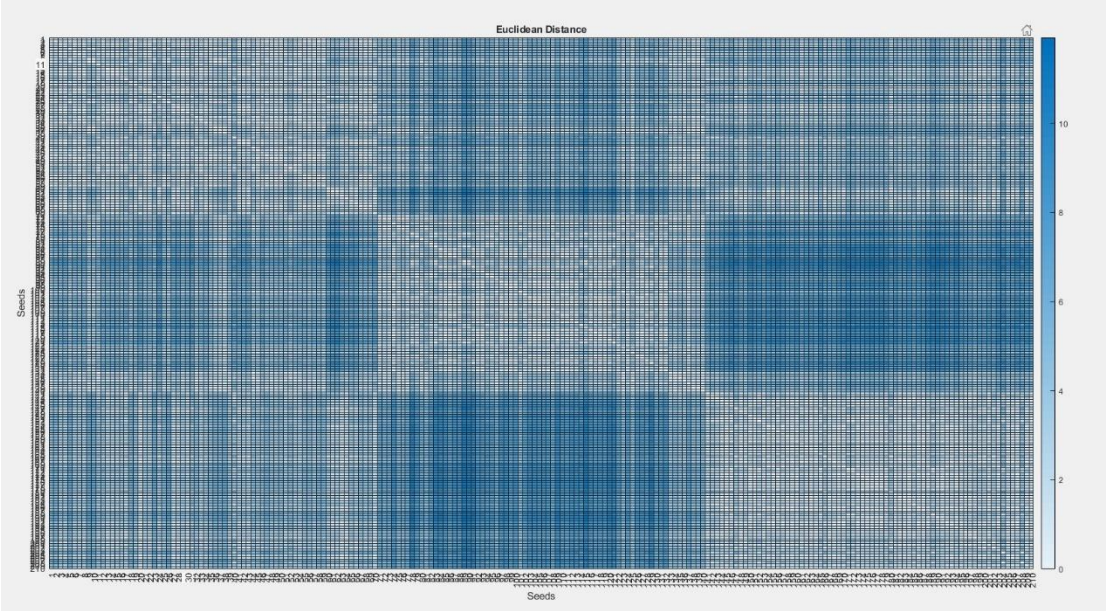

| ΑΦΜ | ΕΠΩΝΥΜΟ | ΟΝΟΜΑ | ΕΞΑΜΗΝΟ |
|---|---|---|---|
| **58352** | ΤΟΚΑΤΛΙΔΗΣ | ΓΕΩΡΓΙΟΣ | 9$^{Ο}$ |

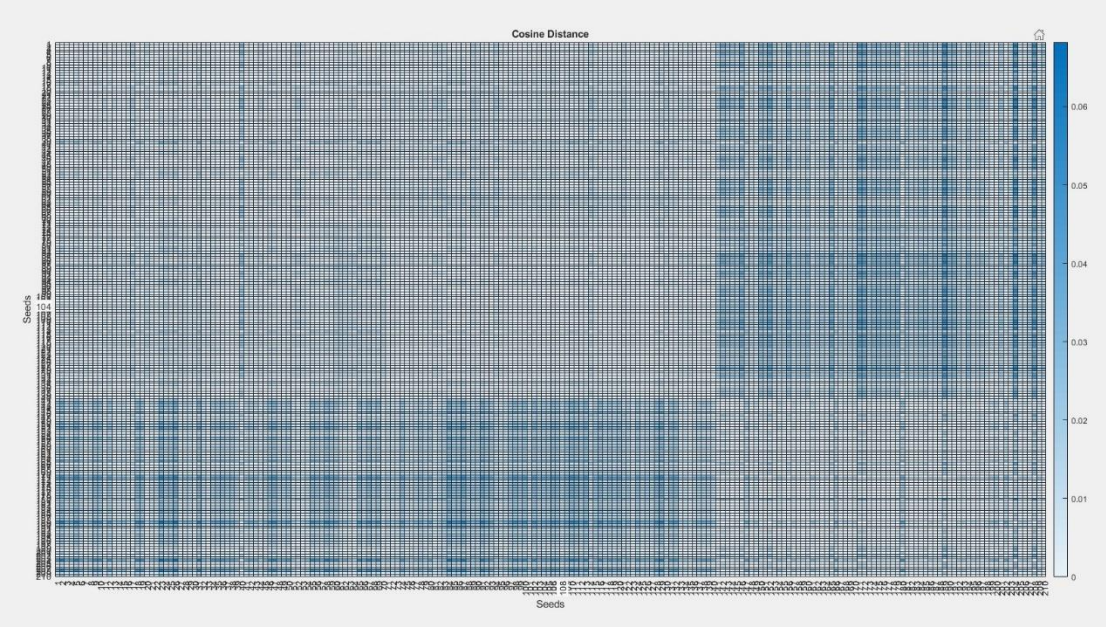
# ΕΡΓΑΣΙΑ #3

## ΜΑΘΗΜΑ : ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

## ΑΣΚΗΣΗ 1

Στο πρώτο ερώτημα της άσκησης ζητείται να βρεθούν οι αποστάσεις μεταξύ όλων των ζευγών του dataset και να απεικονιστούν με τον πίνακα απόστασης, όπου κάθε γραμμή αφορά το στοιχείο – κέντρο για το οποίο υπολογίζονται οι αποστάσεις όλων των άλλων σημείων - δειγμάτων από αυτό. Κάθε στήλη αφορά το ι-οστό δείγμα για το οποίο βρίσκεται η απόσταση από το κεντρικό δείγμα. Λαμβάνοντας επίσης υπόψιν πως οι μετρικές απόστασης έχουν μέρος των ιδιοτήτων τους πως d(x,y) = d(y,x) και ότι d(x,x) = 0, ο πίνακας απόστασης θα είναι συμμετρικός με μηδενική διαγώνιο.  Πράγματι,




σχηματίζονται συμμετρικά χρωματικά μοτίβα ανά τμήματα στον πίνακα που οπτικοποιούν ποιοτικά αυτές τις ιδιότητες. Ο λόγος που διαφέρουν αυτά μεταξύ των διαφορετικών μετρικών είναι διότι στην ευκλείδεια απόσταση λογίζεται η απόσταση των σημείων, ενώ στην απόσταση συνημιτόνου η διεύθυνση του εκάστοτε ζεύγους διανυσμάτων.

Στα διαγράμματα, το σκούρο μπλε εκφράζει τη μέγιστη τιμή που παρατηρήθηκε στον πίνακα απόστασης, ενώ το λευκό μηδενική απόσταση. Από τα αποτελέσματα φαίνεται

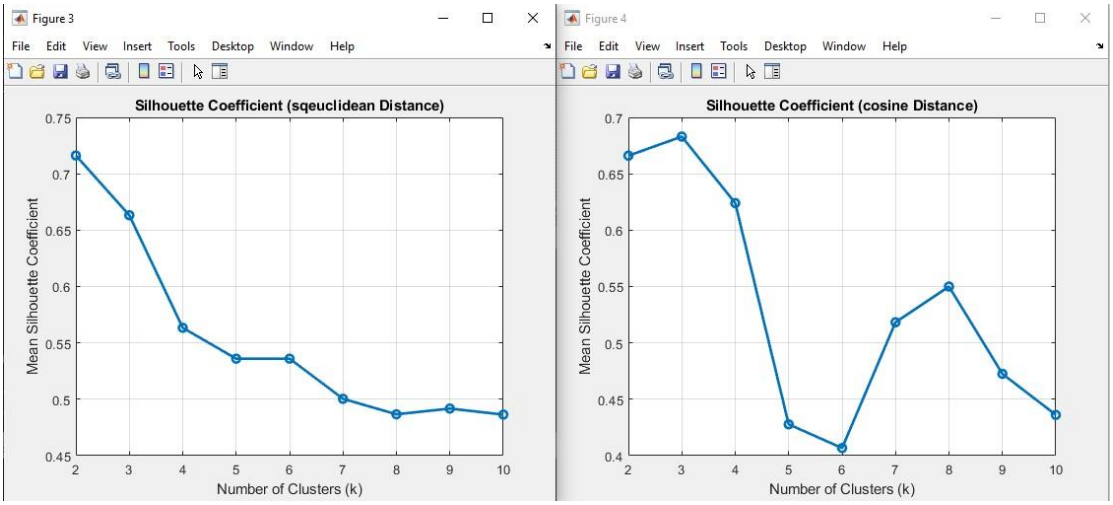
- Για την ευκλείδεια μετρική, πως τα δειγματικά σημεία μεταξύ κλάσεων είναι αρκετά μακριά μεταξύ τους με τα ζεύγη 1-3 και 2-3 να παρουσιάζουν τη μικρότερη και μεγαλύτερη ομοιότητα αντίστοιχα. Επίσης αρκετά ζεύγη σημείων μεταξύ της 1-3 βρίσκονται αρκετά κοντά

- Για την απόσταση συνημιτόνου, φαίνεται πως τα σημεία δεν είναι ιδιαίτερα απλωμένα και παρουσιάζουν σημαντική κατευθυντικότητα, καθώς η μεγαλύτερη απόκλιση στη διεύθυνση που παρουσιάζεται είναι περίπου 0.07 που αντιστοιχεί σε γωνία 21 μοιρών. Συγκριτικά βέβαια η κλάση 3 είναι σε ελαφρώς διαφορετική διεύθυνση από αυτή των κλάσεων 1 και 2.

Συμπερασματικά οι πιο εύκολα διαχωρίσιμες κλάσεις φαίνεται να είναι οι κλάσεις 2-3.

Σύμφωνα με αυτά τα στοιχεία είναι δυνατό να θεωρηθεί πως οι κλάσεις 1 και 2 ή 1 και 3 ίσως μπορούν να ομαδοποιηθούν σε μια ενιαία κλάση ή πιο σωστά, cluster-συστάδα. Αυτό θα διαπιστωθεί και από τα διαγράμματα silhouette coefficient. Το μέγεθος αυτό εκφράζει κατά πόσο ένα σημείο ανήκει στην συστάδα που ταξινομήθηκε (απόσταση από το κέντρο του) και κατά πόσο απέχει από τις υπόλοιπες. Παίρνοντας τον μέσο όρο κάθε σημείου της ίδιας ομάδας βρίσκουμε το silhouette coefficient ανά cluster και κατ' επέκταση ο μέσος όρος αυτών (ίδιο με τον μέσο όρο των silhouette coefficients όλων των δειγμάτων) δίνει το συνολικό silhouette coefficient. Όσο μεγαλύτερο λοιπόν, τόσο καλύτερη διαχωρισιμότητα υπάρχει μεταξύ των clusters που ορίστηκαν προηγουμένως από κάποια clustering τεχνική.

Στα παρακάτω διαγράμματα αξιοποιείται η μετρική Squared Euclidean, όπως και cosine distance (για κανονικοποιημένα δεδομένα) για k-means clustering και για την εύρεση του silhouette coefficient, έτσι ώστε να είναι αξιόπιστα τα αποτελέσματα, δηλαδή να γίνεται χρήση της ίδιας μετρικής απόστασης κάθε φορά.




Σύμφωνα με τη μετρική Squared Euclidean το βέλτιστο k (αριθμός clusters) είναι 2. Για την απόσταση συνημιτόνου βέλτιστο k είναι το 3 με μικρή διαφορά από το k=2.

Τώρα θα υπολογιστεί το Rand Index για 5 διαφορετικά clustering με kmeans, ένα μέτρο που χρησιμοποιείται για να αξιολογηθεί η ομοιότητα μεταξύ δύο κατατάξεων δεδομένων, συνήθως μεταξύ της προβλεπόμενης ομαδοποίησης και της πραγματικότητας (ground truth). Υπολογίζεται με βάση τις ζεύξεις δεδομένων και μετράει πόσο συμφωνούν οι δύο κατατάξεις όσον αφορά την τοποθέτηση των δεδομένων σε ομάδες. Για κάθε ζεύγος σημείων, ελέγχει αν τα σημεία ανήκουν στην ίδια ομάδα και στις δύο κατατάξεις ή αν ανήκουν σε διαφορετικές ομάδες και στις δύο κατατάξεις. Το αποτέλεσμα του Rand Index κυμαίνεται από 0 έως 1, όπου το 1 σημαίνει πλήρη συμφωνία μεταξύ των δύο κατατάξεων (είναι ταυτόσημες) και το 0 πλήρη διαφωνία (εντελώς διαφορετικές).

```
Mean Rand Index (Squared Euclidean): 0.8732
Variance of Rand Index (Squared Euclidean): 0.0000
Mean Rand Index (Cosine): 0.8223
Variance of Rand Index (Cosine): 0.0084
```

Βάση και του μέσου όρου και της διακύμανσης η Squared Euclidean είναι πιο αξιόπιστη καθώς είναι πολύ κοντά στην εκτιμώμενη τιμή του rand index πάντα και είναι υψηλότερη από αυτές της cosine μετρικής, πράγμα αναμενόμενο σύμφωνα με την κατευθυντικότητα των δεδομένων που διαπιστώθηκε προηγουμένως.

Για το τελευταίο ερώτημα της άσκησης, θεωρείται πως έχουμε έναν αρκετά μεγάλο σύνολο δεδομένων για εκπαίδευση και είναι απαραίτητο να γίνεται ταξινόμηση με NN, το οποίο σημαίνει πως θα πρέπει κάθε φορά να υπολογίζονται οι αποστάσεις του αγνώστου δείγματος από όλα τα ταξινομημένα δείγματα και να ταξινομείται στη βάση του πλησιέστερου γείτονα, πράγμα που δίνει πολυπλοκότητα $O(N^2)$ για κάθε νέο στοιχείο.

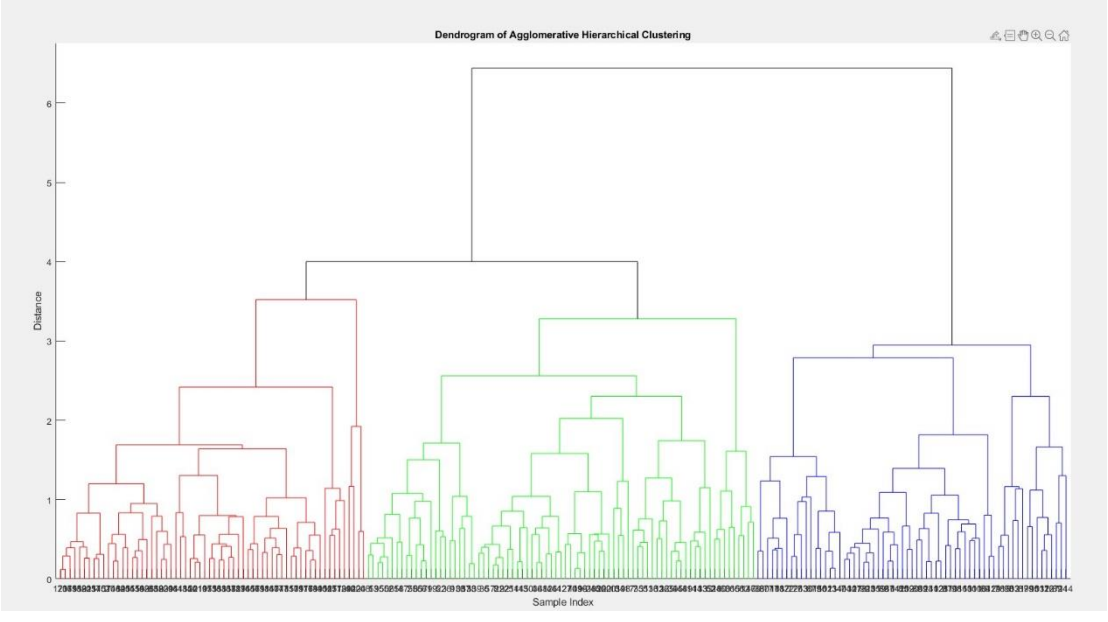
Ένας απλός τρόπος να γίνεται γρήγορα ταξινόμηση βάση του NN είναι να εκπαιδευτεί το σύστημα φτιάχνοντας k ομάδες με κάποια γνωστή τεχνική clustering π.χ k-means. Έπειτα για κάθε νέο δεδομένο αντί να υπολογίζεται η απόστασή του από κάθε δείγμα στον χώρο, μπορεί απλά να υπολογίζεται η απόστασή του από όλα τα κέντρα των ομάδων και να ταξινομείται βάση της μικρότερης απόστασης από κάθε κέντρο, ακριβώς όπως γίνεται και στη διαδικασία k-means clustering.

## ΑΣΚΗΣΗ 2

Σε αυτή την άσκηση γίνεται χρήση ιεραρχικών μεθόδων ομαδοποίησης. Μεγάλο πλεονέκτημα τέτοιων μεθόδων είναι πως ο αλγόριθμος για ένα δεδομένο πρόβλημα παράγει ένα δενδρόγραμμα που δείχνει πως ομαδοποιούνται τα δεδομένα από την αρχή μέχρι το τέλος του αλγορίθμου δίνοντας ευελιξία και εποπτική δυνατότητα για το πόσες ομάδες είναι βέλτιστο να υπάρχουν, χωρίς να υπάρχει ανάγκη για προκαθορισμό των επιθυμητών ομάδων. Αυτό γίνεται, διότι κατά τυχαίο τρόπο μπορεί ο δημιουργός του συστήματος να διατηρήσει διαφορετικά layers (δηλαδή συστήματα με διαφορετικό αριθμό clusters) και να μελετήσει τη συμπεριφορά τους άμεσα. Έτσι ακόμα και για πλήρη άγνοια του πόσες περίπου πρέπει να είναι οι ομάδες του συστήματος αυτές οι μέθοδοι παρέχουν άμεσο τρόπο δοκιμής και επαλήθευσης διαφορετικών τοπολογιών.

Σε αυτή την εφαρμογή χρησιμοποιείται μέθοδος συνάθροισης (agglomerative), στην οποία κάθε δείγμα είναι και μια ομάδα και σταδιακά με κάποιο κριτήριο απόστασης μεταξύ ομάδων γίνεται η ένωσή τους κατά ζεύγη, ώσπου να υπάρχει στο τέλος μόνο μια ομάδα.

Κριτήριο ομοιότητας μέγιστης, ελάχιστης, μέσης απόστασης των σημείων 2 ομάδων ή η απόσταση μεταξύ κέντρων είναι μερικοί τρόποι συνένωσης κλάσεων. Ωστόσο αυτό μπορεί να διαφοροποιηθεί και αν είναι γνωστή κάποια πληροφορία – ιδιαιτερότητα των δεδομένων που να δίνει πιο ταιριαστό heuristic. Πειραματικά φάνηκε πιο αποδοτικό το average linkage method, στο οποίο λαμβάνονται όλες οι αποστάσεις όλων των σημείων μιας ομάδας από τα σημεία της άλλης ομάδας, και λαμβάνεται ο μέσος όρος αυτών. Σαν μετρική απόστασης χρησιμοποιήθηκε η ευκλείδεια απόσταση.




```
Rand Index (Hierarchical Clustering): 0.8865
Rand Index (K-means Clustering): 0.8714
```

Βάση αποτελεσμάτων η ιεραρχική μέθοδος συνένωσης επιτυγχάνει οριακά καλύτερη ομαδοποίηση των δεδομένων από τον K-Means.

## ΑΣΚΗΣΗ 3

Σε αυτό το πρόβλημα γίνεται μείωση των διαστάσεων του προβλήματος με τη χρήση PCA (Principal Component Analysis) και LDA (Linear Discriminant Analysis). Η μέθοδος PCA μέσω του πίνακα συνδιακύμανσης βρίσκει τα ιδιοδιανύσματά του, που εκφράζουν τους άξονες πάνω στους οποίους παρουσιάζεται η μεγαλύτερη διασπορά των χαρακτηριστικών, και είναι στην ουσία γραμμικός συνδυασμός των χαρακτηριστικών. Οι άξονες-ιδιοδιανύσματα κατατάσσονται σε φθίνουσα σειρά, με βάση την ιδιοτιμή που τα συνοδεύει και δείχνει τον βαθμό διακύμανσης που παρουσιάζουν τα χαρακτηριστικά σε αυτόν τον άξονα. Κατά αυτόν τον τρόπο δίνεται η δυνατότητα να μετασχηματιστεί το πρόβλημα στο νέο ορθοκανονικό σύστημα και στη συνέχεια να απαλειφθούν άξονες όπου τα δεδομένα παρουσιάζουν μικρή διακύμανση. Έτσι δεν χάνεται σημαντική πληροφορία του προβλήματος και διεργασίες που επιτελούνται μετέπειτα γίνονται σαφώς πιο γρήγορες ή/και αποδοτικές (π.χ. επεξεργασία σήματος, συμπίεση εικόνας).

Ο LDA λύνει ένα αντίστοιχο πρόβλημα ιδιοτιμών το οποίο επιστρέφει νέους άξονες συστήματος συντεταγμένων, οι οποίοι είναι γραμμικοί συνδυασμοί των αρχικών χαρακτηριστικών, με σκοπό τη μέγιστη διαχωρισιμότητα κλάσεων.

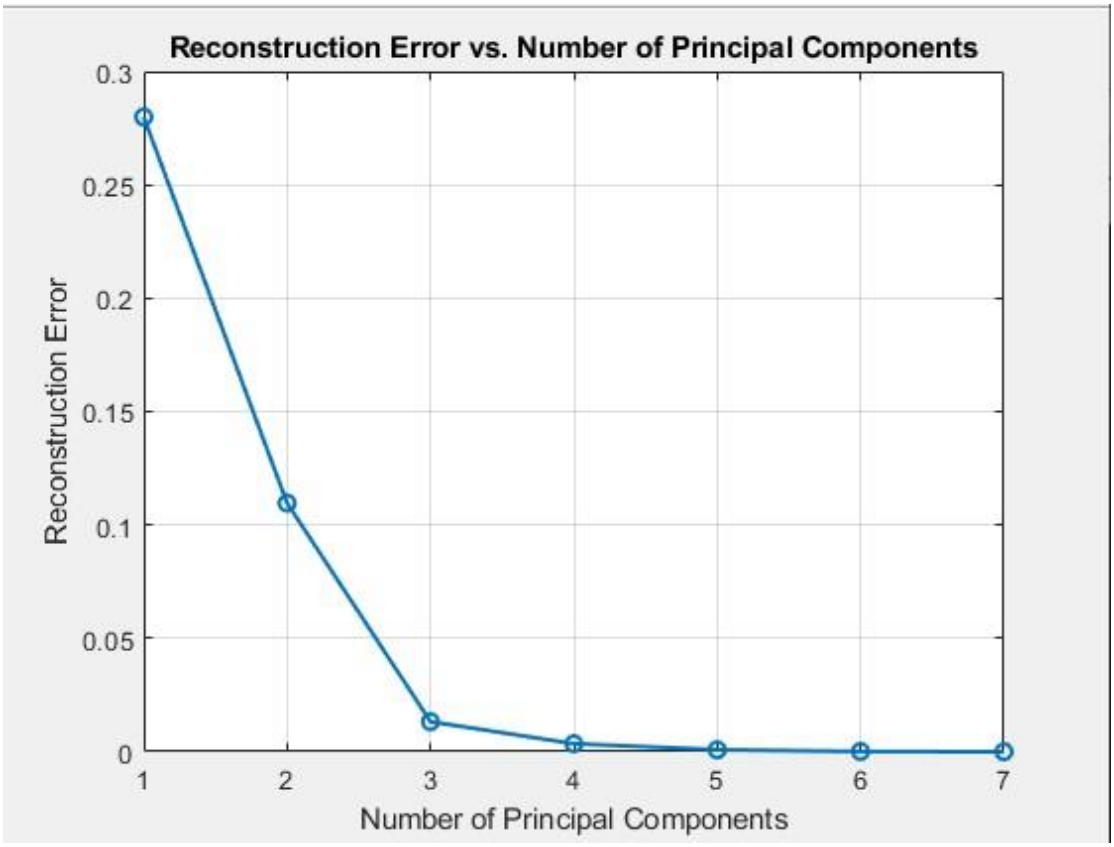
Όπως αναφέρθηκε, τα ιδιοδιανύσματα αναδιατάσσονται με βάση την ιδιοτιμή τους, σε φθίνουσα σειρά. Διατηρούμε τα πρώτα k Principal Components, έτσι ώστε η συνολική διακύμανση να είναι ένα επιθυμητό ποσοστό της αρχικής. Για τα ζητούμενα ποσοστά, λαμβάνουμε τα πρώτα 2 και 4 Principal Components αντίστοιχα.

```
Number of components for 80% variance: 2
Number of components for 99.5% variance: 4
```

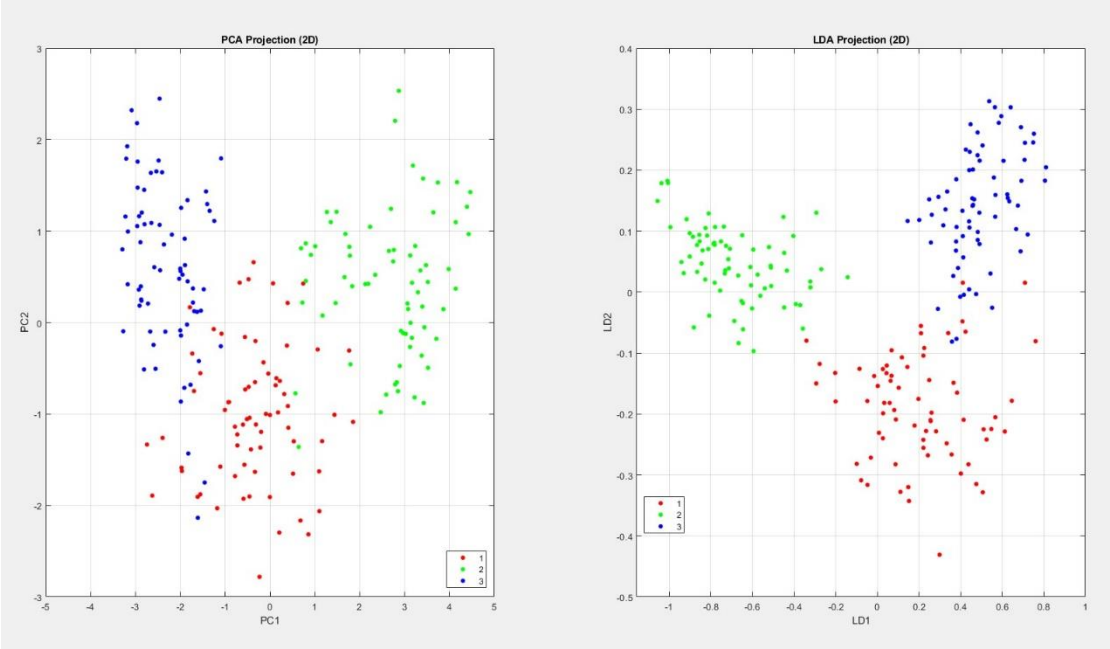
Πραγματοποιώντας ανακατασκευή του αρχικού πίνακα δεδομένων γνωρίζοντας πως

τα scores είναι ο πίνακας Z = U*S που είναι τα προβαλλόμενα δεδομένα στο νέο σύστημα, από την SVD πως $X = U\ S\ V^T$ και στην περίπτωση που τα δεδομένα έχουν κανονικοποιηθεί

ισχύει W = V, όπου W περιέχει σε στήλες του τα principal components (δηλαδή ο πίνακας principal component coefficients) θα έχουμε X’ = Z $W^T$. Το σφάλμα ανακατασκευής είναι



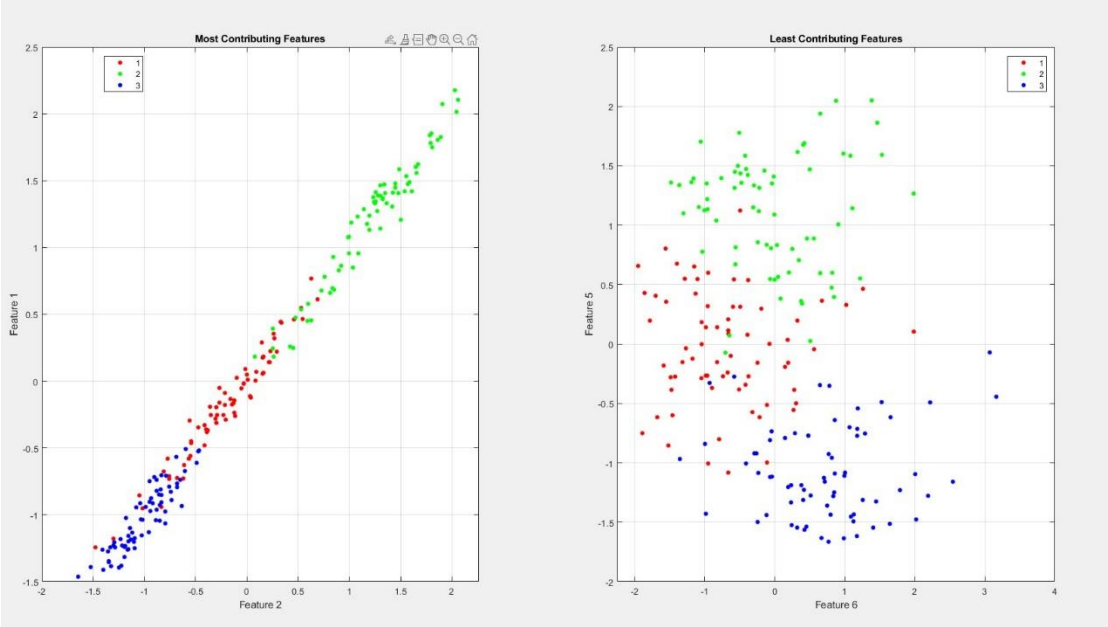

Παίρνοντας τις πρώτες 2 στήλες του score από την PCA έχουμε την δισδιάστατη προβολή στους 2 κύριους άξονες (πρώτα 2 Principal Components). Για την LDA, επιλέγονται σαν axis οι άξονες που επιτυγχάνουν καλύτερο διαχωρισμό στις κλάσεις. Για να γίνει αυτό, πρέπει να υπολογιστούν τα inter-cluster και intra-cluster scatter matrices για να καθοριστεί η διαχωριστική ικανότητα κάθε discriminant.



Από το σχήμα επαληθεύεται πως ο PCA στοχεύει στη μέγιστη διασπορά, ενώ ο LDA στη μέγιστη διαχωρισιμότητα.

Για να βρεθούν τα χαρακτηριστικά με τη μεγαλύτερη συνεισφορά πρέπει να λάβουμε πάλι τους συντελεστές - ιδιοτιμές για κάθε linear discriminant που αντιστοιχούν στη συνεισφορά διαχωρισιμότητας των χαρακτηριστικών και με μια μετρική να λάβουμε το μέτρο των διανυσμάτων που προκύπτουν από τις 3 ιδιοτιμές ανά χαρακτηριστικό (μία για τον διαχωρισμό 1-2, 2-3 και 1-3 αντίστοιχα). Βρίσκουμε κατά αυτόν τον τρόπο

```
Most contributing features: 2, 1
Least contributing features: 6, 5
```



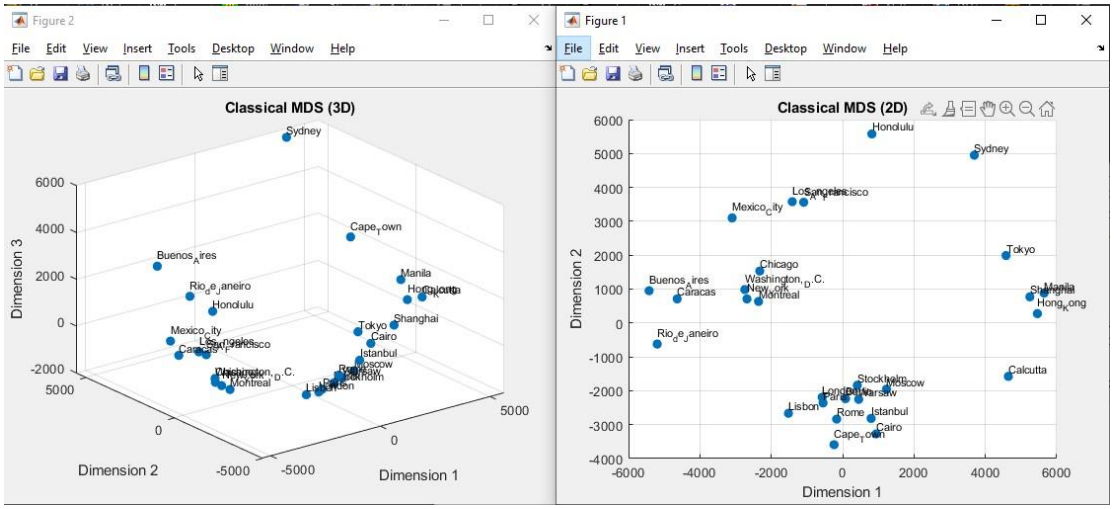

Παρατηρούμε από το αριστερό γράφημα πως υπάρχει γραμμική σχέση μεταξύ των πιο σημαντικών χαρακτηριστικών, και το πώς αυτά κατατάσσουν τα δεδομένα στις διακριτές ομάδες καθιστώντας το πρόβλημα εύκολα διαχωρίσιμο με χρήση λιγότερων διαστάσεων.

Στη δεξιά εικόνα συμβαίνει ακριβώς το αντίθετο, δηλαδή δεν υπάρχει ξεκάθαρη συσχέτιση μεταξύ των χαρακτηριστικών και των ομάδων του προβλήματος, οπότε και δεν συνεισφέρουν σημαντική πληροφορία στο πρόβλημα διαχωρισιμότητας με αποτέλεσμα να λειτουργούν περισσότερο σαν θόρυβος.

## ΑΣΚΗΣΗ 4

Η μέθοδος MDS είναι μια ακόμα μέθοδος ελάττωσης διαστάσεων ενός προβλήματος, αξιοποιώντας τον πίνακα αποστάσεων των δεδομένων και μειώνοντας τις διαστάσεις αυτού χωρίς να χαθεί σημαντική πληροφορία. Αυτό το επιτυγχάνει μετασχηματίζοντας τις θέσεις των σημείων έτσι, ώστε οι σχετικές αποστάσεις να διατηρούνται και η αναλογία στις ομοιότητες/διαφορές μεταξύ των δεδομένων να παραμένουν αναλλοίωτες.

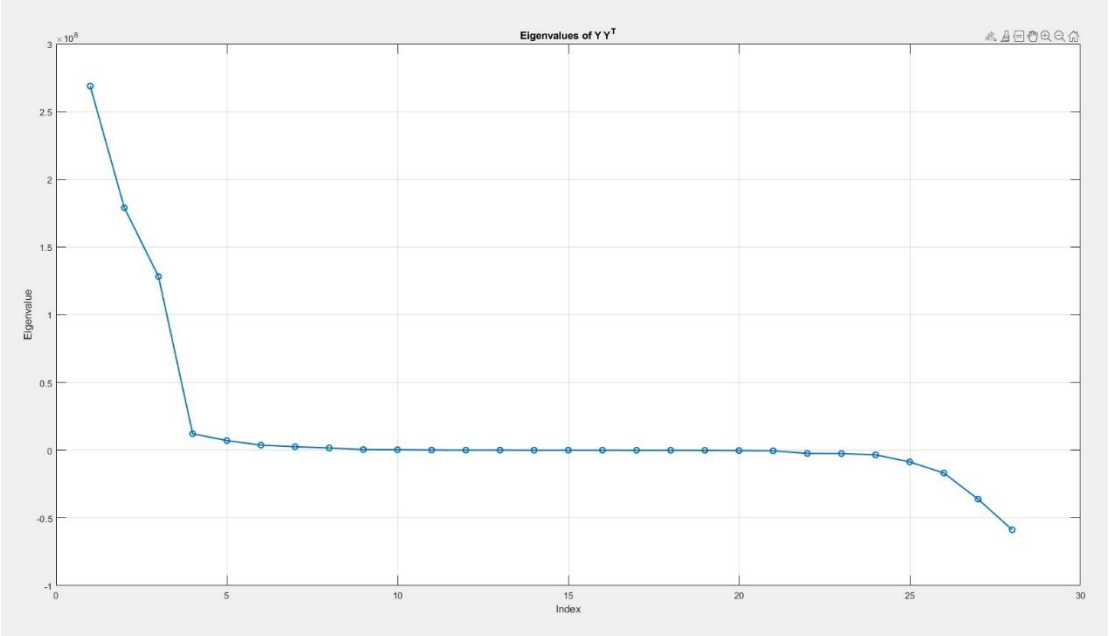
Παρακάτω φαίνεται η αναπαράσταση του World Distance Matrix




Τα αποτελέσματα σαφώς δεν μπορούν να είναι συναφή με τις γεωγραφικές τοποθεσίες των πόλεων που απεικονίζονται, καθώς αυτό εξαρτάται από το τι είδους χαρακτηριστικά επεξεργαζόμαστε, αλλά αυτό που αναμένεται και είναι σαφές και στην εικόνα είναι πως όλες οι πόλεις των διαφορετικών ηπείρων φαίνεται να ομαδοποιούνται αναλόγως. Φαίνεται πόλεις τις Ασίας όπως Τόκιο, Χονγκ Κονγκ, Σανγκάη να είναι μια ομάδα. Αντίστοιχα πόλεις της Ευρώπης και της Βόρειας και Νότιας Αμερικής να σχηματίζουν δικές τους ομάδες, ενώ πόλεις από πιο απομακρυσμένες χώρες όπως η Αυστραλία να είναι outlier.

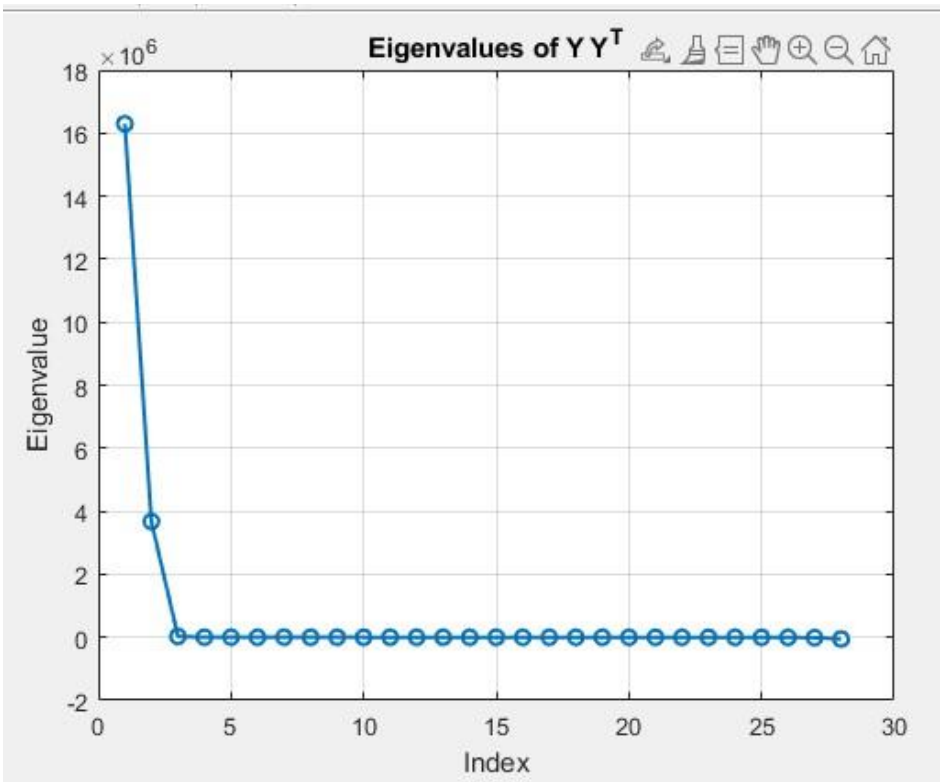
Ήδη ποιοτικά, φαίνεται πως οι 2 και 3 διαστάσεις κάνουν επαρκή δουλειά σε θέμα διαχωρισιμότητας και διατήρησης της σημαντικής πληροφορίες για τις σχετικές θέσεις των διαφορετικών τοποθεσιών. Στις 3 διαστάσεις ειδικά, φαίνεται ορισμένες ασυνέπειες της δισδιάστατης αναπαράστασης να διορθώνονται. Οπότε αναμένεται πως δεν θα είναι απαραίτητες παραπάνω από 3 διαστάσεις για την επαρκή εποπτεία του προβλήματος.

Βρίσκοντας τις ιδιοτιμές λύνοντας για όλα τα χαρακτηριστικά έχουμε




Όπου πράγματι οι 3 πρώτες διαστάσεις-components-ιδιοδιανύσματα φαίνεται να είναι και οι πιο σημαντικές για την διατήρηση της αρχικής πληροφορίας.

Πράγματι για το dataset των αμερικάνικων πόλεων πάνω από 3 διαστάσεις παρατηρούμε μηδενικές ιδιοτιμές, δηλαδή μηδενική συνεισφορά στη πληροφορία των δεδομένων.



Ο λόγος έγκειται στον εξής λόγο. Προφανώς οι πρώτες 3 διαστάσεις-χαρακτηριστικά πιθανόν να εκφράζουν σε καρτεσιανές συντεταγμένες τις θέσεις των πόλεων οπότε αυτές οι πληροφορίες είναι και οι πιο σημαντικές. Τα υπόλοιπα χαρακτηριστικά μπορεί να αφορούν δημογραφικά όπως πολιτικές πεποιθήσεις, συνήθειες και άλλα χαρακτηριστικά συνδεδεμένα με τον πολιτισμό ενός λαού. Όταν λοιπόν μελετώνται πόλεις μιας πιο στενευμένης γεωγραφικής τοποθεσίας είναι πιθανό αυτά τα χαρακτηριστικά να μη παρουσιάζουν καμία σημαντική διακύμανση ώστε να μπορέσουν να δώσουν και κάποια δυνατότητα διάκρισης δειγμάτων με βάση αυτά τα χαρακτηριστικά.