix

f) Consider sequential feature selection on this data set, with average Euclidean distance between the classes as the criterion function.
Set up the value of the criterion function for the two original features independently.

Answer: J(first feature)=mu21-mu11 = 1.8-(-1.8) = 3.6
J(second feature) = mu22-mu12 = 2.1-(-2.1) = 4.2

g) Select the best feature with this criterion function, and draw the decision boundary for this feature set on Figure 1.

Answer: The criterion is maximum average class distance, so for one feature the best feature is feature 2, with a distance of 4.2. The boundary is midway between the class centers for feature 2, see plot. (boundary y=0) 8 samples are misclassified using feature 2. See fig. 1

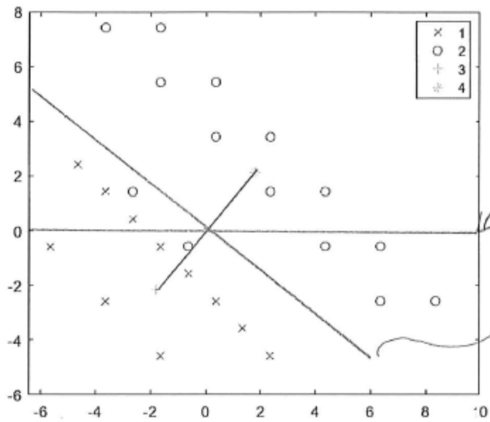
h) Perform a projection down onto the principal component by sketching the projection (no computation needed) on Figure 2.

Answer: See plot

i) Do a classification based on Euclidean distance of the first principal component and count the number of misclassified samples.

Answer: 12 samples are misclassified based on the first principal component.

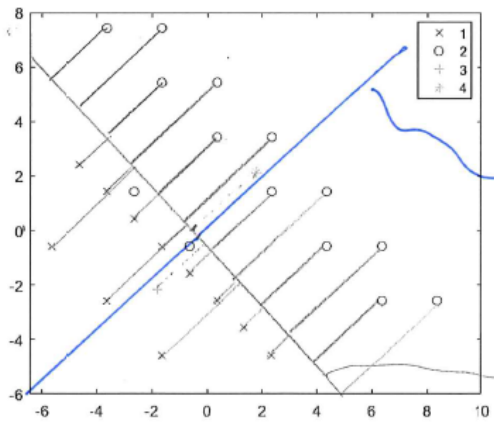
Figure 1



g) Decision boundary for feature 2

Boundary $\Sigma = I$ original features

Figure 2



PCA dir: $v_2 = \begin{bmatrix} -0.65 \\ 0.76 \end{bmatrix}$

Decision boundary based on PCA 1

PCA 1

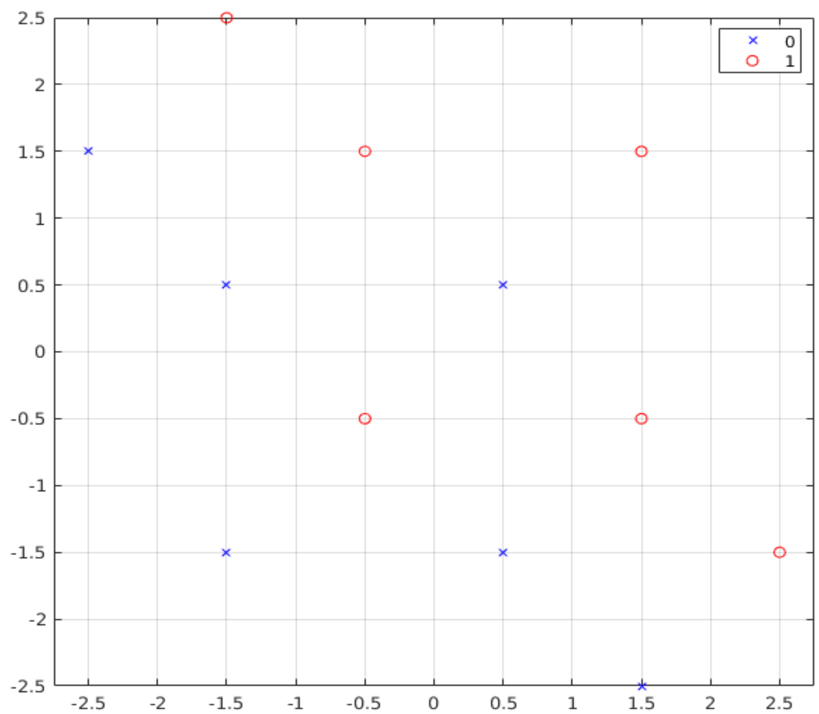
Exercise 4 (from 2018 exam) : PCA and classification

A set of samples from two classes are given as (a scatter plot is also given below):

Class1: $\begin{bmatrix} -2.5 & 1.5 \\ -1.5 & 0.5 \\ 0.5 & -1.5 \\ 1.5 & -2.5 \\ 0.5 & 0.5 \\ -1.5 & -1.5 \end{bmatrix}$ $\mu_1 = \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}$ $\Sigma_1 = \begin{bmatrix} 2.4 & -1.6 \\ -1.6 & 2.4 \end{bmatrix}$ $\Sigma^{-1}{}_1 = \begin{bmatrix} 0.75 & 0.5 \\ -0.5 & 0.75 \end{bmatrix}$

- Class2: $\begin{bmatrix} -1.5 & 2.5 \\ -0.5 & 1.5 \\ 1.5 & -0.5 \\ 2.5 & -1.5 \\ -0.5 & -0.5 \\ 1.5 & 1.5 \end{bmatrix}$ $\mu_2 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ $\Sigma_2 = \begin{bmatrix} 2.4 & -1.6 \\ -1.6 & 2.4 \end{bmatrix}$ $\Sigma^{-1}{}_2 = \begin{bmatrix} 0.75 & 0.5 \\ -0.5 & 0.75 \end{bmatrix}$

- Global mean: $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$
- Global covariance matrix: $\begin{bmatrix} 2.45 & -1.18 \\ -1.18 & 2.45 \end{bmatrix}$
- Eigenvalues 1.27 and 3.65 with eigenvectors: $\begin{bmatrix} -0.7 \\ -0.7 \end{bmatrix}$ and $\begin{bmatrix} -0.7 \\ 0.7 \end{bmatrix}$



a) Find the direction of the first principal component of this data set.
This is the eigenvector associated with the largest eigenvalue [-0.7, 0.7]

b) Sketch the eigenvector on the plot

c) Compute the new feature values using this component for all samples.

*For class 1: Feature vector transpose*eigenvector:*

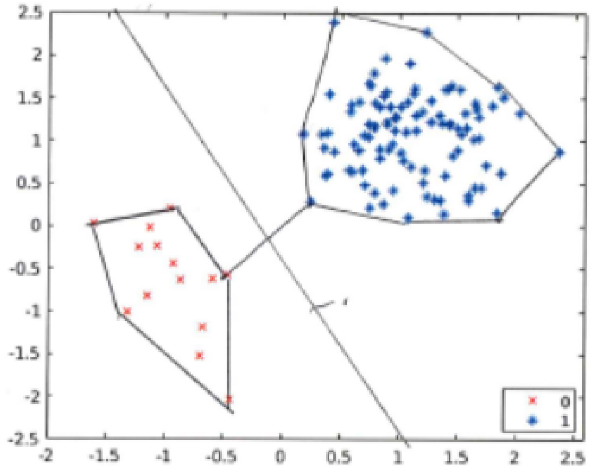
$$\begin{bmatrix} 2.5 * 0.7 + 1.5 * 0.7 = 2.8 \\ 1.5 * 0.7 + 0.5 * 0.7 = 1.4 \\ -0.5 * 0.7 - 1.5 * 0.7 = -1.4 \\ -1.5 * 0.7 - 2.5 * 0.7 = -2.8 \\ -1.5 * 0.6 - 1.5 * 0.7 = 0 \\ -0.5 * 0.7 + 0.5 * 0.7 = 0 \end{bmatrix}$$

For class 2:

$$\begin{bmatrix} 1.5 * 0.7 + 2.5 * 0.7 = 2.8 \\ 0.5 * 0.7 + 1.5 * 0.7 = 1.4 \\ -1.5 * 0.7 - 0.5 * 0.7 = -1.4 \\ -2.5 * 0.7 - 2.5 * 0.7 = -2.8 \\ -0.5 * 0.7 + 0.5 * 0.7 = 0 \\ -1.5 * 0.7 + 1.5 * 0.7 = 0 \end{bmatrix}$$

d) Plot the points in the figure and discuss how the first principal component performs in this case.
The two classes collapse to equal points.

Figur 1



Figur 2