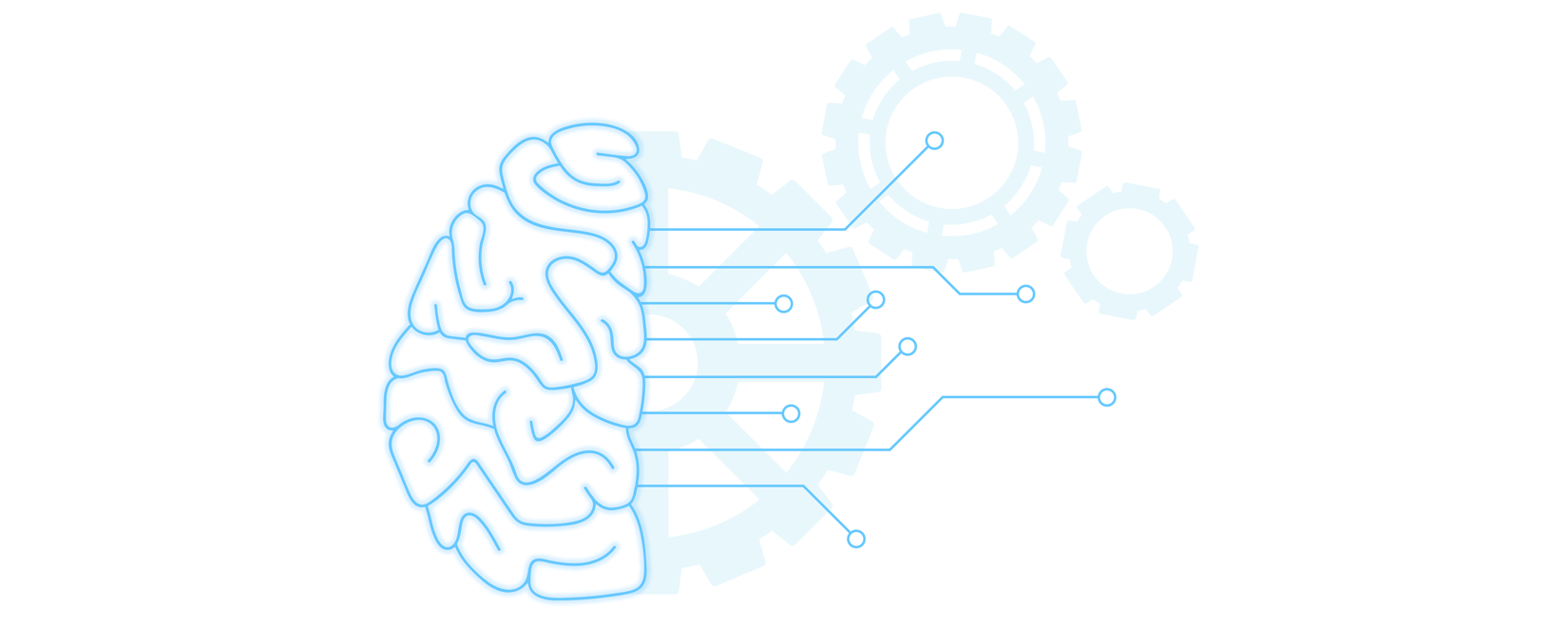
ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ  
  
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧ. ΚΑΙ ΜΗΧ. ΥΠΟΛΟΓΙΣΤΩΝ

# ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΑΙ ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

# ΤΗΛ 311



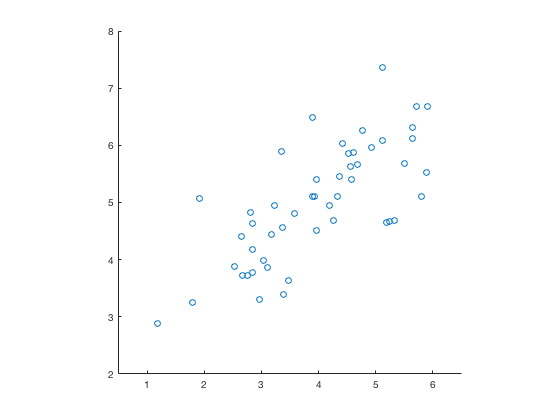
## 1η ΣΕΙΡΑ ΑΣΚΗΣΕΩΝ

## ΑΝΑΦΟΡΑ

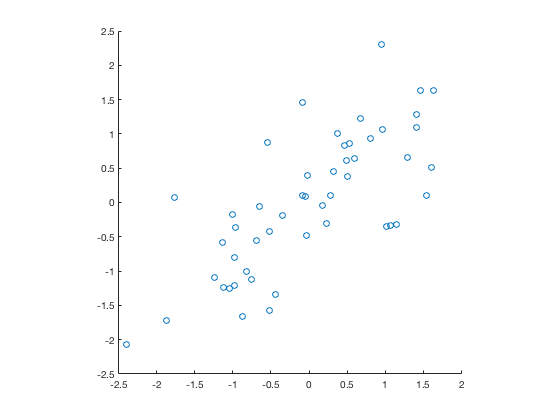
## ΚΥΡΙΑΖΑΚΗΣ ΚΛΕΑΝΘΗΣ – 2015030086

## **ΘΕΜΑ 1: Principal Component Analysis**

**Μέρος 1**

**a)** Στο πρώτο μέρος καλούμαστε να υλοποιήσουμε την μέθοδο PCA έτσι ώστε να μειώσουμε τις διστάσεις των δεδομένων μας. Αρχικά ξεκινάμε φορτώνοντας το αρχείο 'ex1\_1\_data1.mat' το οποίο περιέχει 50 δείγματα με 2 feature το καθένα. Έπειτα εμφανίστηκαν στον δισδιάστατο χώρο με την εντολή scatter:

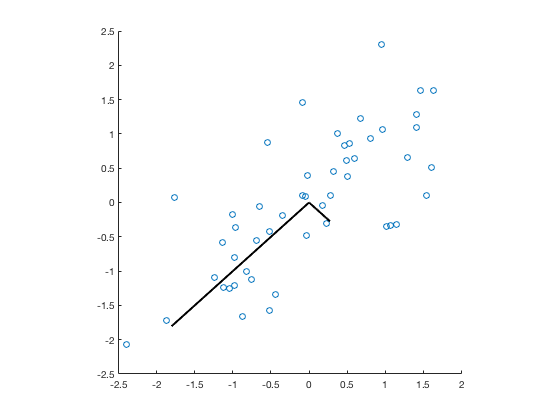
**b)** Πριν γίνει κάποια επεξεργασία στα δεδομένα, πρέπει να γίνει κανονικοποίηση με μέση τιμή μηδέν και διασπορά ένα έτσι ώστε να υπάρχει μία κοινή κλίμακα για όλα τα δεδομένα. Αυτό γίνεται στην συνάρτηση ‘featureNormalize’ μέσα στην οποία υπολογίζεται η μέση τιμή και τυπική απόκλιση των δεδομένων, έπειτα γίνεται η πράξη η οποία φέρνει τις τιμές στο επιθυμητό εύρος τιμών χωρίς να χάσουμε πληροφορία.



Στην παραπάνω εικόνα φαίνονται τα κανονικοποιημένα δείγματα. Παρατηρείται πως είναι ακριβώς ίδιο ‘pattern’ με τα αρχικά δείγματα, μόνο που πλέον οι τιμές βρίσκονται γύρω από το μηδέν και στους δύο άξονες.

**c)** Στη συνέχεια, περνάμε τα κανονικοποιημένα δεδομένα στην συνάρτηση ‘myPCA’ η οποία θα υπολογίσει τις ιδιοτιμές τα ιδιοδιανίσματα. Με την έτοιμη συνάρτηση του MATLAB ‘eig’ επιστρέφονται οι ιδιοτιμές στην διαγώνιο της μεταβλητής D και τα ιδιοδιανίσματα στην μεταβλητή V. Παρατηρείται πως στην συγκεκριμένη περίπτωση οπού έχουμε δυο features, οι μεταβλητές V και D είναι πίνακες διαστάσεων 2x2. Έπειτα, γίνεται ταξινόμηση των ιδιοτιμών κρατώντας τα index έτσι ώστε να γίνει αντίστοιχη ταξινόμηση και στα ιδιοδιανίσματα. Τέλος, η συνάρτηση επιστρέφει τις μεταβλητές που υπολόγισε (Τις ιδιοτιμές στον πίνακα S διάστασης 2x1 διότι δεν είναι πλέον στην διαγώνιο, και τα ιδιοδιανίσματα στον πίνακα U διάστασης 2x2).

Έπειτα, συνεχίζοντας την ροή του προγράμματος, εκτυπώνονται στο διάγραμμα με τις κανονικοποιημένες τιμές και τα ιδιοδιανίσματα PC1 και PC2 όπως φαίνεται στην παρακάτω εικόνα:



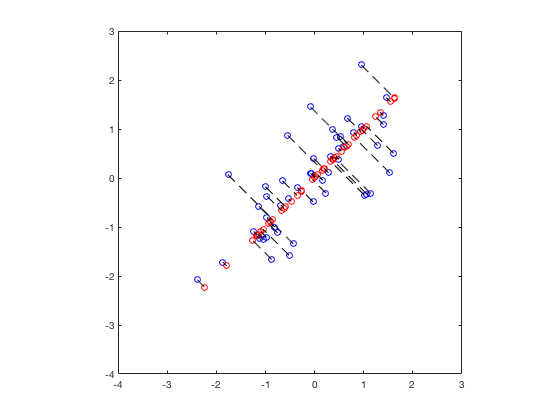
**d)** Συνεχίζοντας με τον υπολογισμό της συνεισφοράς της κάθε κυρίας συνιστώσας στην συνολική διακύμανση παίρνουμε τα εξής αναμενόμενα αποτελέσματα:

PCvariance(1,1) = 0.8678

PCvariance(2,1) = 0.1322

Ο στόχος μας πλέον είναι να μειώσουμε τις δύο διαστάσεις σε μία, και αυτό θα το πετύχουμε καλώντας την συνάρτηση ‘projectData’. Οπότε θέτουμε την μεταβλητή Κ = 1 (που συμβολίζει την επιθυμητή διάσταση) και την περνάμε σαν όρισμα στην συνάρτηση μαζί με τα κανονικοποιημένα δεδομένα καθώς και την μεταβλητή που περιέχει τα ιδιοδιανίσματα. Έτσι, επιστρέφονται οι καινούριες τιμές στην μεταβλητή Ζ η οποία έχει διάσταση 50x1 σε αντίθεση με την X\_norm η οποία είχε 50x2.

**e)** Τώρα που έχουμε τα δεδομένα σε χώρο μίας διάστασης, μπορούμε αν θέλουμε να προσεγγίσουμε ξανά τα δεδομένα δύο διαστάσεων που είχαμε πριν. Αυτό υλοποιείται στην συνάρτηση ‘recoverData’ στην οποία ουσιαστικά προβάλλουμε τα δεδομένα στην Ζ σε όλα τα ιδιοδιανίσματα (Στην περίπτωσή μας PC1, PC2). Στην συνέχεια εκτυπώνονται σε κοινό διάγραμμα τα αρχικά δείγματα, τα δείγματα που προβάλλαμε στην κύρια συνιστώσα καθώς και γραμμές που τα ενώνουν.



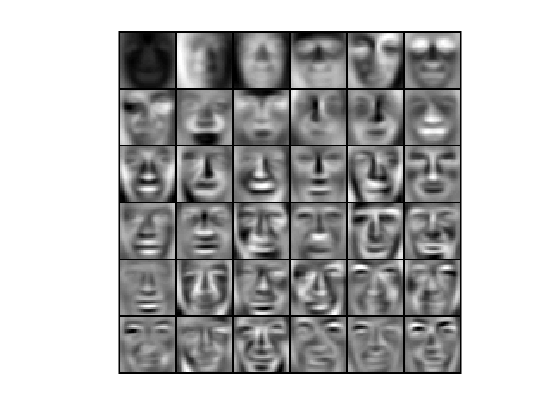
Παρατηρώντας την εικόνα συμπεραίνουμε πως η διαδικασία PCA για μείωση διαστάσεων ήταν επιτυχής αφού όλα τα δείγματά μας προβλήθηκαν στην ευθεία κατά την οποία τα δεδομένα μας είχαν την μέγιστη διασπορά.

**Μέρος 2**

Στο δεύτερο μέρος του θέματος πρέπει να επαναλάβουμε την ίδια διαδικασία του πρώτου μέρους χρησιμοποιώντας αυτή την φορά τα δεδομένα 5000 εικόνων στο αρχείο ‘ex1faces.mat’. Σκοπός είναι να παραμετροποιηθεί ο κώδικας του προηγούμενου μέρους για να μπορεί να χρησιμοποιηθεί σε όλες τις περιπτώσεις.

Στην συγκεκριμένη περίπτωση έχουμε 5000 εικόνες μεγέθους 32x32 που σημαίνει ότι έχουμε 1024 features. Προφανώς δεν θα μας είναι όλα χρήσιμα και θα χρειαστεί να γίνει μείωση της διάστασης έτσι ώστε να παραμείνουν μόνο τα πιο σημαντικά principal components. Αυτό σημαίνει ότι ο αλγόριθμος που αναπτύξαμε προηγουμένως θα δέχεται πληροφορία 1024 διαστάσεων και θα επιστρέφει πληροφορία μειωμένης διάστασης ανάλογα με την τιμή που θα αναθέσουμε στην μεταβλητή Κ.

**f)** Αφού φορτωθεί το dataset, γίνεται εμφάνιση των 100 πρώτων εικόνων. Η διαδικασία αυτή γίνεται με την εντολή ‘load('ex1\_1\_faces.mat')’ και έπειτα ‘displayData(X(1:100, :))’ και παράγεται το παρακάτω αποτέλεσμα:

**g)** Στην συνέχεια, όπως και πριν, περνάμε τα δεδομένα από την ‘featureNormalize’ για την κανονικοποίηση και μετά με την συνάρτηση ‘myPCA’ παίρνουμε τα ιδιοδιανίσματα και τις ιδιοτιμές. Στο κοινό διάγραμμα παρακάτω εμφανίστηκαν οι πρώτες 36 κύριες συνιστώσες που βρέθηκαν.

Παρατηρούμε πως όλες οι συνιστώσες μοιάζουν με πρόσωπα και μας δείχνουν τα πιο σημαντικά ‘χαρακτηριστικά’ στις εικόνες

**h)** Σε αυτό το ερώτημα, γίνεται μείωση διαστάσεων με προβολή των σημείων στις Κ κύριες συνιστώσες. Χρησιμοποιήθηκε Κ=20, Κ=100 και Κ=200 για να δούμε τυχών διαφορές. Άρα, σε αντίθεση με τα αρχικά δεδομένα που ήταν σε πίνακα διαστάσεων 5000x1024 τώρα θα έχουμε πίνακα διαστάσεων 5000xK (όπου Κ=20 ή 100 ή 200).

**i)** Πλέον, αφού έχουμε ελαττώσει τις διαστάσεις των feature μας, μπορούμε με την χρήση της ‘recoverData’ να κάνουμε ανάκτηση των αρχικών δειγμάτων (με 1024 διαστάσεις) και να συγκρίνουμε τυχών αποκλίσεις.

Κ = 20



Κ = 100



Κ = 200

Παρατηρούμε πως όσο το Κ αυξάνεται τόσο περισσότερη πληροφορία κρατάμε για το κάθε πρόσωπο το οποίο είναι αναμενόμενο. Όσο μικραίνει ο αριθμός των διαστάσεων που κρατάμε τόσο λιγότερο χώρο και επεξεργαστική ισχύ απαιτεί η επεξεργασία των δεδομένων.