ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΑΙ ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ
1ο ΣΕΤ ΑΣΚΗΣΕΩΝ

ΜΙΧΑΗΛΙΔΗΣ ΣΤΕΡΓΙΟΣ 2020030080 – ΜΟΥΣΤΑΚΑΣ ΙΩΑΝΝΗΣ 2020030120
ΟΜΑΔΑ 60

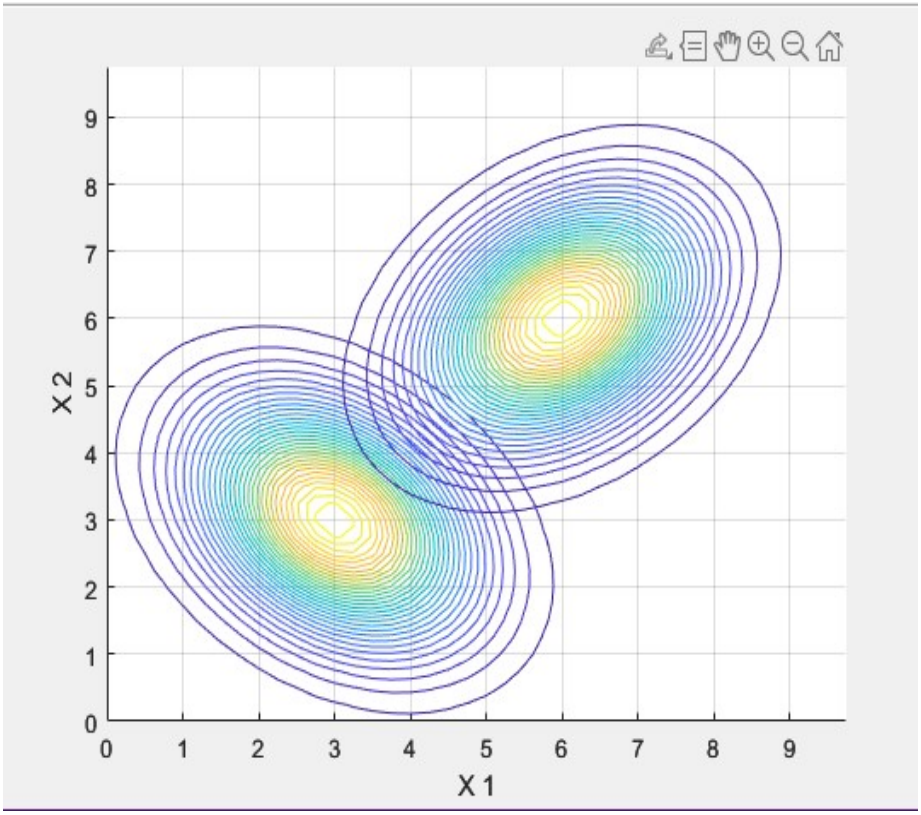
Άσκηση 1

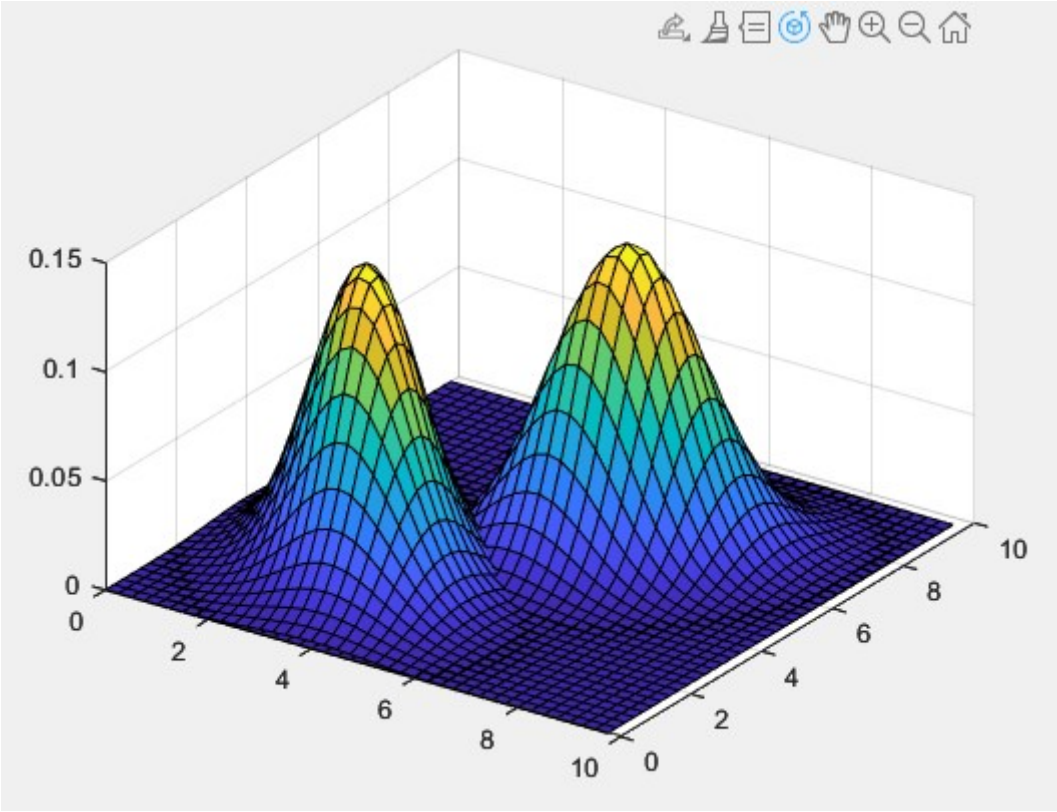
Σκοπός είναι η κατηγοριοποίηση 2D δειγμάτων σε δύο κλάσεις ω_1 ή ω_2.
Μετά από επεξεργασία και ελαχιστοποίηση σφάλματος , καταλήγουμε στο όριο απόφασης :

$$x_1 x_2 + 1.7 \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) - 18 = 0$$ **(μορφή υπερβολής)**

(Για τις πράξεις που οδηγούν στα σύνορα αποφάσεων, ανατρέξτε στα συνημμένα αρχεία…)
Για τυχαίο διάνυσμα X τ.ω. x_1x_2 μεγαλύτερο από το όριο απόφασης, αποφασίζουμε υπέρ της κλάσης ω_1 ενώ για x_1x_2 μικρότερο από το όριο αποφασίζουμε υπέρ της ω_2.

b) Οι ισοϋψείς καμπύλες που ζητούνται (σε 2D) και οι αντίστοιχες κατανομές στο R^3 :

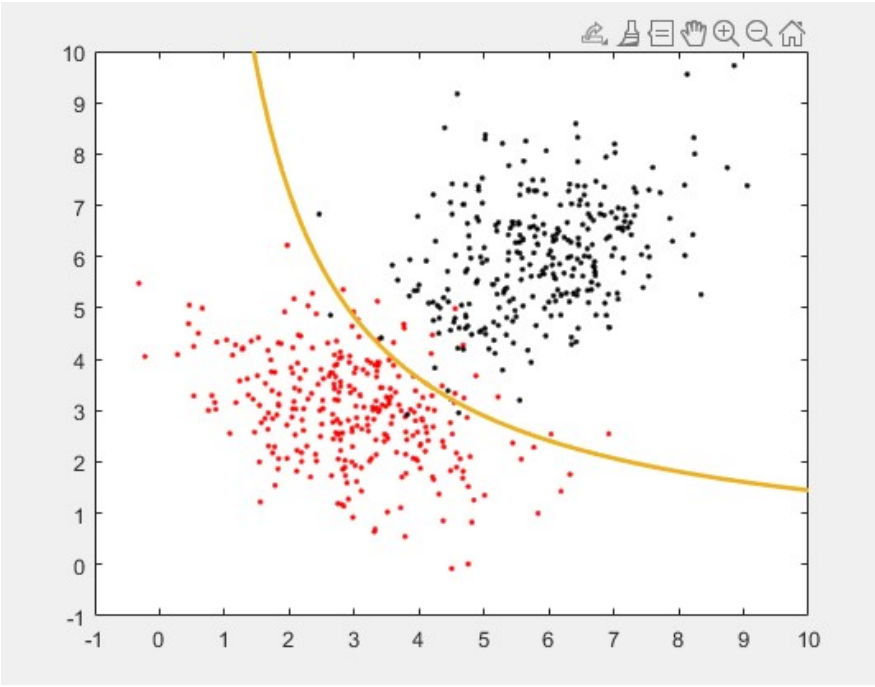




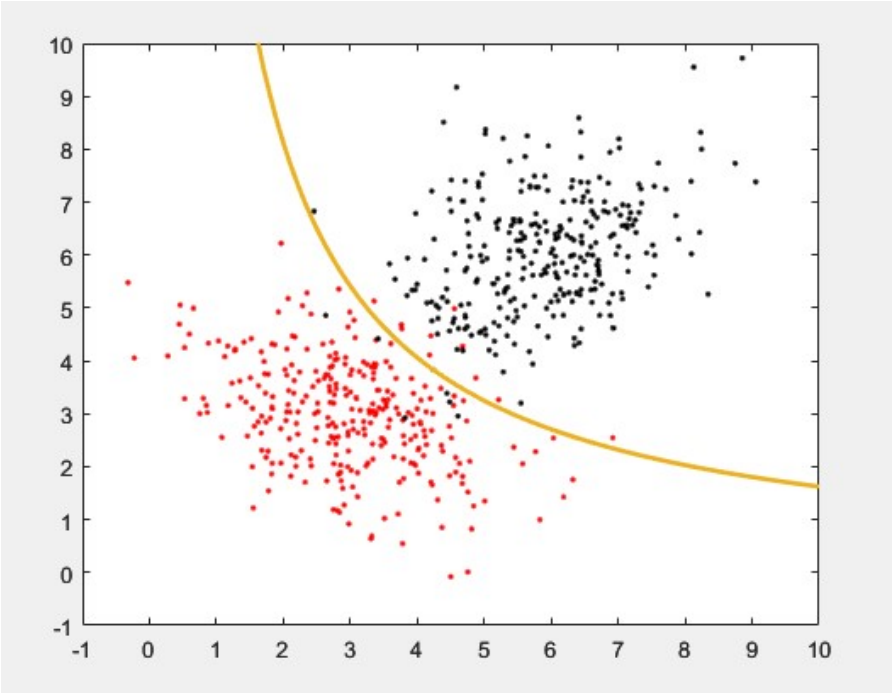
c) Στη συνέχεια δημιουργούμε 300 δείγματα από κάθε κατανομή , τα σχεδιάζουμε σε κοινό plot που συμπεριλαμβάνει την καμπύλη απόφασης και παρατηρούμε ότι η κατηγοριοποίηση είναι ακριβής.

<u>Έστω $P(\omega_1) = \pi_1$:</u>

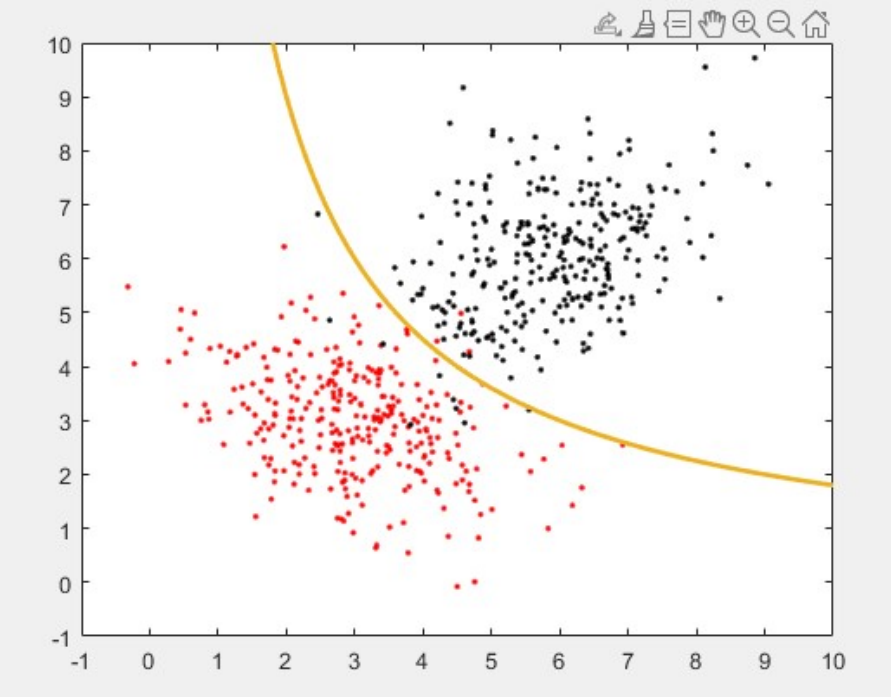
Για $\pi_1 = 0.1$:



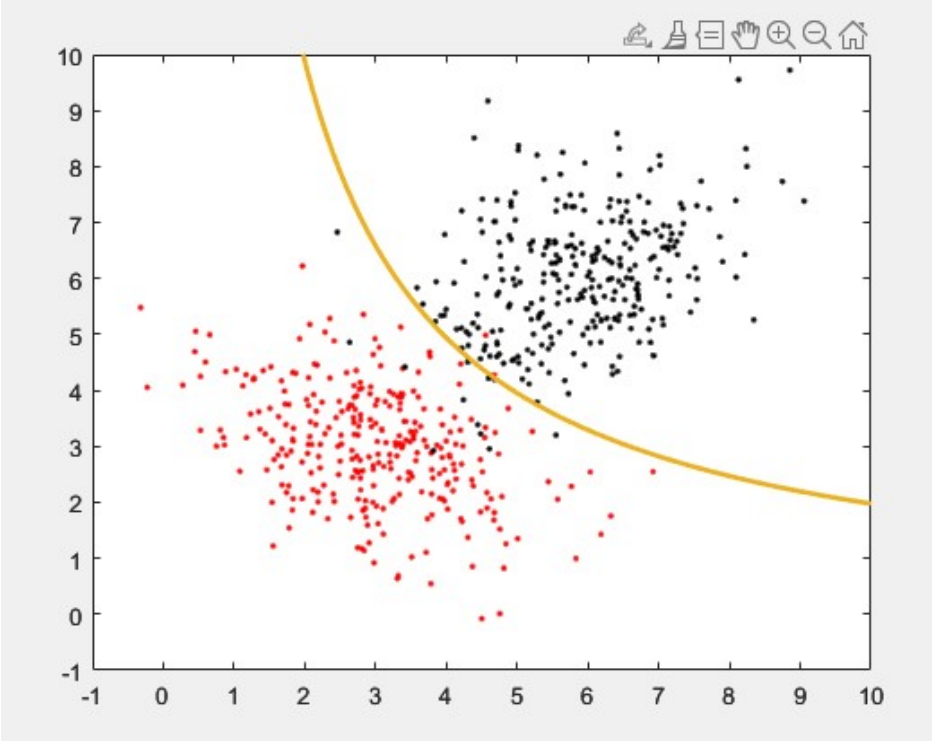
Για $\pi_1 = 0.25$:



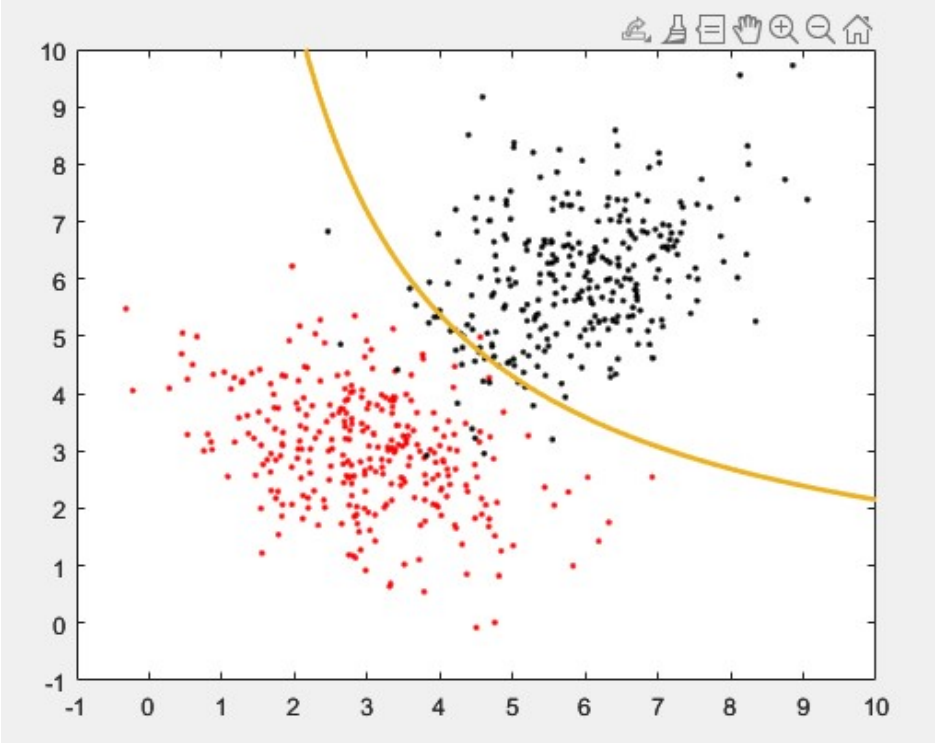
Για $\pi_1 = 0.5$:



Για $\pi_1 = 0.75$:



Για $\pi_1 = 0.9$:



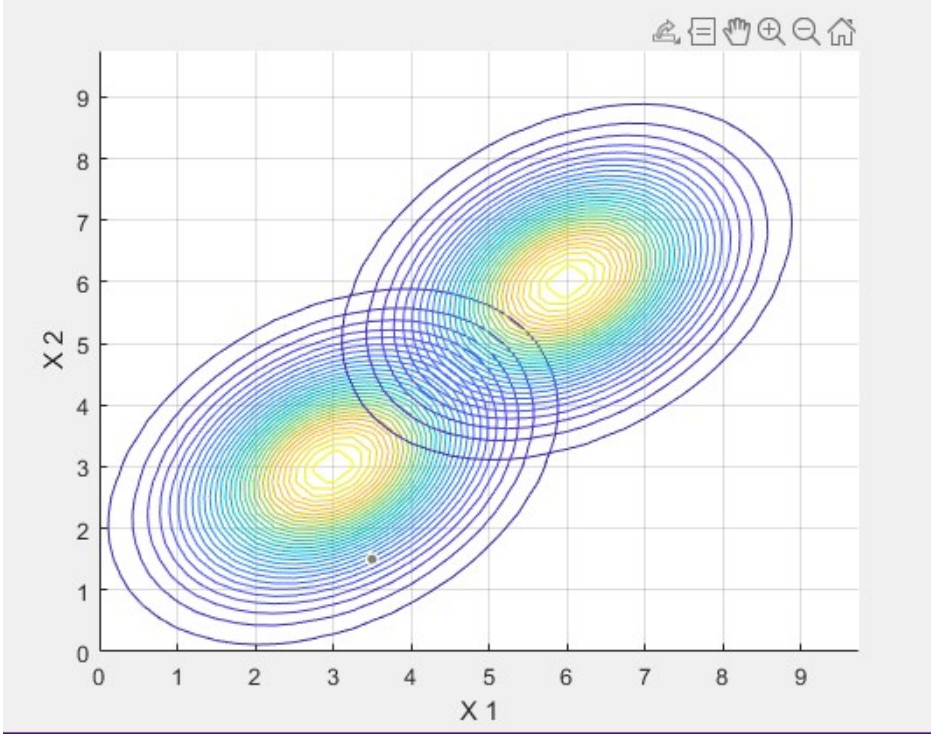
Παρατηρούμε πως καθώς αυξάνεται η *a-priori* πιθανότητα π_1 (και επομένως μειώνεται αντιστοίχως η *a-priori* πιθανότητα $P(\omega_2) = \pi_2 = 1 - \pi_1$) το σύνορο απόφασης μετακινείται προς το νέφος δεδομένων που αντιστοιχεί στην κλάση $\boldsymbol{\omega_2}$ το οποίο είναι και το λογικό (η κλάση με την μεγαλύτερη *a-priori* πιθανότητα "σπρώχνει" το decision boundary μακριά της).

d) Επαναλαμβάνουμε τα παραπάνω βήματα για $\Sigma_1=\Sigma_2=\Sigma$:

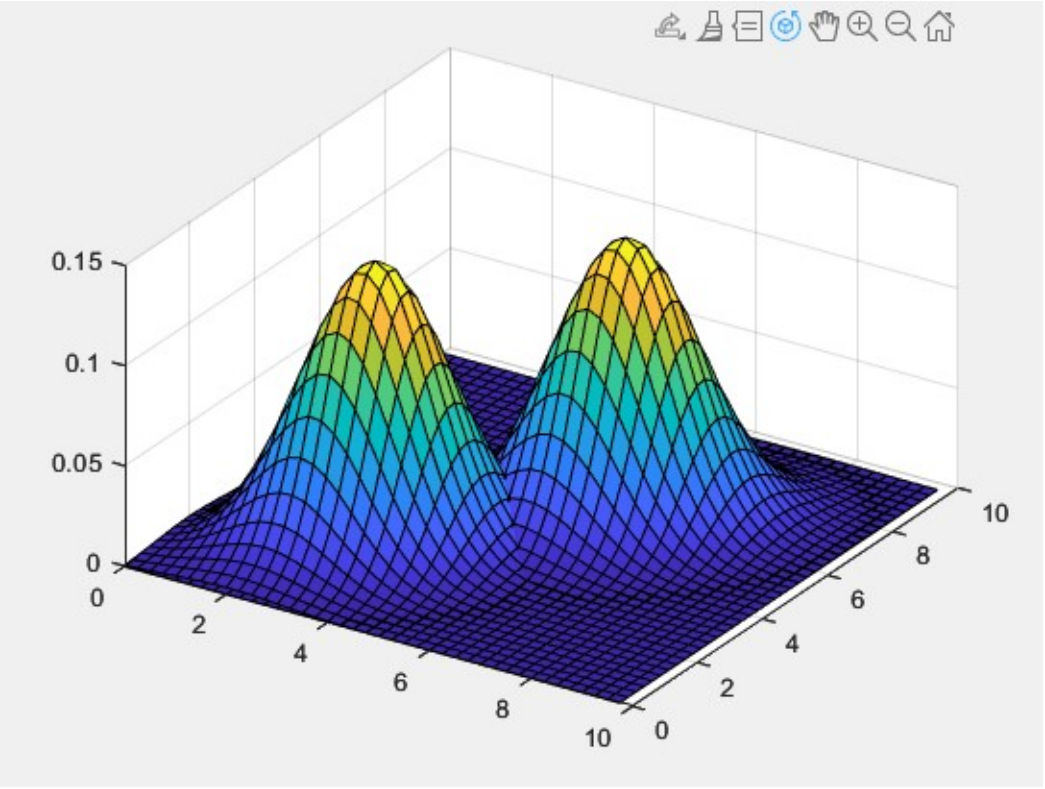
με νέο κανόνα απόφασης :

$$x_1 + x_2 = 9 - \frac{8}{15}\ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) \quad \textbf{(ευθεία)}$$

Contour plots:

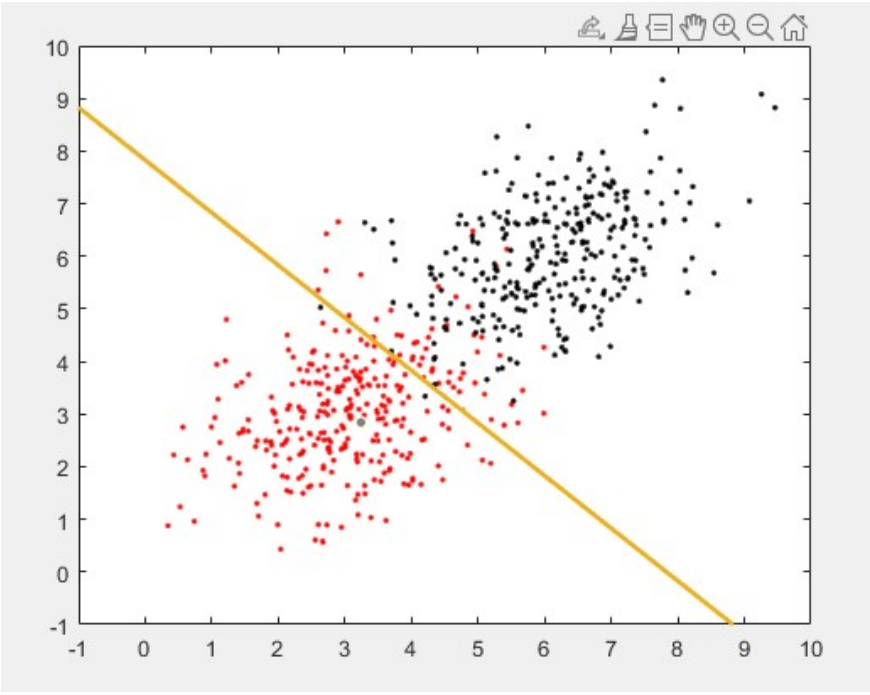


Επιφάνειες στον R^3 :

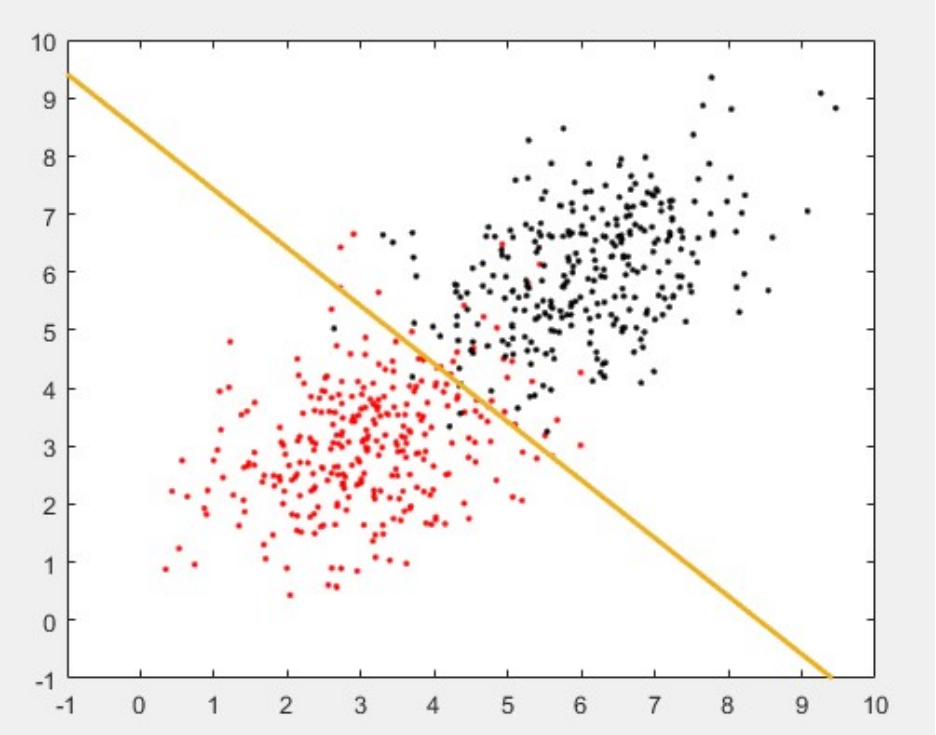


Δημιουργούμε εκ νέου δείγματα, τα απεικονίζουμε σε κοινό plot με την ευθεία και παρατηρούμε το ίδιο φαινόμενο όπως προηγουμένως, δηλαδή μια ακριβή κατηγοριοποίηση και την επιθυμητή κίνηση του decision boundary ανάλογα την π_1 .

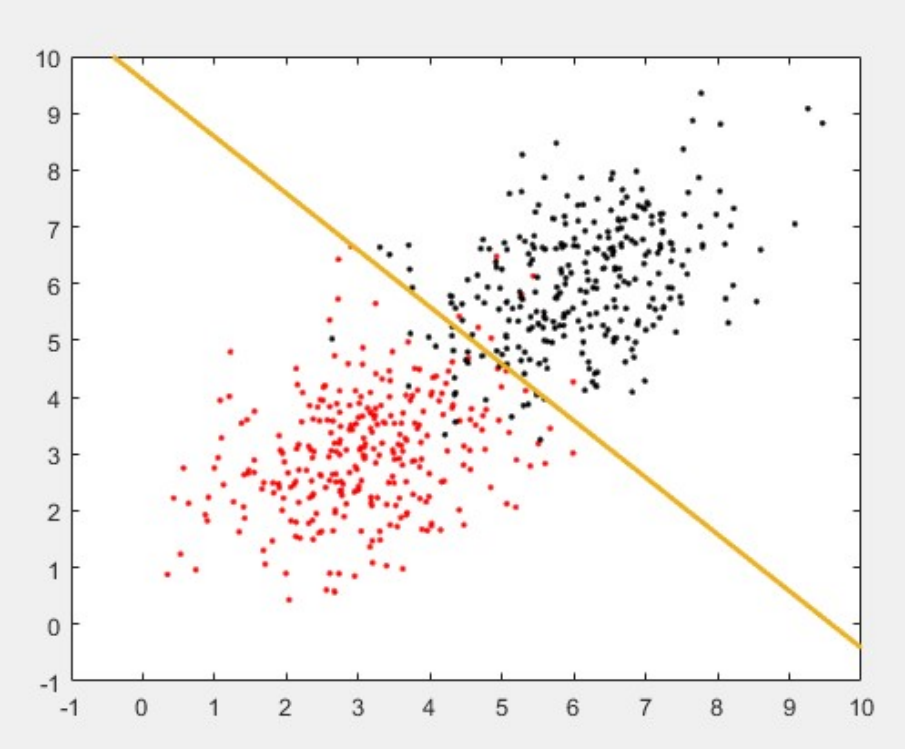
Για $\pi_1 = 0.1$:



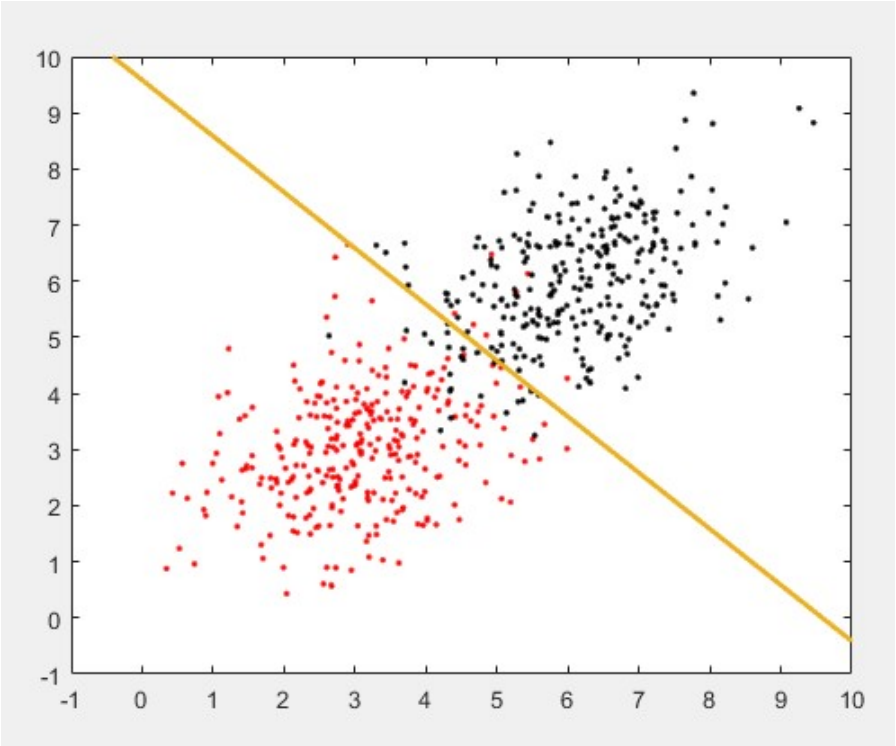
Για $\pi_1 = 0.25$:



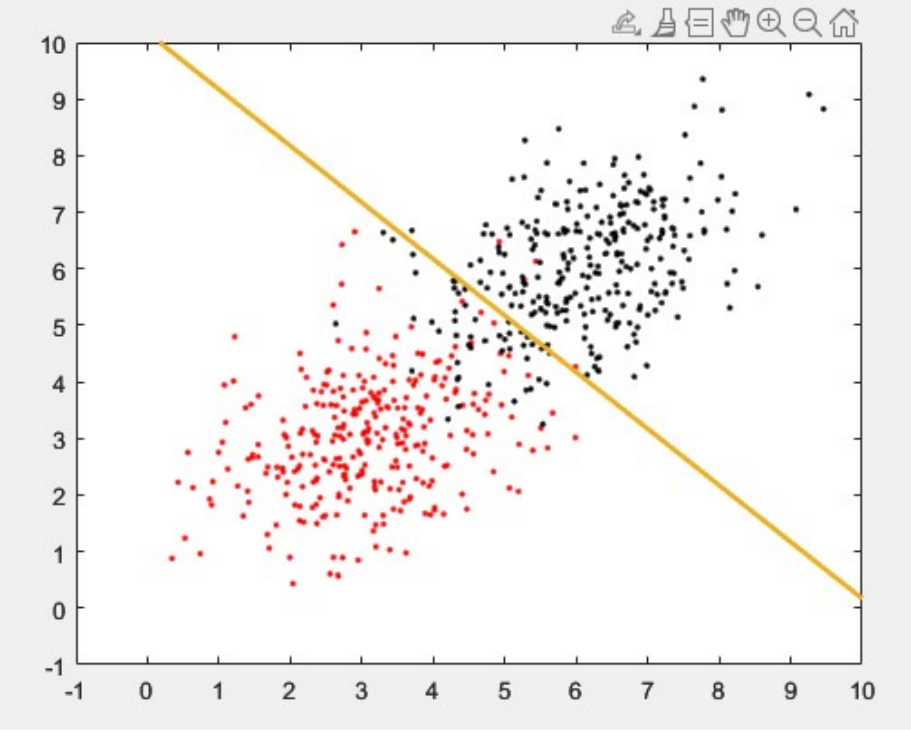
Για $\pi_1 = 0.5$:



Για $\pi_1 = 0.75$:



Για $\pi_1 = 0.9$:

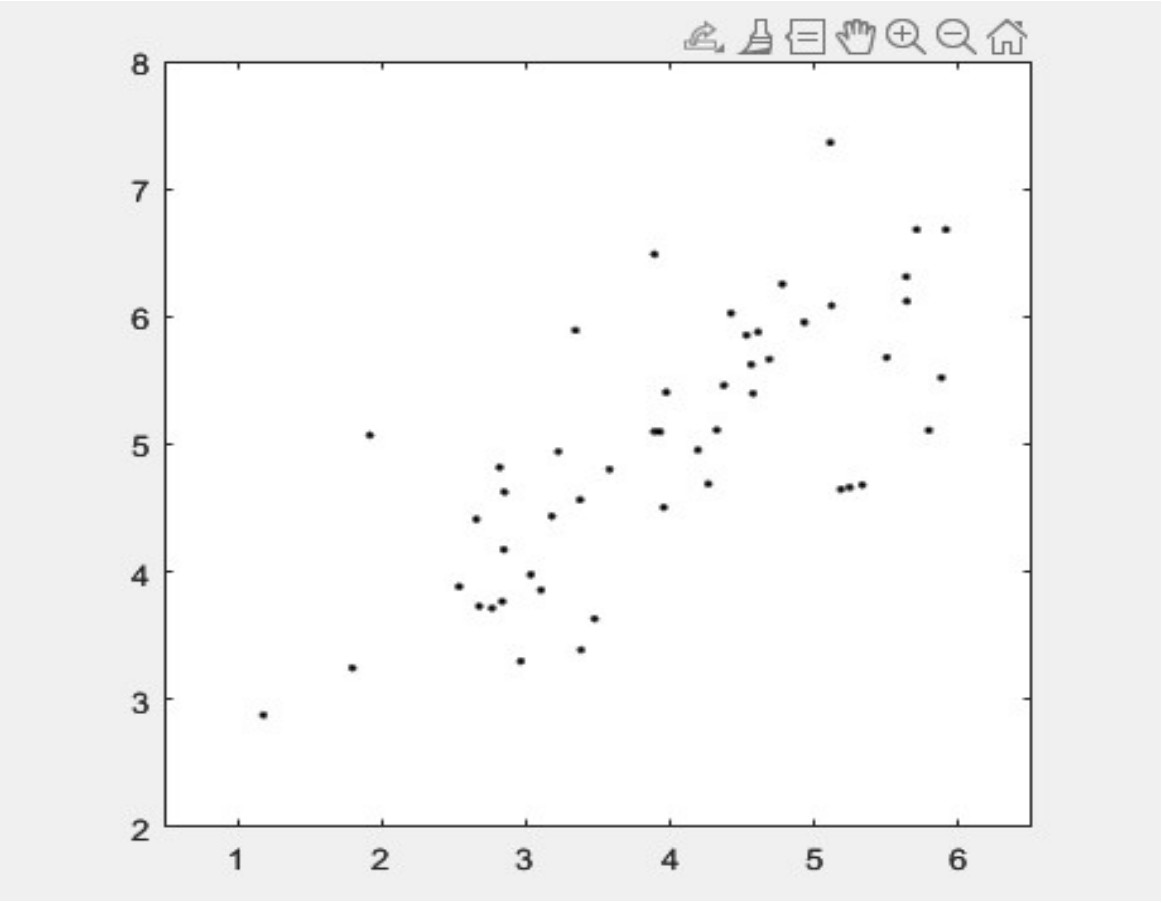


Άσκηση 5

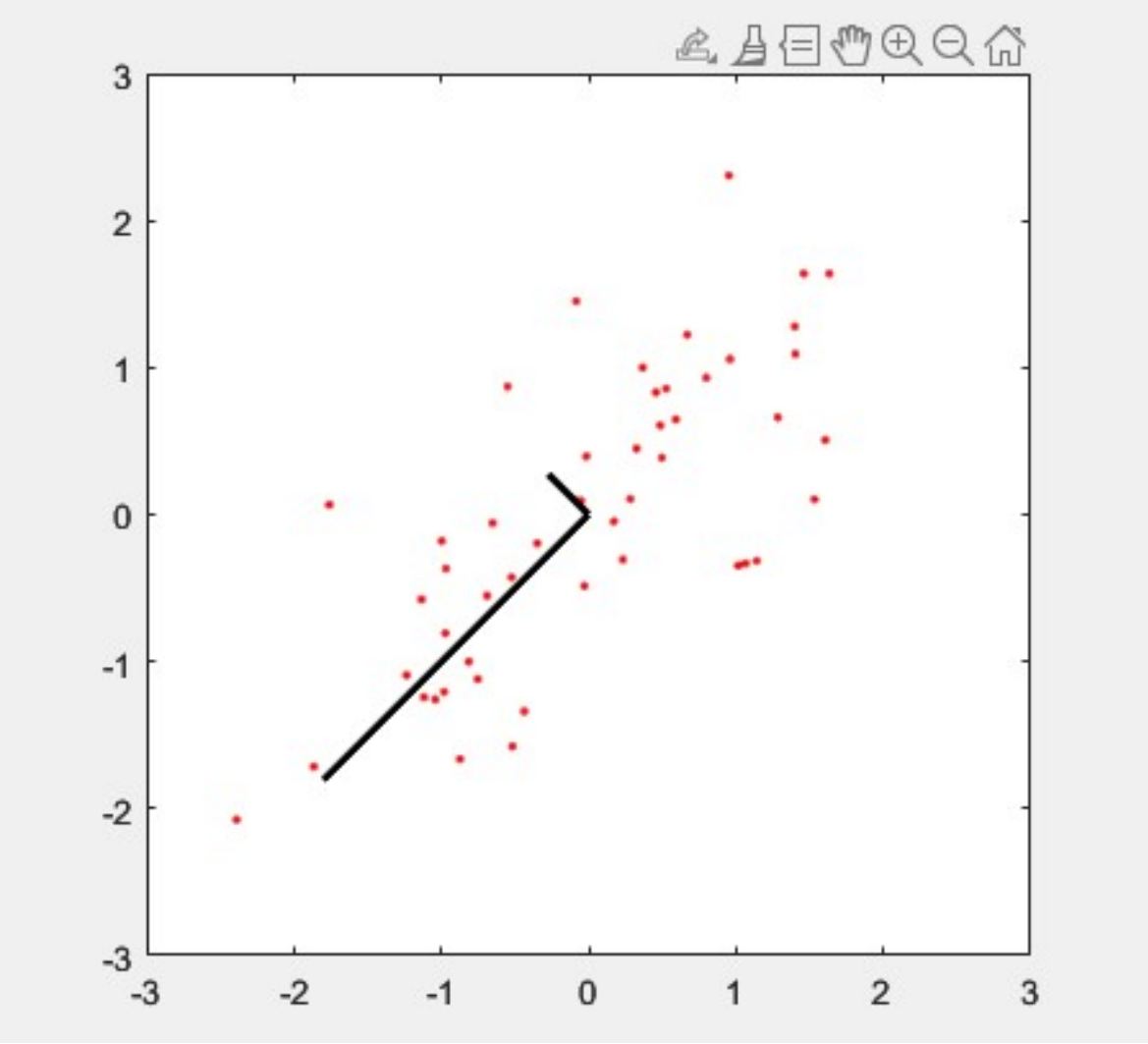
Επιθυμούμε να χρησιμοποιήσουμε την μέθοδο PCA για να μειώσουμε τις διαστάσεις , για αρχή σε ένα 2D dataset και ύστερα σε ένα dataset από εικόνες προσώπων , δηλαδή ένα dataset με πολλά features.

Μέρος 1 :
Το Dataset στο οποίο θα εφαρμόσουμε dimensionality reduction μέσω PCA

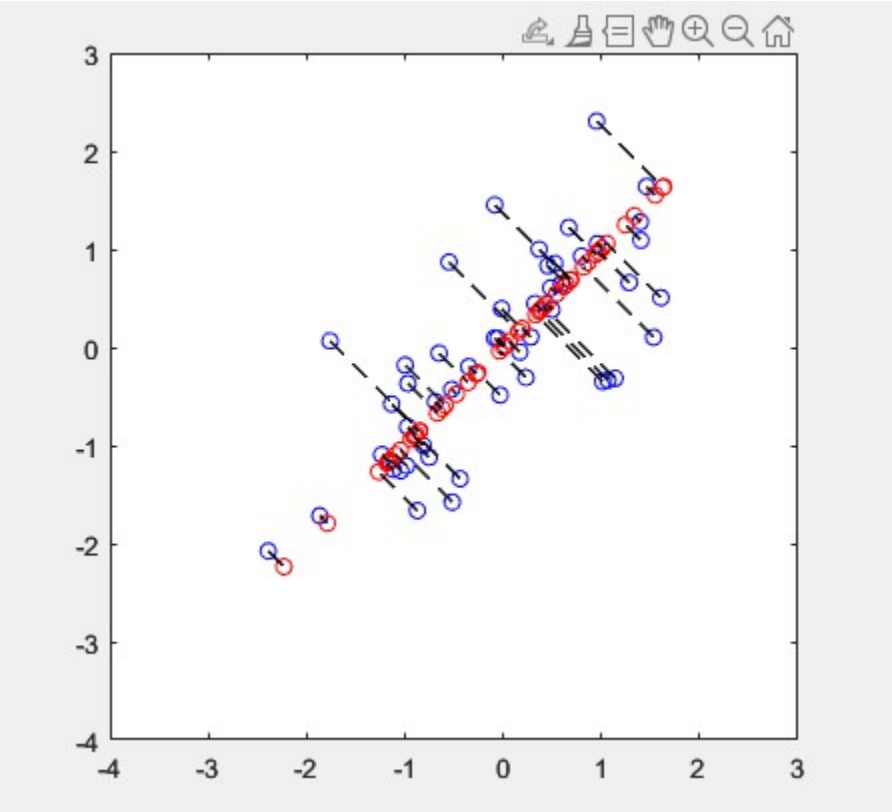
Μετά από Data Standardization (Αφαίρεση της μέσης τιμής για να έχουμε μηδενική μέση τιμή στο νέο σετ και διαίρεση με την μέση τυπική απόκλιση) :



Οι αποστάσεις των δειγμάτων μεταξύ τους δεν αλλάζει, αλλάζει μόνο η θέση τους σχετικά με τους άξονες.
Χρειαζόμαστε τον πίνακα συνδιασποράς , τις ιδιοτιμές και τα ιδιοδιανύσματά του.
Σχεδιάζουμε τα ιδιοδιανύσματα που αντιστοιχούν στις ιδιοτιμές σε κοινό plot με τα κανονικοποιημένα δείγματα :



Θα μειώσουμε τις διαστάσεις χρησιμοποιώντας το ιδιοδιάνυσμα που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή. Μετά, θα προβάλουμε τα δεδομένα πάνω σε αυτό :
$\mathbf{y = w^T x}$



Αρχικά, τα δεδομένα μας ήταν 2D άρα εύλογα το z είναι μονοδιάστατο.
Για την ανάκτηση του X θα χρησιμοποιήσουμε το μειωμένης διάστασης x :
$\mathbf{X_{revovered} = zw}$

Μέρος 2 :

Επαναλαμβάνουμε για το νέο Dataset (5000) εικόνες προσώπων. Τα πρώτα 100 πρόσωπα :

Εφαρμόζουμε standardization και PCA και κρατάμε τα 36 κυριότερα χαρακτηριστικά , που αντιστοιχούν στα ιδιοδιανύσματα των μεγαλύτερων ιδιοτιμών του πίνακα συνδιασποράς και περιέχουν τις τιμές των 1024 x 1024 pixels .

Επειδή οι εικόνες των προσώπων είναι ένα πολυδιάστατο Dataset, χρειάζονται πολλά Principal Components για την ακριβότερη αναπαράστασή τους. Αυτό είναι εμφανές όσο αυξάνεται ο αριθμός των διαστάσεων που θα διατηρηθούν.

Για λιγότερες διαστάσεις παρατηρούμε ότι απεικονίζονται μόνο τα βασικότερα χαρακτηριστικά των προσώπων, ενώ όσο αυξάνεται ο αριθμός διαστάσεων οι εικόνες που ανακτάμε γίνονται όλο και πιο ακριβείς.

K=36 :

K = 50 :


Original faces


Recovered faces

K=100


Original faces


Recovered faces

K=200


Original faces


Recovered faces

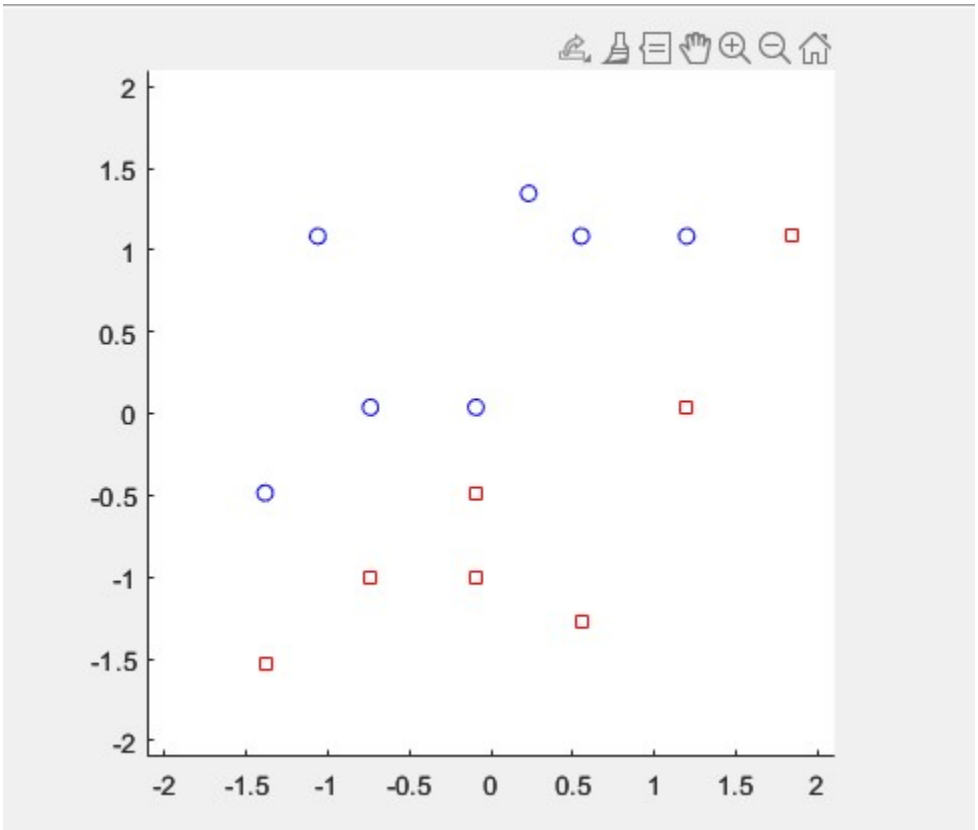
K=500


Original faces


Recovered faces

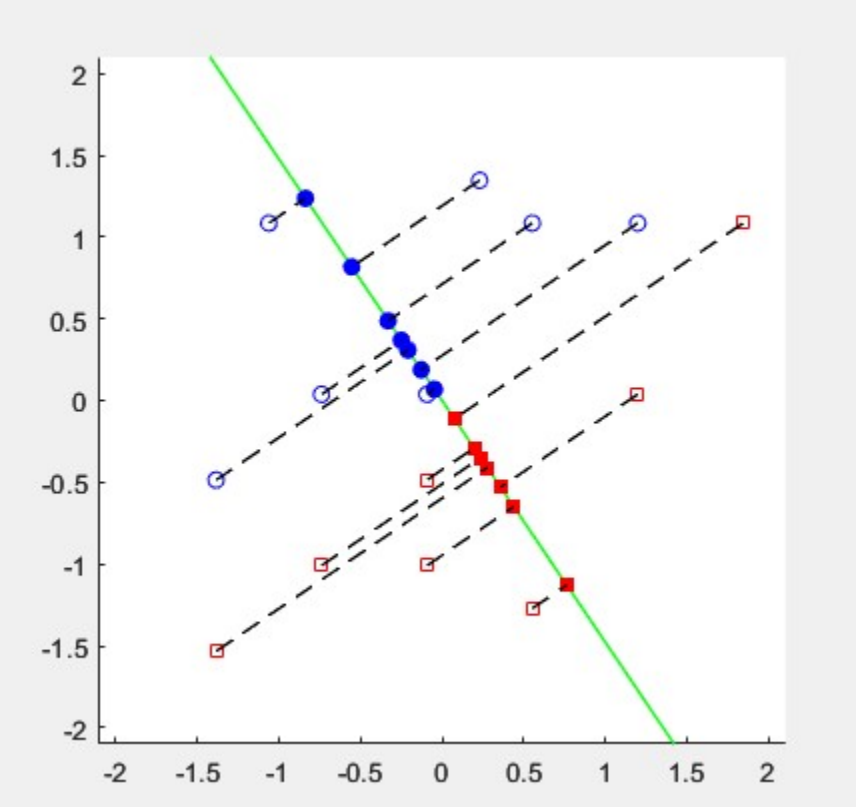
Άσκηση 7

Μέρος 1 :
Ομοίως με την προηγούμενη άσκηση , εκτελούμε standardization και εφαρμόζουμε LDA σε ένα 2D Dataset. Θέλουμε να κατηγοριοποιήσουμε τα δεδομένα σε δύο κλάσεις.
Τα δεδομένα :

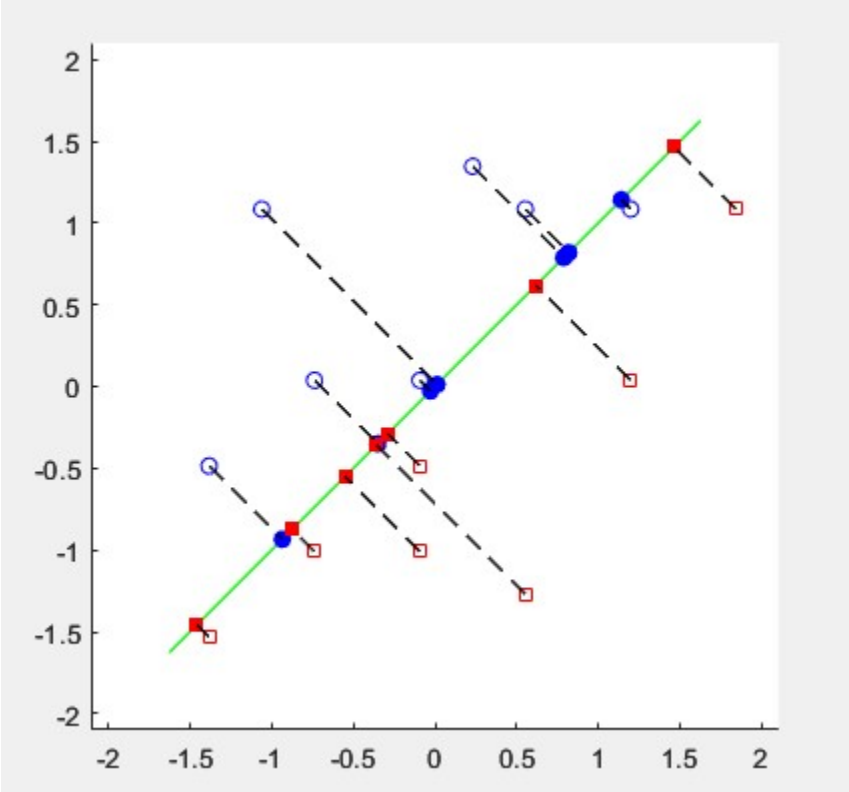


Ο βέλτιστος κανόνας σύμφωνα με τον LDA αλγόριθμο είναι :

$z = w^T x$ με $w = \beta (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2)$, $\beta > 0$
Προβάλουμε τα δεδομένα στην ευθεία :



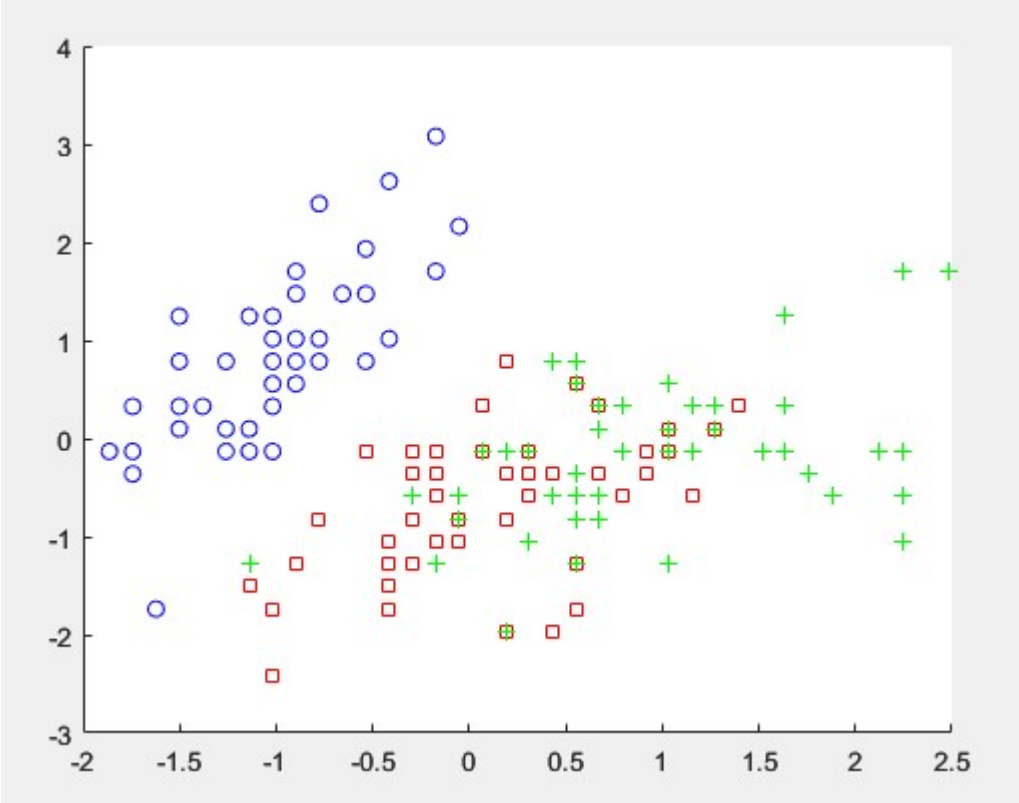
Για να συγκρίνουμε τον αλγόριθμο LDA με τον PCA , προβάλουμε και στην ευθεία που παράγει ο PCA :



Παρατηρούμε ότι ο κανόνας που απορρέει από τον LDA είναι ο βέλτιστος , εφόσον μεγιστοποιείται η απόσταση των κλάσεων μεταξύ τους, ενώ τα δεδομένα που ανήκουν στην ίδια κλάση είναι κοντινότερα μεταξύ τους. Ο PCA αλγόριθμος σε αυτή την περίπτωση θα ήταν κακός και η κατηγοριοποίηση θα οδηγούσε σε πολλά σφάλματα.

Μέρος 2 :

Σε αυτή τη φάση της άσκησης θα εφαρμόσουμε τον αλγόριθμο LDA σε ένα ευρέως γνωστό Dataset, το Dataset Iris. Απεικονίζουμε αρχικά τα πρώτα 2 features κάθε κλάσης , μετά από standardization :



Για να συνεχίσουμε , χρειαζόμαστε :

Within-Class Scatter Matrix :

$\mathbf{S_w = \sum P_i \Sigma_i}$

Between-Class Matrix :

$\mathbf{S_b = \sum P_i(\mu_i - \mu_0)(\mu_i - \mu_0)^T}$ **με i=0,1,..,c και c = numOfClasses**

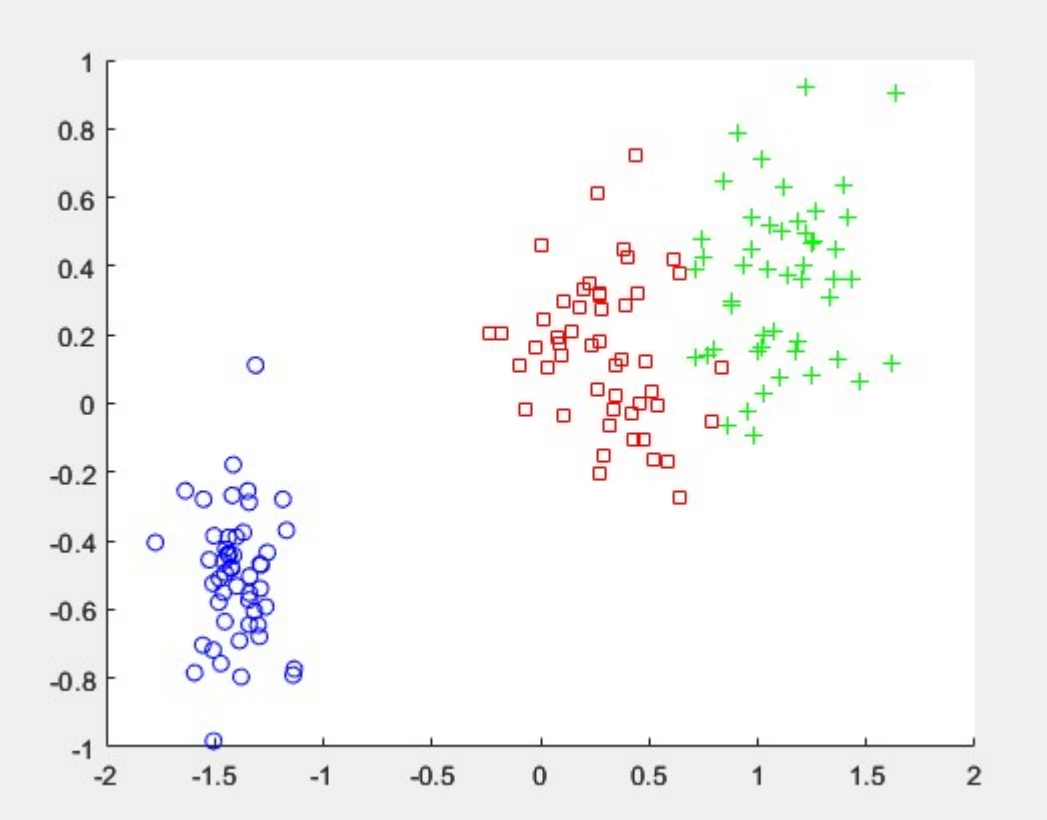
Global mean :

$\mathbf{\mu_0 = (1/N) \sum n_i \mu_i}$

εφαρμόζουμε μείωση διάστασης :

$\mathbf{z = W^T x}$, όπου ο W , όπως πριν αποτελείται από τους Sb, Sw^{-1} (Περιέχει τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές του γινομένου τους)

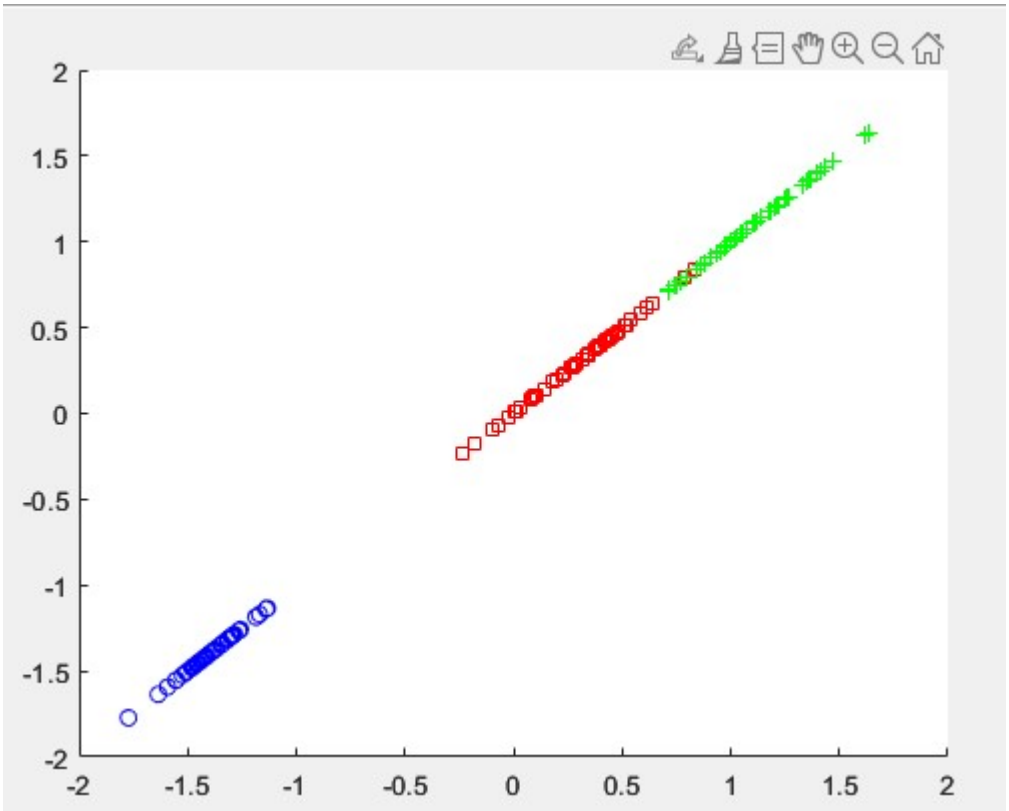
Έτσι , μειώνονται οι διαστάσεις από 4 σε 2 :



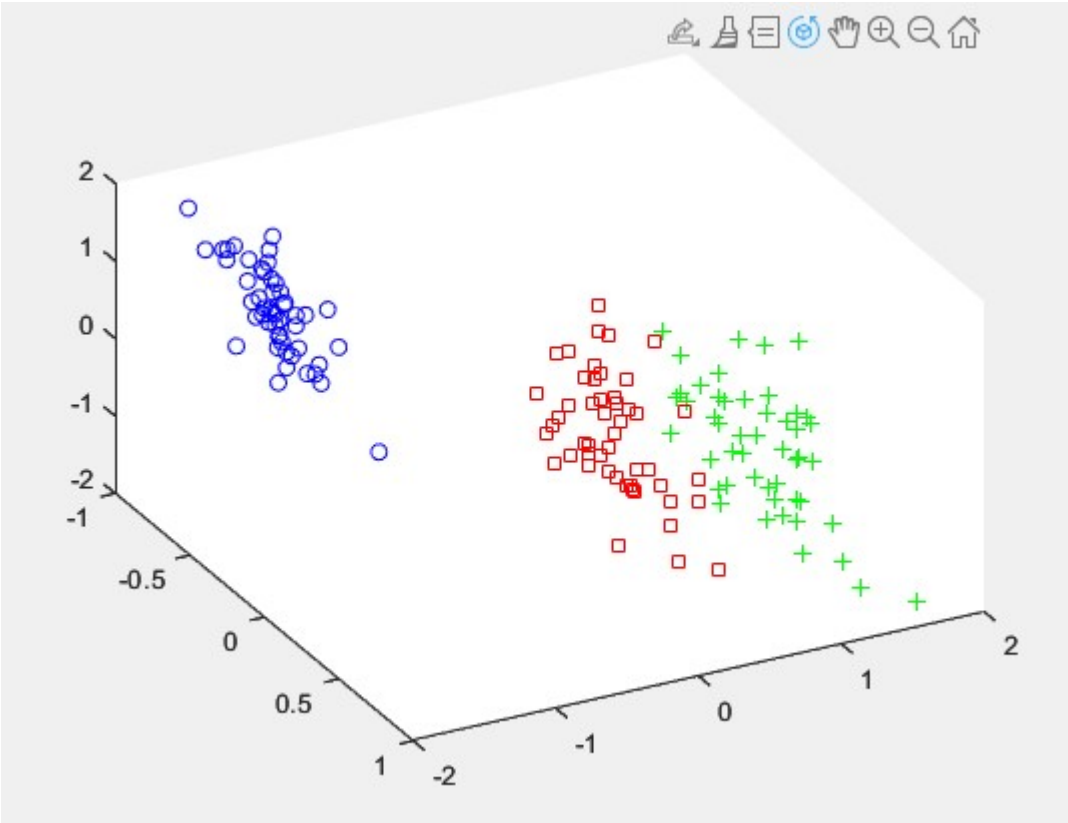
Παρατηρούμε ότι η ταξινόμηση θα έχει μικρό σφάλμα.

Σε άλλες περιπτώσεις :

Από 4 διαστάσεις σε 1 :



Από 4 σε 3 :



Οι ασκήσεις 1,3,4,6 βρίσκονται αναλυτικότερα σε pdf που επισυνάπτεται στο ανεβασμένο zip αρχείο στο e-class , εφόσον περιέχουν μόνο θεωρητικό κομμάτι.