Ινστιτούτο Σχεδιασμού και Ανάλυσης Πειραμάτων του Γ.Π.Α.

3ος Κύκλος Διαλέξεων-Απρίλιος 2024

Πολυμεταβλητή Ανάλυση στην Πράξη

Μ. Κούτρας
Καθηγητής Τμήματος Στατιστικής & Ασφαλιστικής Επιστήμης
Δντης ΠΜΣ στην Εφαρμοσμένη Στατιστική
Πανεπιστήμιο Πειραιώς

Πολυμεταβλητή Ανάλυση:

Γιατί και πότε

- **Ένα** χαρακτηριστικό ή πολλά χαρακτηριστικά, αλλά τα μελετάμε ανεξάρτητα το ένα από το άλλο;

  Κλασικές τεχνικές της περιγραφικής στατιστικής (ιστογράμματα, φυλλογραφήματα, θηκογράμματα, αριθμητικά περιγραφικά μέτρα) και της στατιστικής συμπερασματολογίας (έλεγχοι υποθέσεων, εκτιμητική)

- **Περισσότερα από ένα** χαρακτηριστικά συγχρόνως;

  Τεχνικές της πολυμεταβλητής στατιστικής ανάλυσης (multivariate Statistical Analysis)

Τι θα δούμε μαζί

- Γραφικές τεχνικές για πολυδιάστατα δεδομένα
- Ανάλυση κατά συστάδες ή ομάδες (Cluster Analysis-CA)
- Ανάλυση κυρίων συνιστωσών (Principal Component Analysis-PCA)
- Διαχωριστική ανάλυση (Disciminant Analysis)

Μία μελέτη περίπτωσης (case study)

Fisher Iris Data

N=150 παρατηρήσεις

Ποικιλίες

a: Iris Setosa
b: Iris Versicolor
c: Iris Virginica

Μεταβλητές

SL: Μήκος Σεπάλου
SW: Πλάτος Σεπάλου
PL: Μήκος Πετάλου
PW: Πλάτος Πετάλου

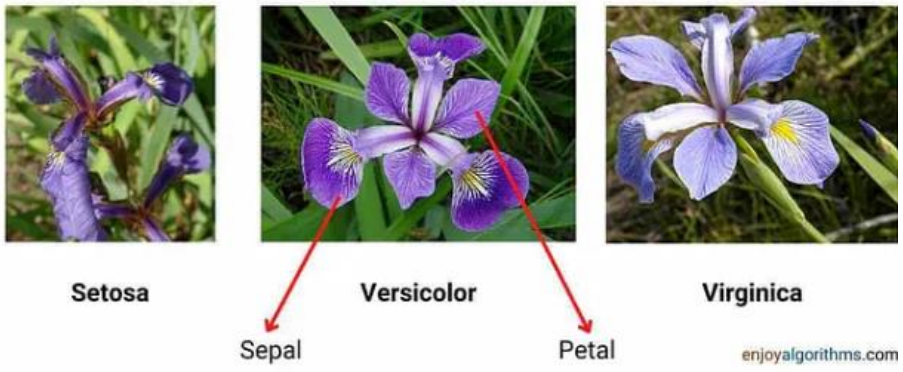
Fisher Iris Data: Δείγμα 30 παρατηρήσεων

Μεταβλητές

SL : Μήκος Σεπάλου
SW: Πλάτος Σεπάλου
PL : Μήκος Πετάλου
PW:Πλάτος Πετάλου

Ποικιλίες

a: Iris Setosa
b: Iris Versicolor
c: Iris Virginica



α/α	Ποικιλία	SL	SW	PL	PW
1	a	5.1	3.5	1.4	.2
2	a	4.9	3.0	1.4	.2
3	a	4.7	3.2	1.3	.2
4	a	4.6	3.1	1.5	.2
5	a	5.0	3.6	1.4	.2
6	a	5.4	3.9	1.7	.4
7	a	4.6	3.4	1.4	.3
8	a	5.0	3.4	1.5	.2
9	a	4.4	2.9	1.4	.2
10	a	4.9	3.1	1.5	.1
11	b	7.0	3.2	4.7	1.4
12	b	6.4	3.2	4.5	1.5
13	b	6.9	3.1	4.9	1.5
14	b	5.5	2.3	4.0	1.3
15	b	6.5	2.8	4.6	1.5
16	b	6.3	3.3	4.7	1.6
17	b	4.9	2.4	3.3	1.0
18	b	6.6	2.9	4.6	1.3
19	b	5.2	2.7	3.9	1.4
20	b	5.7	2.8	4.1	1.3
21	c	6.3	3.3	6.0	2.5
22	c	5.8	2.7	5.1	1.9
23	c	7.1	3.0	5.9	2.1
24	c	6.3	2.9	5.6	1.8
25	c	6.5	3.0	5.8	2.2
26	c	7.6	3.0	6.6	2.1
27	c	4.9	2.5	4.5	1.7
28	c	7.3	2.9	6.3	1.8
29	c	6.7	2.5	5.8	1.8
30	c	7.2	3.6	6.1	1.5

Πίνακες δεδομένων

Έχουμε ένα δείγμα n ατόμων/αντικειμένων από ένα πληθυσμό, και σε κάθε άτομο παρατηρούμε $p \geq 2$ χαρακτηριστικά (τυχαίες μεταβλητές). Οι $n \times p$ παρατηρήσεις μπορούν να συγκεντρωθούν σε ένα πίνακα $X = (x_{ij})$ με n γραμμές και p στήλες

p μεταβλητές

n άτομα

$$\begin{bmatrix} x_{11} & .. & x_{1j} & .. & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & .. & x_{ij} & .. & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & .. & x_{nj} & .. & x_{np} \end{bmatrix}$$

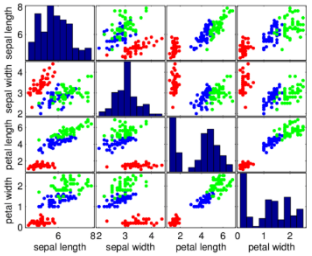
Γραφικές Μέθοδοι

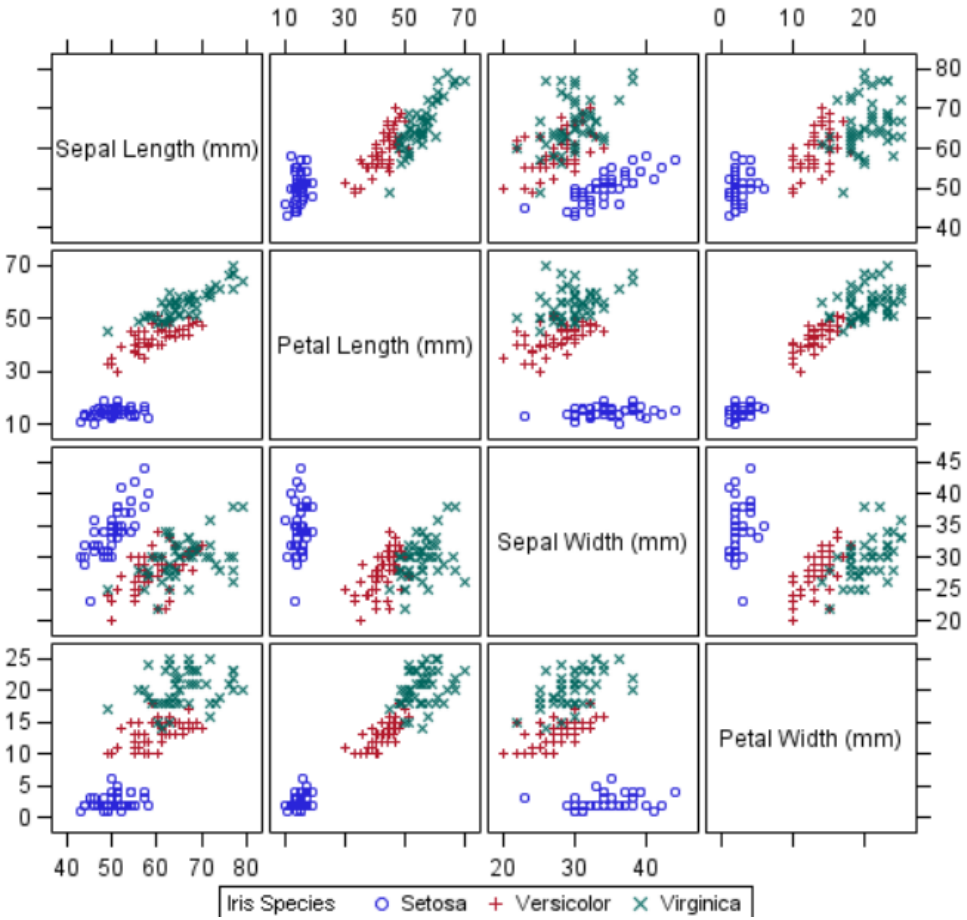
Εμπειρικές Μέθοδοι ομαδοποίησης

- 2D και 3D plots
- Γραφική αναπαράσταση πολυδιάστατων παρατηρήσεων σε δύο διαστάσεις: icon plots, star plots, sun ray plots, Chernoff faces κλπ.
- Οι καμπύλες Andrews μπορούν επίσης να μας δώσουν ένα εναλλακτικό δισδιάστατο γραφικό τρόπο ομαδοποίησης.

2D Plots

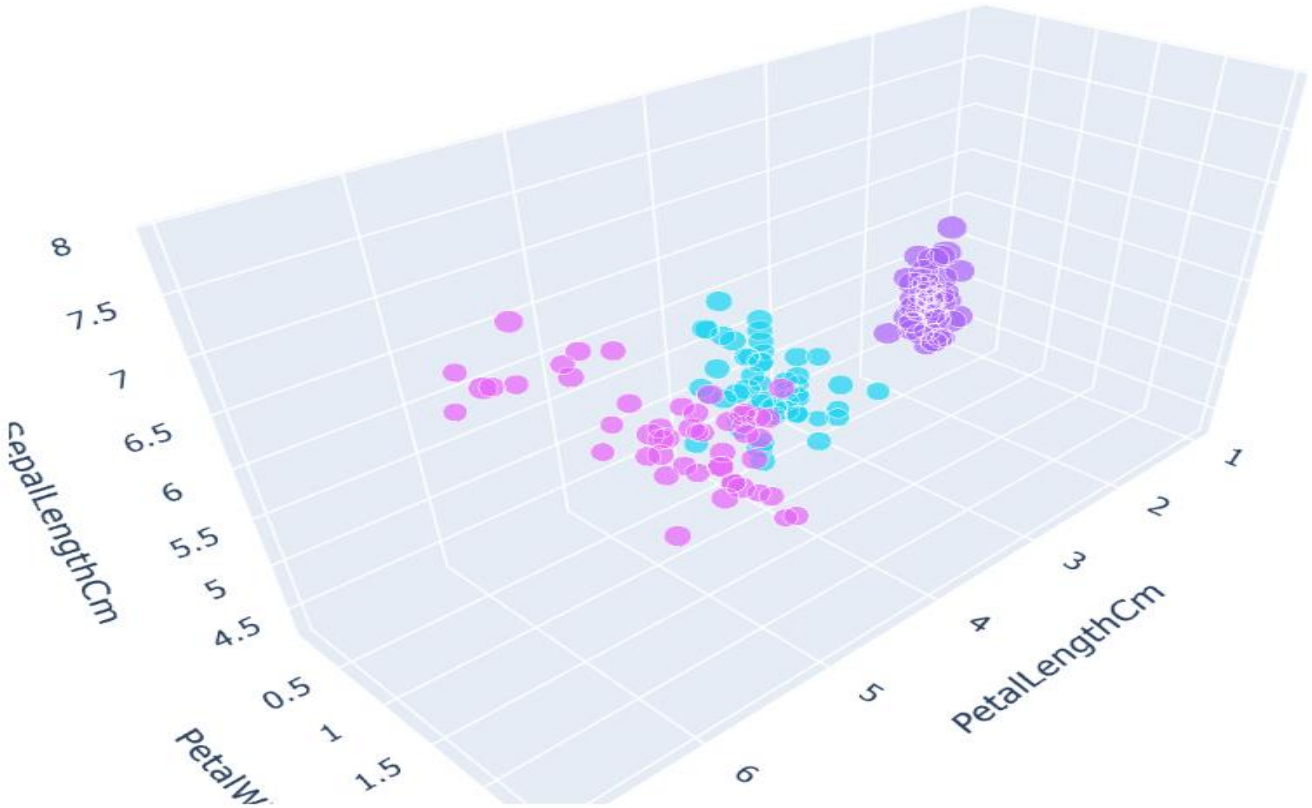
Δύο μεταβλητές μόνο!



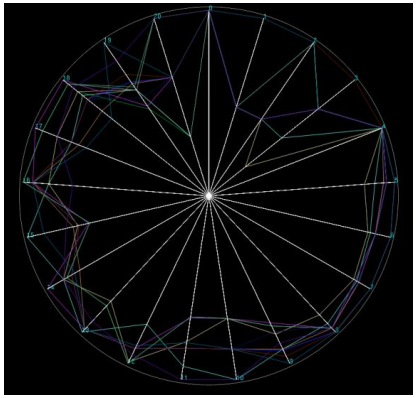


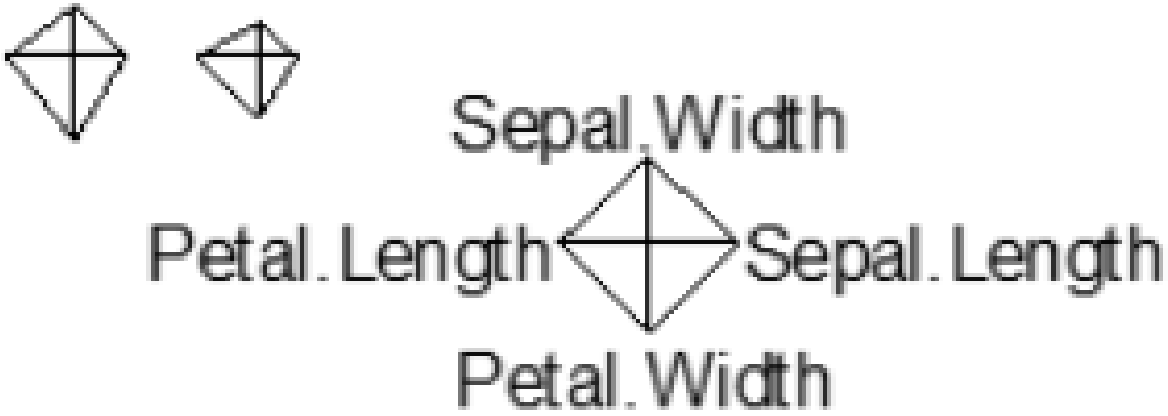
3D Plots

Τρεις μεταβλητές μόνο!

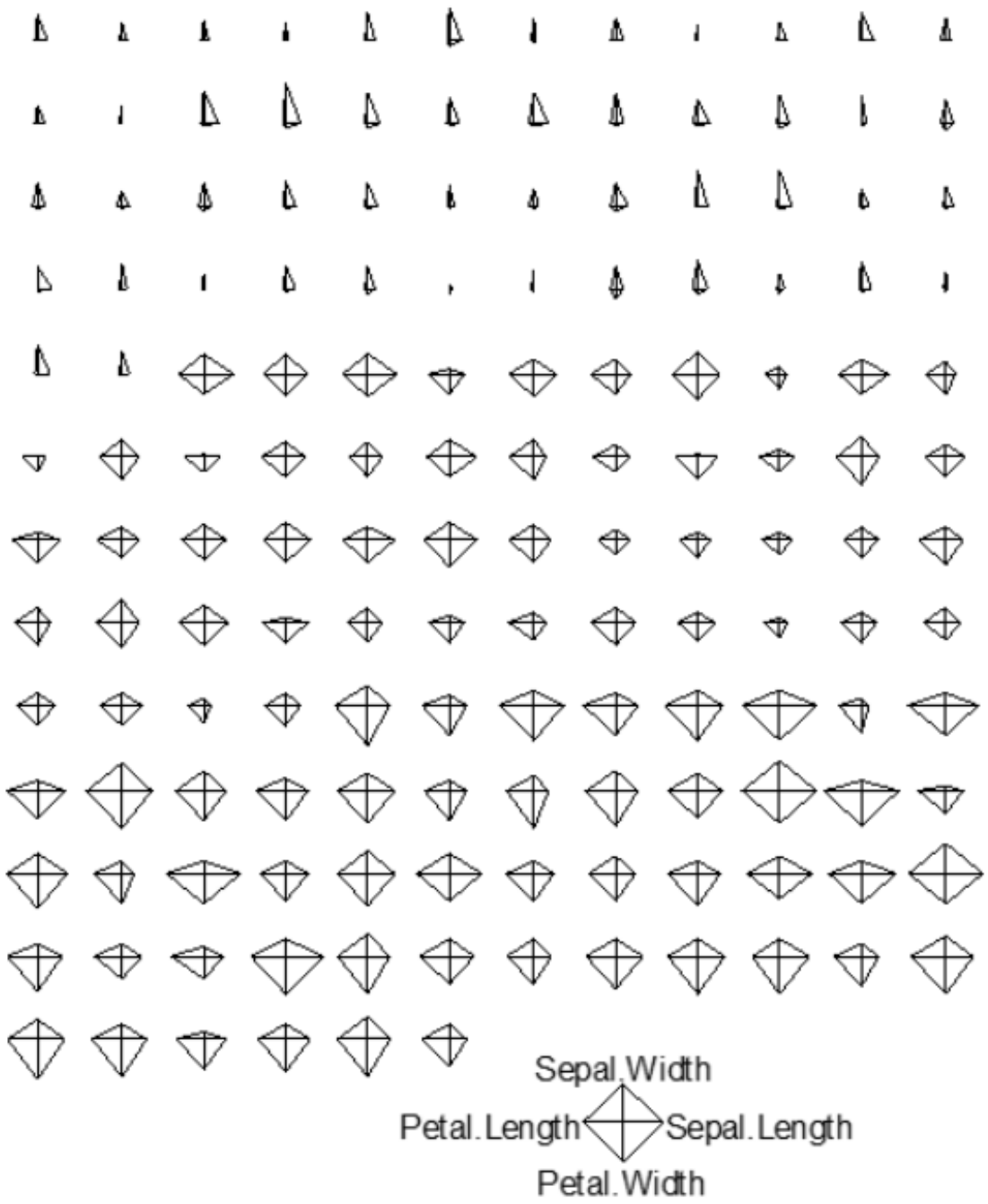


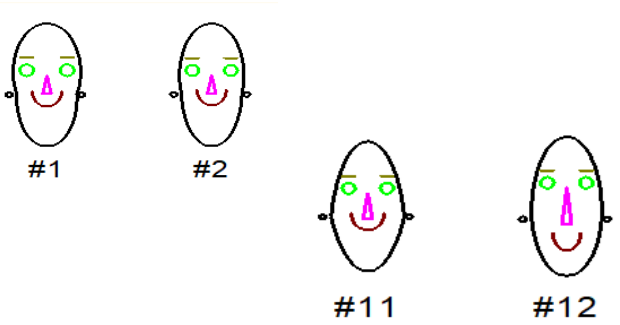
Star Plots





Star Plots

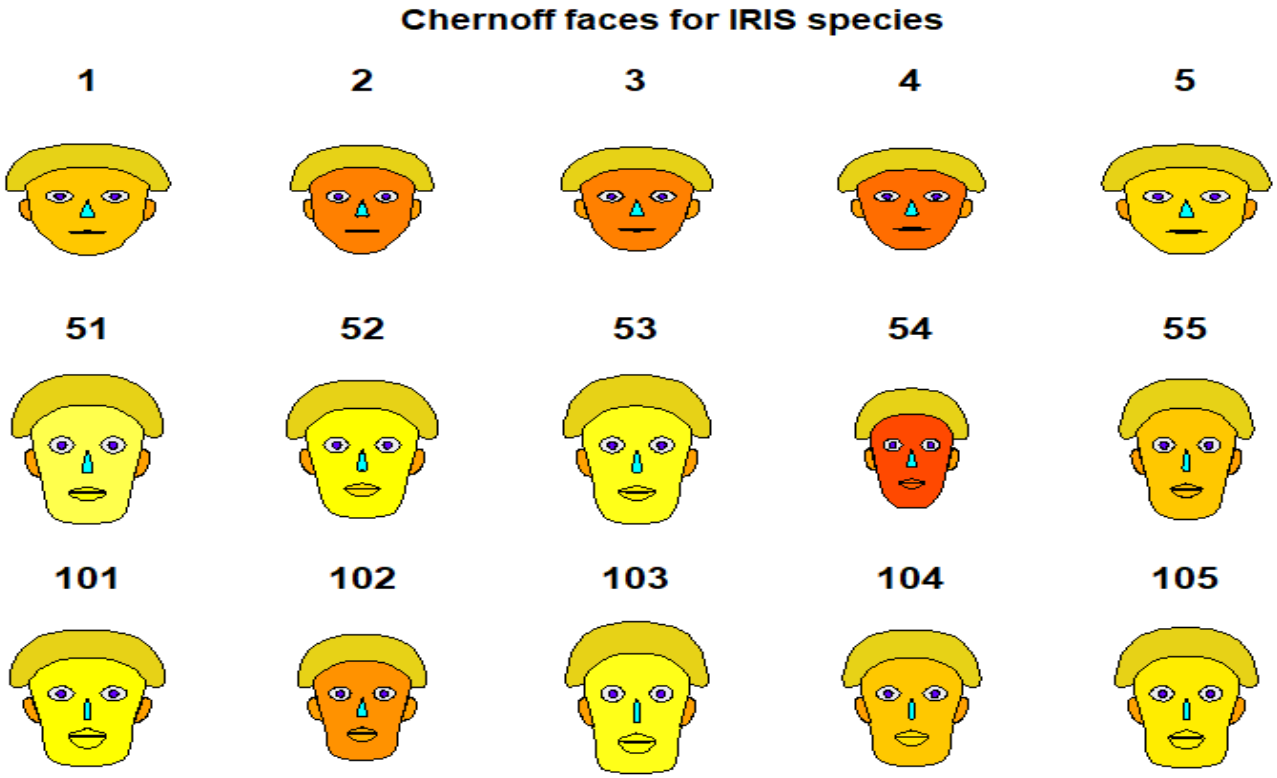




Chernoff faces

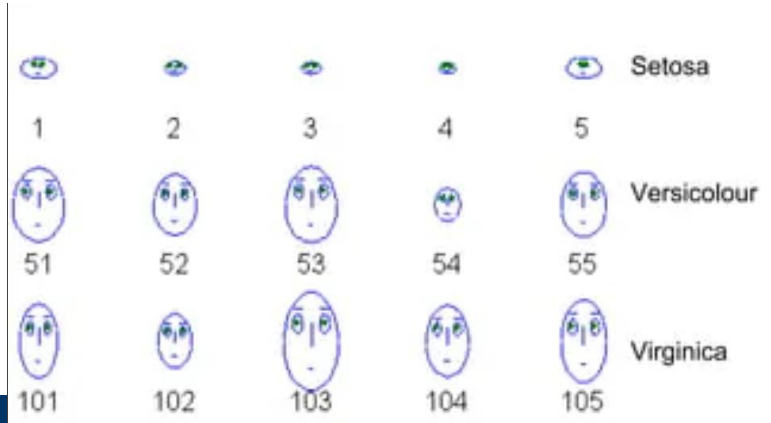
- Εισήχθηκε από τον H. Chernoff το 1973
- Σε κάθε χαρακτηριστικό του προσώπου αντιστοιχίζεται μια μεταβλητή
- Η τεχνική στηρίζεται στη ευαισθησία του ανθρώπινου ματιού να εντοπίζει ακόμα και τις μικρότερες διαφορές στα πρόσωπα
- Εντοπίζει συσχετίσεις μεταξύ των μεταβλητών αλλά και ακραίες παρατηρήσεις

Chernoff faces: Iris data



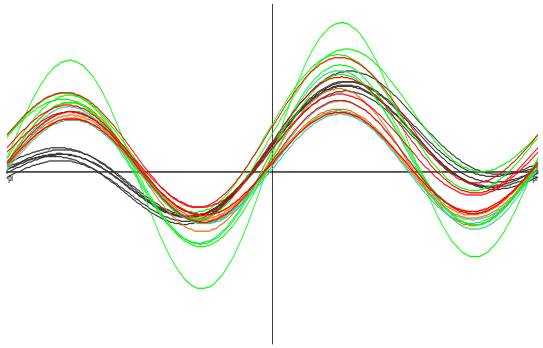
Chernoff faces: Iris data

Chernoff faces



- Εφαρμόζεται σε περιορισμένο πλήθος μεταβλητών
- Η τελική μορφή κάθε προσώπου εξαρτάται από την αντιστοίχηση των μεταβλητών στα χαρακτηριστικά του
- Κάποια χαρακτηριστικά έχουν μεγαλύτερη αναγνωρισιμότητα από άλλα

Καμπύλες του Andrews



- Παρουσιάστηκαν από τον Andrews το 1972
- Κάθε πολυμεταβλητή παρατήρηση απεικονίζεται στο επίπεδο με τη μορφή καμπύλης μέσω μιας συνάρτησης

Παράδειγμα

$$X = \begin{bmatrix} 12 & 13 & 14 & 17 & 16 & 18 \\ 10 & 14 & 12 & 18 & 14 & 16 \\ 20 & 19 & 18 & 20 & 16 & 18 \\ 13 & 12 & 11 & 11 & 16 & 18 \\ 15 & 7 & 5 & 14 & 3 & 10 \end{bmatrix}$$

Βαθμολογίες 5 μαθητών (n =5) σε 6 διαγωνίσματα (p =6)

$$f_1(t) = 12/\sqrt{2} + 13\sin t + 14\cos t + 17\sin 2t + 16\cos 2t + 18\sin 3t$$

$$f_2(t) = 10/\sqrt{2} + 14\sin t + 12\cos t + 18\sin 2t + 14\cos 2t + 16\sin 3t$$

$$f_3(t) = 20/\sqrt{2} + 19\sin t + 18\cos t + 20\sin 2t + 16\cos 2t + 18\sin 3t$$

$$f_4(t) = 13/\sqrt{2} + 12\sin t + 11\cos t + 11\sin 2t + 16\cos 2t + 18\sin 3t$$

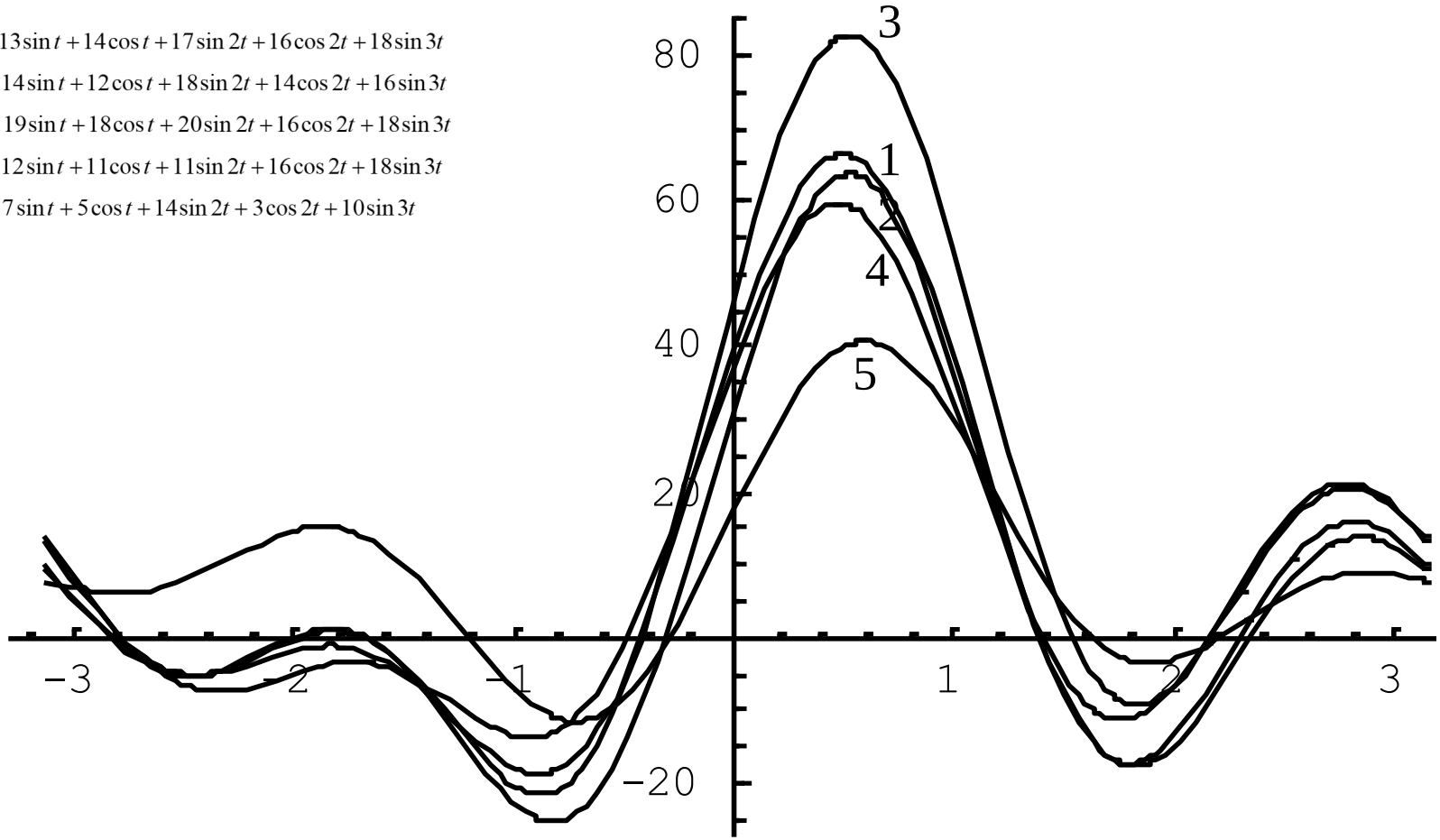
$$f_5(t) = 15/\sqrt{2} + 7\sin t + 5\cos t + 14\sin 2t + 3\cos 2t + 10\sin 3t$$

Παράδειγμα

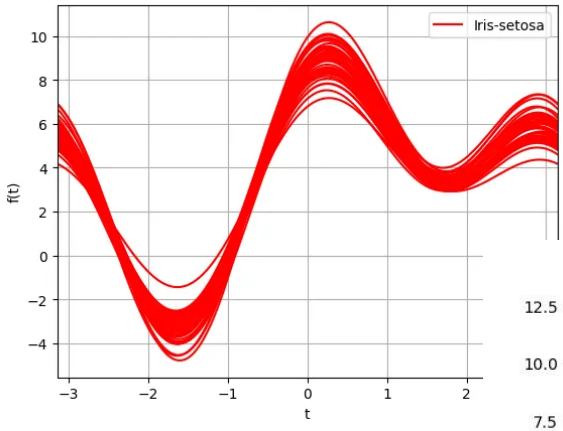
$$X = \begin{bmatrix} 12 & 13 & 14 & 17 & 16 & 18 \\ 10 & 14 & 12 & 18 & 14 & 16 \\ 20 & 19 & 18 & 20 & 16 & 18 \\ 13 & 12 & 11 & 11 & 16 & 18 \\ 15 & 7 & 5 & 14 & 3 & 10 \end{bmatrix}$$

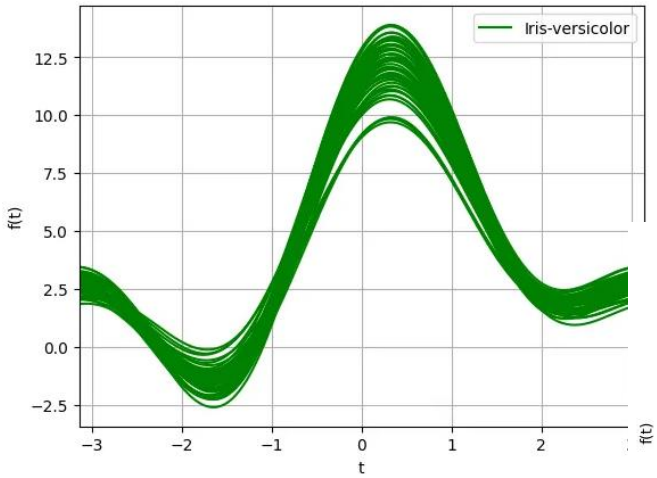
Βαθμολογίες 5 μαθητών (n =5) σε 6 διαγωνίσματα (p =6)

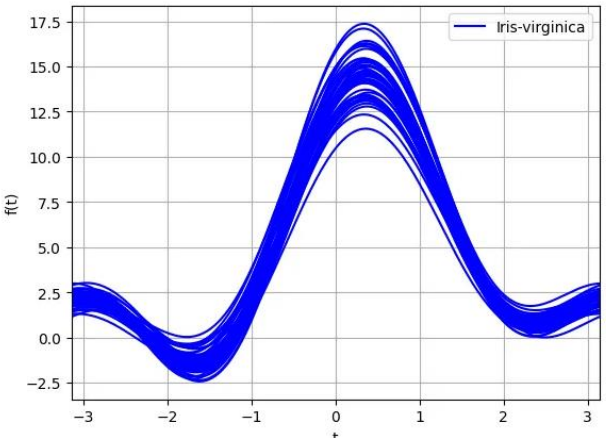
$$f_1(t) = 12/\sqrt{2} + 13\sin t + 14\cos t + 17\sin 2t + 16\cos 2t + 18\sin 3t$$
$$f_2(t) = 10/\sqrt{2} + 14\sin t + 12\cos t + 18\sin 2t + 14\cos 2t + 16\sin 3t$$
$$f_3(t) = 20/\sqrt{2} + 19\sin t + 18\cos t + 20\sin 2t + 16\cos 2t + 18\sin 3t$$
$$f_4(t) = 13/\sqrt{2} + 12\sin t + 11\cos t + 11\sin 2t + 16\cos 2t + 18\sin 3t$$
$$f_5(t) = 15/\sqrt{2} + 7\sin t + 5\cos t + 14\sin 2t + 3\cos 2t + 10\sin 3t$$



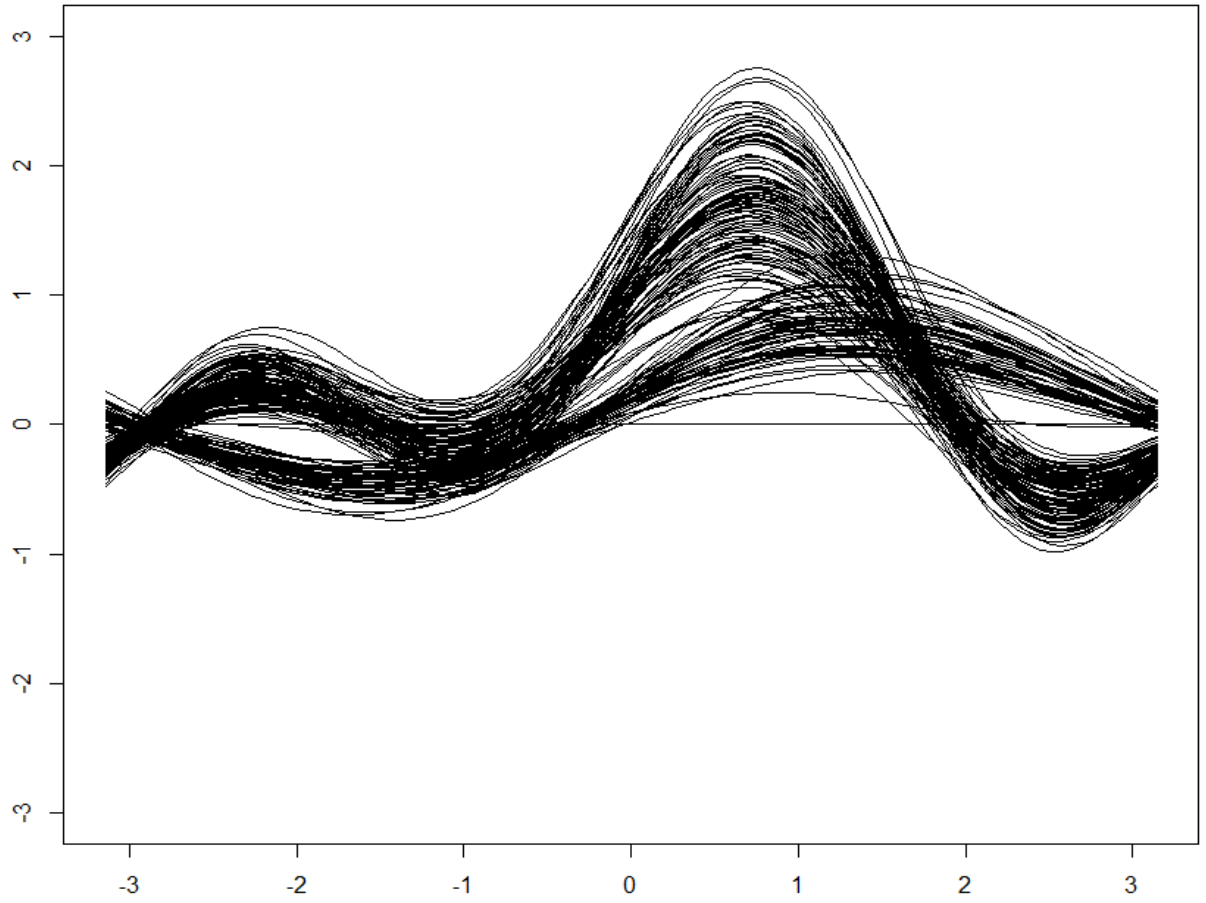
Καμπύλες του Andrews: Iris data



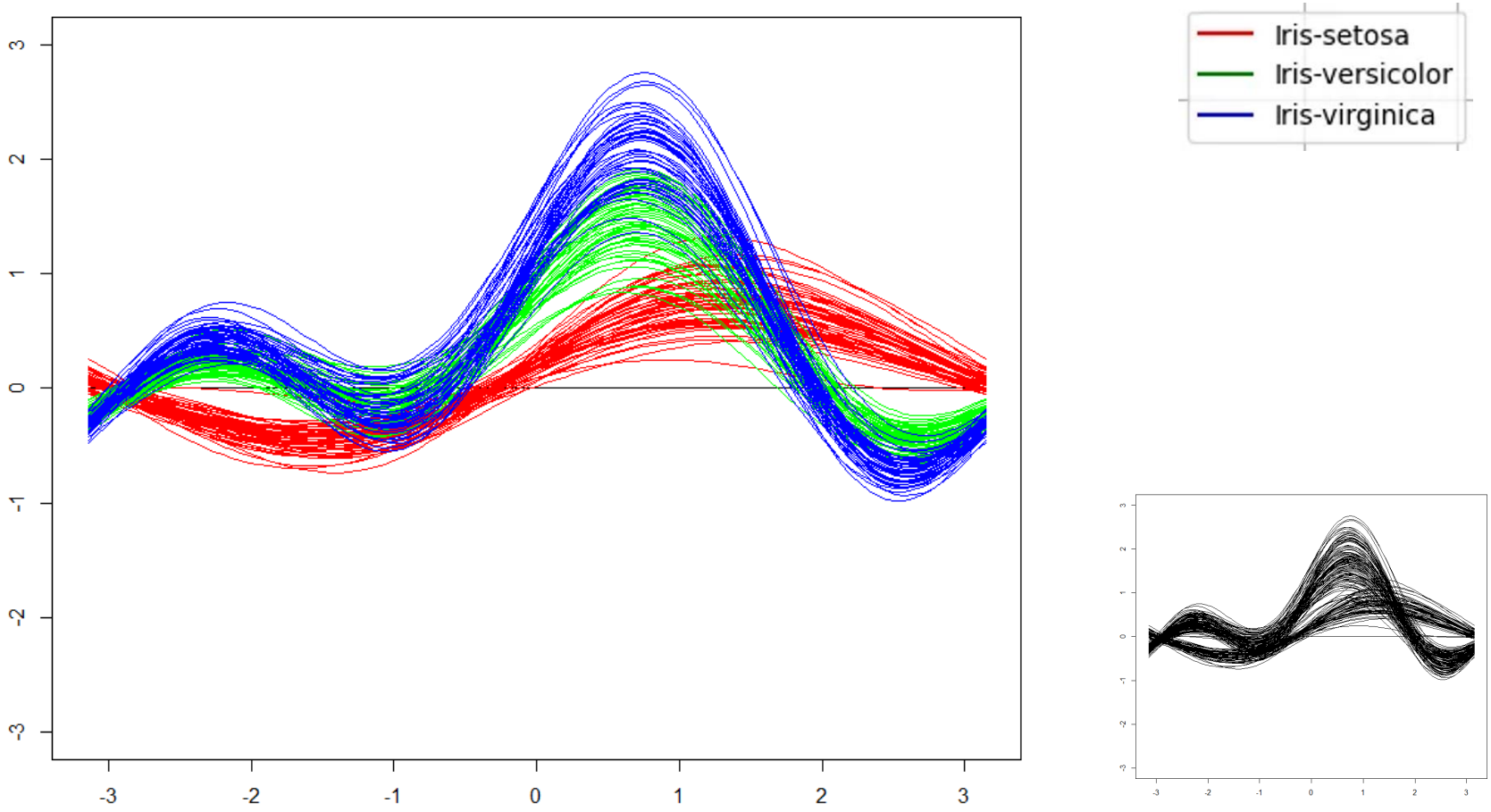




Καμπύλες του Andrews: Iris data



Καμπύλες του Andrews: Iris data



Ανάλυση κατά συστάδες

Εισαγωγικά

Ανάλυση κατά συστάδες (ομάδες) ή ομαδοποίηση δεδομένων (Cluster Analysis)

- Εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς έναν αριθμό μεταβλητών με σκοπό να δημιουργήσει συστάδες (ομάδες) από παρατηρήσεις που μοιάζουν μεταξύ τους

- Μια επιτυχημένη εφαρμογή των τεχνικών της θα πρέπει να καταλήξει σε ομάδες για τις οποίες
  - οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς
  - παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο.

Με αυτόν τον τρόπο επιτυγχάνουμε την ευκολότερη και αποδοτικότερη επεξεργασία των δεδομένων που διαθέτουμε.

α/α	Ποικιλία	SL	SW	PL	PW
1	a	5.1	3.5	1.4	.2
2	a	4.9	3.0	1.4	.2

Πίνακες δεδομένων & αποστάσεις

$\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$: το διάνυσμα των παρατηρήσεων (για τις p μεταβλητές) που αφορά το i άτομο ($i = 1,2,...,n$).

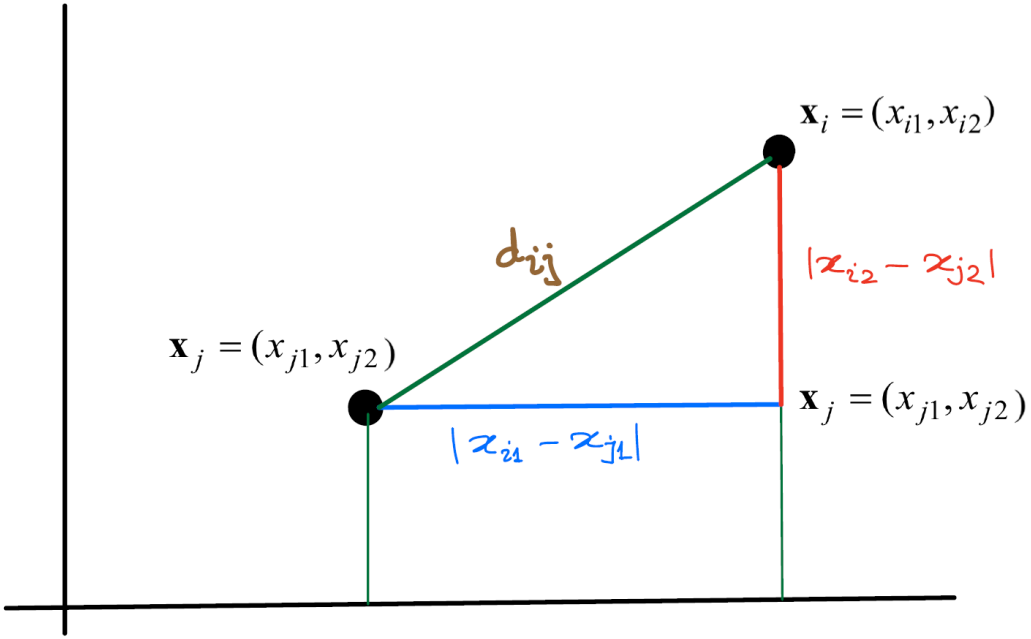
Το πιο γνωστό μέτρο απόστασης μεταξύ δύο παρατηρήσεων

$$\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip}) \text{ και } \mathbf{x}_j = (x_{j1}, x_{j2}, ..., x_{jp})$$

είναι η **ευκλείδεια απόσταση**, η οποία ορίζεται από τον τύπο

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2} \ .$$

Ευκλείδια απόσταση στο επίπεδο



$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$$

$p = 2$

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$$

Η απόσταση του Pearson

Αν συμβολίσουμε με s_r τη διακύμανση της r μεταβλητής

$$s_r = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_{ir} - \bar{x}_r)^2 \right]^{1/2}, \qquad \bar{x}_r = \frac{1}{n} \sum_{i=1}^{n} x_{ir}$$

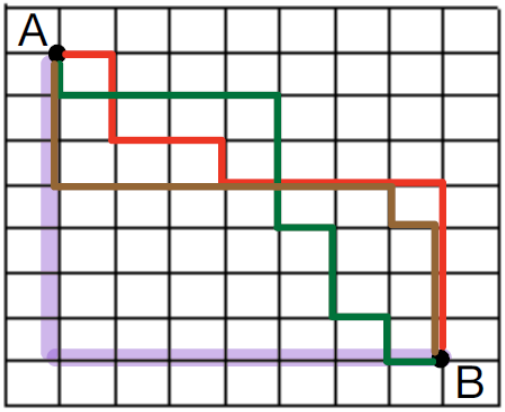
τότε η απόσταση του Pearson έχει τη μορφή

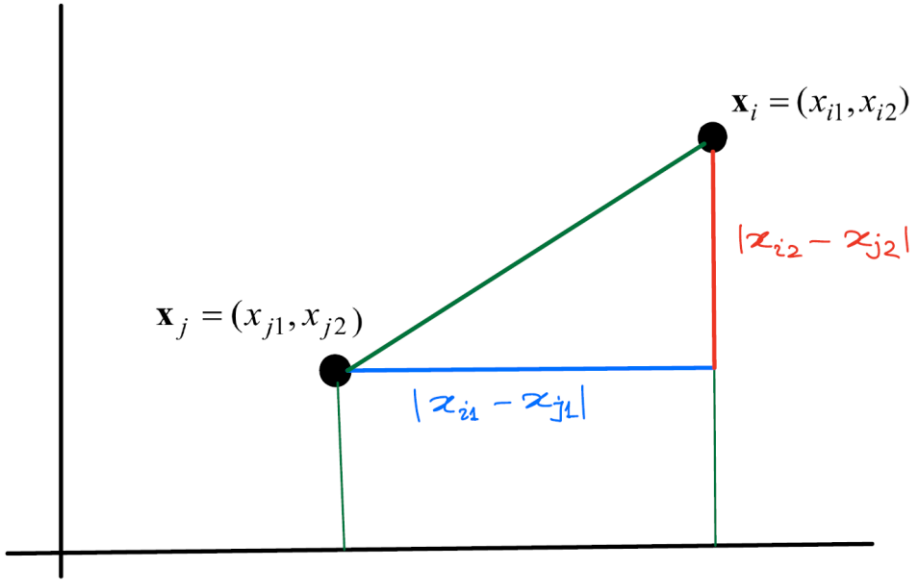
$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{r=1}^{p} \frac{(x_{ir} - x_{jr})^2}{s_r^2}} = \sqrt{\sum_{r=1}^{p} \left(\frac{x_{ir} - x_{jr}}{s_r} \right)^2}$$

Απόσταση Manhattan ή City-block metric

α/α	Ποικιλία	SL	SW	PL	PW
1	a	5.1	3.5	1.4	.2
2	a	4.9	3.0	1.4	.2

$$d_{ij} = \sum_{r=1}^{p} \left| x_{ir} - x_{jr} \right|$$



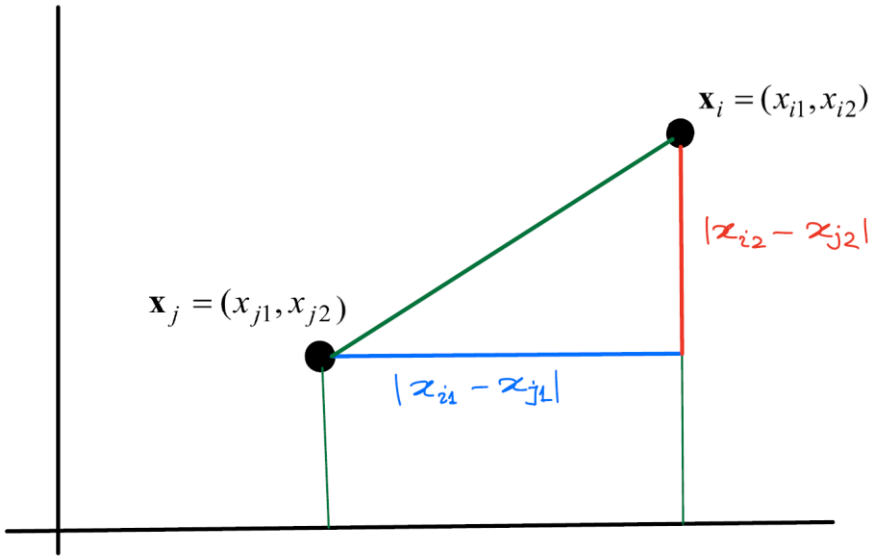


$$d_{ij} = \sum_{r=1}^{p} \left| x_{ir} - x_{jr} \right| \qquad p = 2$$

Απόσταση max ή απόσταση του Chebyshev

α/α	Ποικιλία	SL	SW	PL	PW
1	a	5.1	3.5	1.4	.2
2	a	4.9	3.0	1.4	.2

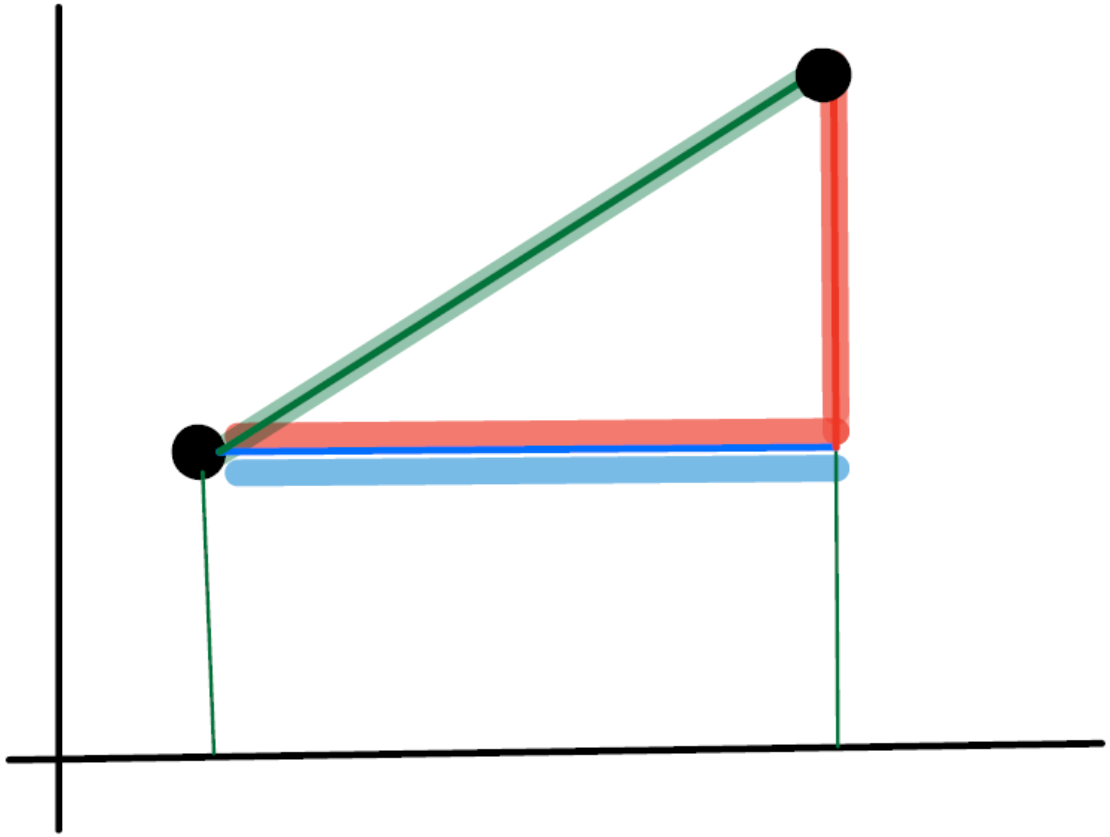
$$d_{ij} = \max_{r=1,2,\ldots,p} \left| x_{ir} - x_{jr} \right|$$



$$d_{ij} = \max_{r=1,2,\ldots,p} \left| x_{ir} - x_{jr} \right|$$

$p = 2$

Οι τρεις βασικές αποστάσεις



Μέθοδοι Ομαδοποίησης

Κατάταξη των μεθόδων ομαδοποίησης

Α. Ιεραρχικές μέθοδοι

- **συσσωρευτικές μέθοδοι**
- **διαιρετικές μέθοδοι**

Β. Μη ιεραρχικές μέθοδοι

Μη ιεραρχικές μέθοδοι ομαδοποίησης

Στόχος: να ομαδοποιήσουν τις *n* μονάδες των δεδομένων σε *k* ομάδες όπου το *k* είναι καθορισμένο από την αρχή (περιορισμός της μεθόδου)

Τρόποι αντιμετώπισης: Δοκιμάζουμε με διαφορετικές επιλογές ως προς το πλήθος των ομάδων ή με κάποιον άλλο τρόπο αποφασίζουμε τον σωστό αριθμό των ομάδων.

Μη ιεραρχικές μέθοδοι ομαδοποίησης

Μηχανισμός λειτουργίας

θεωρούν *k* συγκεκριμένα άτομα (μητρικά σημεία - seed points) και γύρω από αυτά ταξινομούν τα υπόλοιπα στοιχεία έως ότου διαμορφωθούν οι επιθυμητές ομάδες

Μη ιεραρχικές μέθοδοι ομαδοποίησης

Τρόποι δημιουργίας μητρικών σημείων

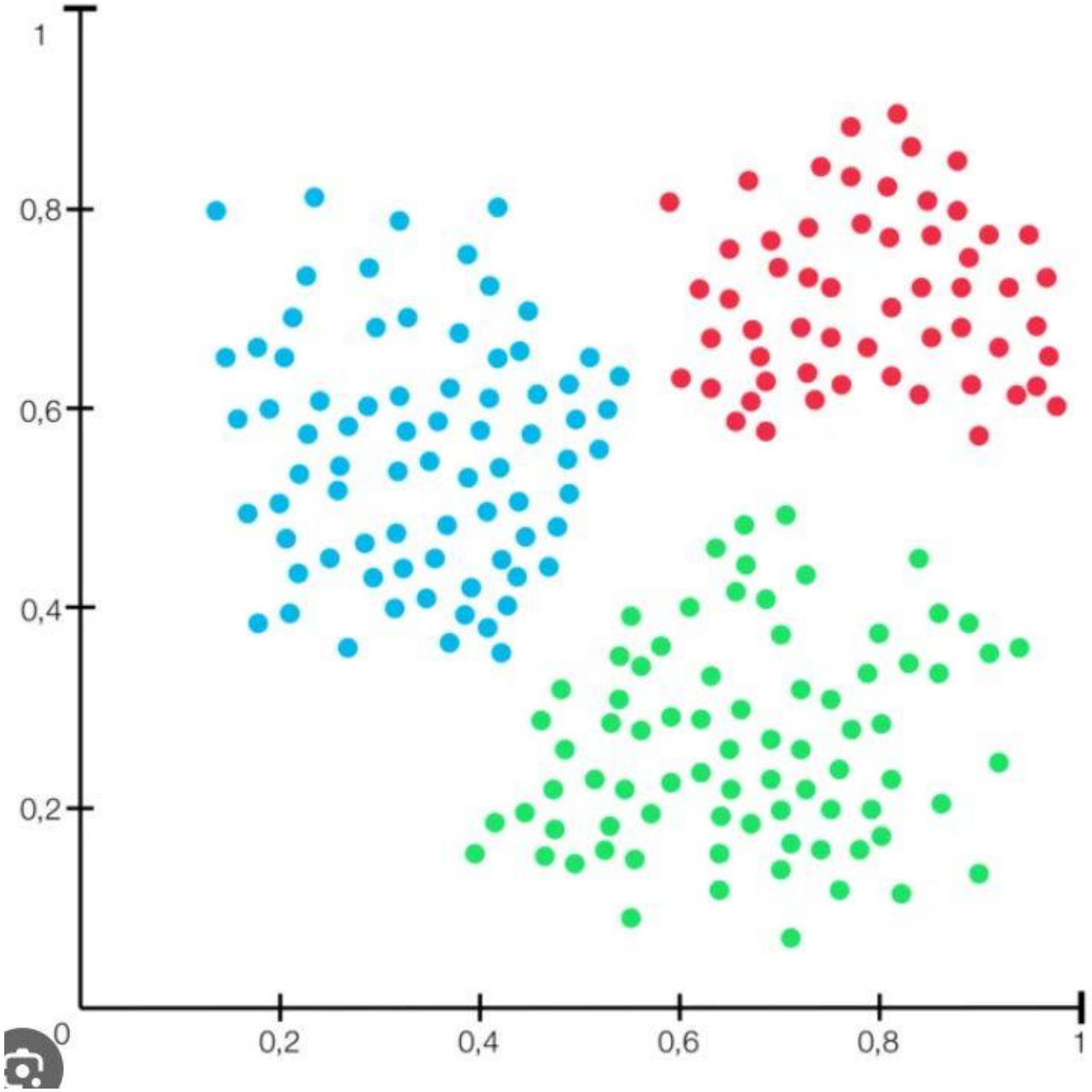
- Διαλέγουμε τα πρώτα *k* στη σειρά άτομα από τα δεδομένα (αυτός είναι ο πιο εύκολος και οικονομικός τρόπος).
- Αριθμούμε τα άτομα από το 1 έως το *n* και διαλέγουμε αυτά με την αρίθμηση *n/k*, 2*n/k*, ..., και *n*.
- Αριθμούμε τα δεδομένα από το 1 έως το *n*, δημιουργούμε *k* διαφορετικούς τυχαίους αριθμούς από 1 έως το *n* και επιλέγουμε τα άτομα που αντιστοιχούν στους αριθμούς αυτούς.

Μη ιεραρχικές μέθοδοι ομαδοποίησης

Τρόποι δημιουργίας μητρικών σημείων (Astrahan)

- Υπολογίζουμε την 'πυκνότητα' ('density') για κάθε άτομο η οποία ορίζεται ως το πλήθος των ατόμων που βρίσκονται γύρω από αυτό μέσα σε μια καθορισμένη περιοχή ακτίνας (απόστασης) d_1.
- Διατάσσουμε όλα τα στοιχεία με βάση την 'πυκνότητα' και επιλέγουμε αυτό με τη μεγαλύτερη 'πυκνότητα' ως το πρώτο μητρικό σημείο.
- Επιλέγουμε διαδοχικά μητρικά σημεία με σειρά τέτοια ώστε να ελαττώνεται η 'πυκνότητα' και συγχρόνως κάθε νέο μητρικό σημείο να απέχει το λιγότερο μια ελάχιστη απόσταση, έστω d_2 από όλα τα προηγούμενα μητρικά σημεία που έχουμε επιλέξει. Συνεχίζουμε να επιλέγουμε μητρικά σημεία μέχρι όλα τα εναπομείναντα στοιχεία να έχουν 'πυκνότητα' μηδέν (δηλαδή να έχουν απόσταση τουλάχιστον d_1 από κάθε άλλο στοιχείο).
- Αν προκύψουν περισσότερα μητρικά σημεία από όσα θέλουμε, τότε ιεραρχικά ομαδοποιούμε τα μητρικά σημεία έως ότου έχουμε ακριβώς k.

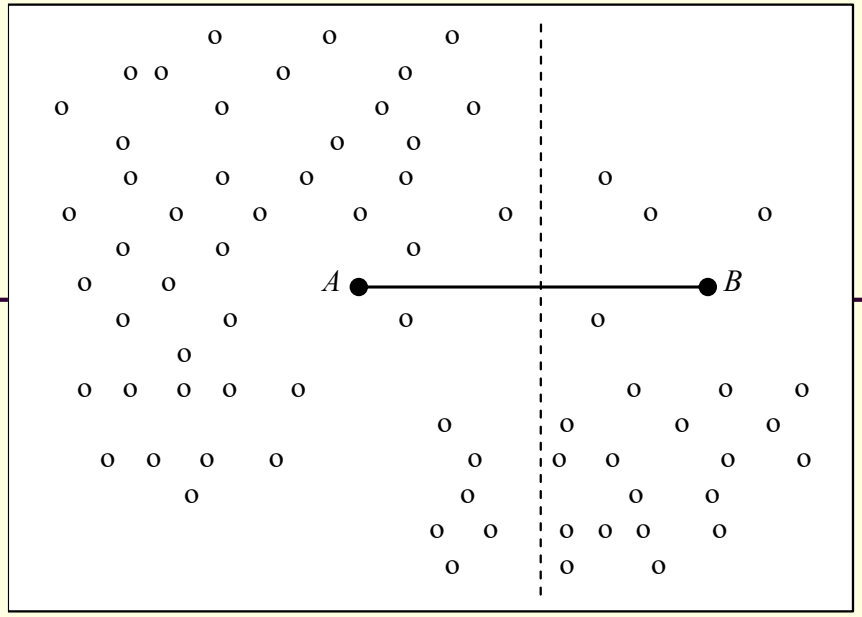
($d_2 > d_1$).

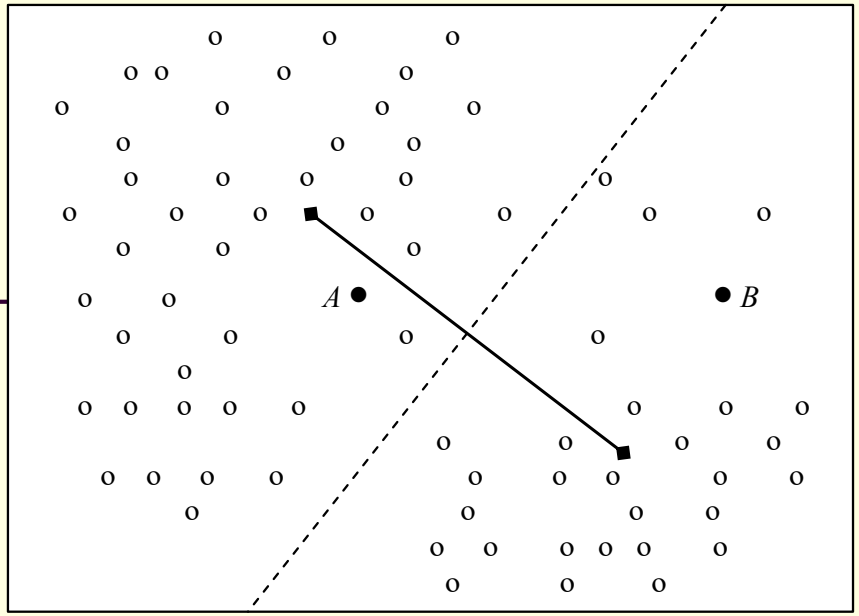

1
0,8
0,6
0,4
0,2
0
0,2
0,4
0,6
0,8
1

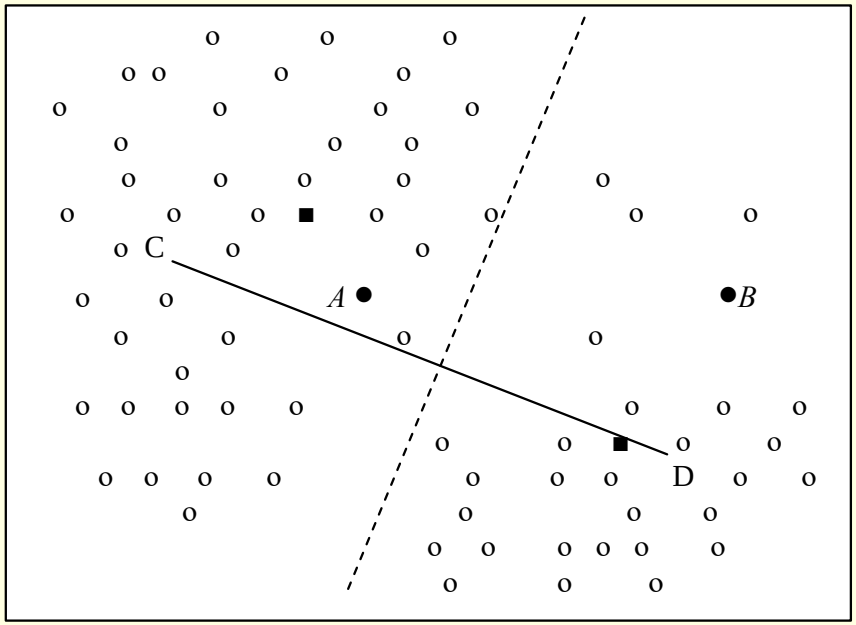
Αλγόριθμοι υλοποίησης μη-Ιεραρχικών μεθόδων : οι γενικές αρχές

- Οι μέθοδοι δουλεύουν επαναληπτικά και χρησιμοποιούν την έννοια του κέντρου μιας ομάδας (κέντρου βάρους, centroid) το οποίο αντιστοιχεί στο διάνυσμα των μέσων ανα μεταβλητή για όλες τις παρατηρήσεις της ομάδας.
- οι παρατηρήσεις κατατάσσονται σε ομάδες ανάλογα με την απόστασή τους από τα κέντρα των ήδη διαμορφωμένων ομάδων.
- Η διαφοροποίηση των μεθόδων έγκειται στο σημείο όπου γίνεται η ανανέωση των κέντρων των ομάδων και η ταξινόμηση των υπολοίπων παρατηρήσεων σε αυτές.
- Συνήθως η απόσταση που χρησιμοποιείται για την κατάταξη των παρατηρήσεων είναι η ευκλείδεια.

Η μέθοδος του Forgy







Η μέθοδος του MacQueen (*k*-means method)

Βήμα 1ο. Καθόρισε αρχικά έναν αρχικό σύνολο από k **μητρικά σημεία** χρησιμοποιώντας k από τα n άτομα που είναι διαθέσιμα.

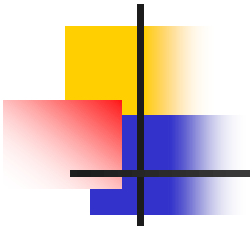
Βήμα 2ο. Κατάταξε καθένα από τα εναπομείναντα $n-k$ άτομα στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από το άτομο. Μετά από κάθε τοποθέτηση υπολόγισε ξανά το κέντρο βάρους της αλλαγμένης πλέον ομάδας.

Βήμα 3ο. Όταν όλα τα άτομα έχουν τοποθετηθεί σε ομάδες μέσω του βήματος 2, θεώρησε τα δημιουργηθέντα κέντρα βάρους ως μητρικά σημεία και εκτέλεσε μια ακόμη σάρωση στα δεδομένα τοποθετώντας κάθε άτομο των δεδομένων στο πλησιέστερο μητρικό σημείο.

Ιεραρχικές μέθοδοι

Παράγουν μια ιεραρχία «δενδροειδούς μορφής» όπου στα διάφορα στάδια το πλήθος k των ομάδων παίρνει όλες τις δυνατές τιμές από το 1 έως το n.

Στο ένα άκρο της ιεραρχίας υπάρχει μια μόνο ομάδα που περιέχει n άτομα και στο άλλο άκρο υπάρχουν n ομάδες όπου η καθεμιά περιέχει ένα μόνο άτομο.



Είδη ιεραρχικών μεθόδων

- **Συσσωρευτικές μέθοδοι** (agglomerative methods)

  Ξεκινούν με *n* ομάδες και με διαδοχικές συγχωνεύσεις καταλήγουν σε μια ομάδα που περιέχει όλα τα άτομα που υπάρχουν στα δεδομένα.

- **Διαιρετικές μέθοδοι** (divisive methods)

  Εκτελούν την αντίθετη διεργασία δηλαδή ξεκινούν με μια μόνο ομάδα που περιέχει *n* άτομα και διαιρούν τα δεδομένα σε όλο και μικρότερες ομάδες έως ότου κάθε ομάδα να περιέχει ένα και μόνο άτομο

Συσσωρευτικές μέθοδοι

Βήμα 1ο. Ξεκίνα με n ομάδες (clusters) του ενός ατόμου η καθεμιά, και με τον nxn πίνακα των αποστάσεων **D**=[d_{ij}] (εναλλακτικά θα μπορούσε να χρησιμοποιηθεί ένας πίνακας μέτρων ομοιότητας).

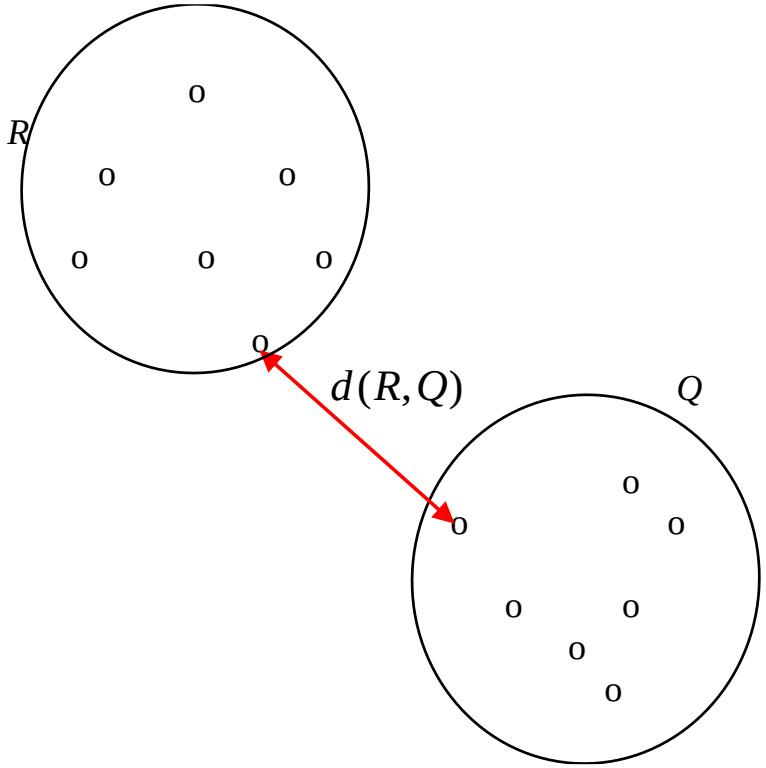
Βήμα 2ο. Εντόπισε στον πίνακα **D** το ζεύγος των πλησιέστερων (πιο όμοιων) ομάδων έστω Q και R.

Βήμα 3ο. Συγχώνευσε τις ομάδες Q και R σε μια ομάδα, την P =(QR) μειώνοντας έτσι τον αριθμό των ομάδων κατά ένα. Ανανέωσε τον πίνακα αποστάσεων **D** διαγράφοντας τις γραμμές και στήλες που αντιστοιχούσαν στις ομάδες Q και R, και προσθέτοντας μια γραμμή και μια στήλη που περιέχει τις αποστάσεις της ομάδας P=(QR) από τις υπόλοιπες ομάδες.

Βήμα 4ο. Επανάλαβε τα βήματα 2 και 3 συνολικά $n-1$ φορές έτσι ώστε με τη λήξη του αλγορίθμου, όλα τα άτομα να αποτελούν μία μόνο ομάδα. Σε κάθε βήμα κατάγραψε τις λεπτομέρειες συνένωσης, την ταυτότητα των ομάδων και τα επίπεδα (distances ή similarities) στα οποία πραγματοποιούνται οι συγχωνεύσεις.

Μέθοδος της απλής συνένωσης (Single Linkage Method)

$$d(R,Q) = \min_{i \in R,\ j \in Q} d_{ij}$$



Το… παρατσούκλι

«**μέθοδος του πλησιέστερου (κοντινότερου) γείτονα**» (nearest neighbor method)

Παράδειγμα

άτομο	1	2	3	4	5
1	0				
2	14	0			
3	8	12	0		
4	11	10	14	0	
5	16	15	7	13	0

Ας θεωρήσουμε ένα σύνολο 5 ατόμων $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$ με τον παραπάνω πίνακα αποστάσεων

- Ξεκινάμε με 5 ομάδες, κάθε ομάδα αποτελούμενη από ένα άτομο δηλαδή τις (1), (2), (3), (4), (5).
- $\min d_{ij} = 7 = d_{35}$
- (1), (2), (3,5), (4).

Ομάδα	(3,5)	(1)	(2)	(4)
(3,5)	0			
(1)	8	0		
(2)	12	14	0	
(4)	13	11	10	0

$$d((3,5),(1)) = \min(d_{31}, d_{51}) = \min(8,16) = 8$$

- $d((3,5),(2)) = \min(d_{32}, d_{52}) = \min(12,15) = 12$

$$d((3,5),(4)) = \min(d_{43}, d_{54}) = \min(14,13) = 13.$$

Παράδειγμα 3.2.1

Ομάδα	(3,5)	(1)	(2)	(4)
(3,5)	0			
(1)	8	0		
(2)	12	14	0	
(4)	13	11	10	0

- Η ελάχιστη τιμή που εμφανίζεται στο νέο πίνακα είναι η

$$d((1),(3,5)) = 8$$

- Συγχωνεύονται οι ομάδες (1) και (3,5) δημιουργώντας έτσι την ομάδα (1,3,5).

Ομάδα	(1,3,5)	(2)	(4)
(1,3,5)	0		
(2)	12	0	
(4)	11	10	0

Παράδειγμα 3.2.1

Ομάδα	(1,3,5)	(2)	(4)
(1,3,5)	0		
(2)	12	0	
(4)	11	10	0

- Η ελάχιστη τιμή του πίνακα είναι η

$$d((2),(4)) = 10$$

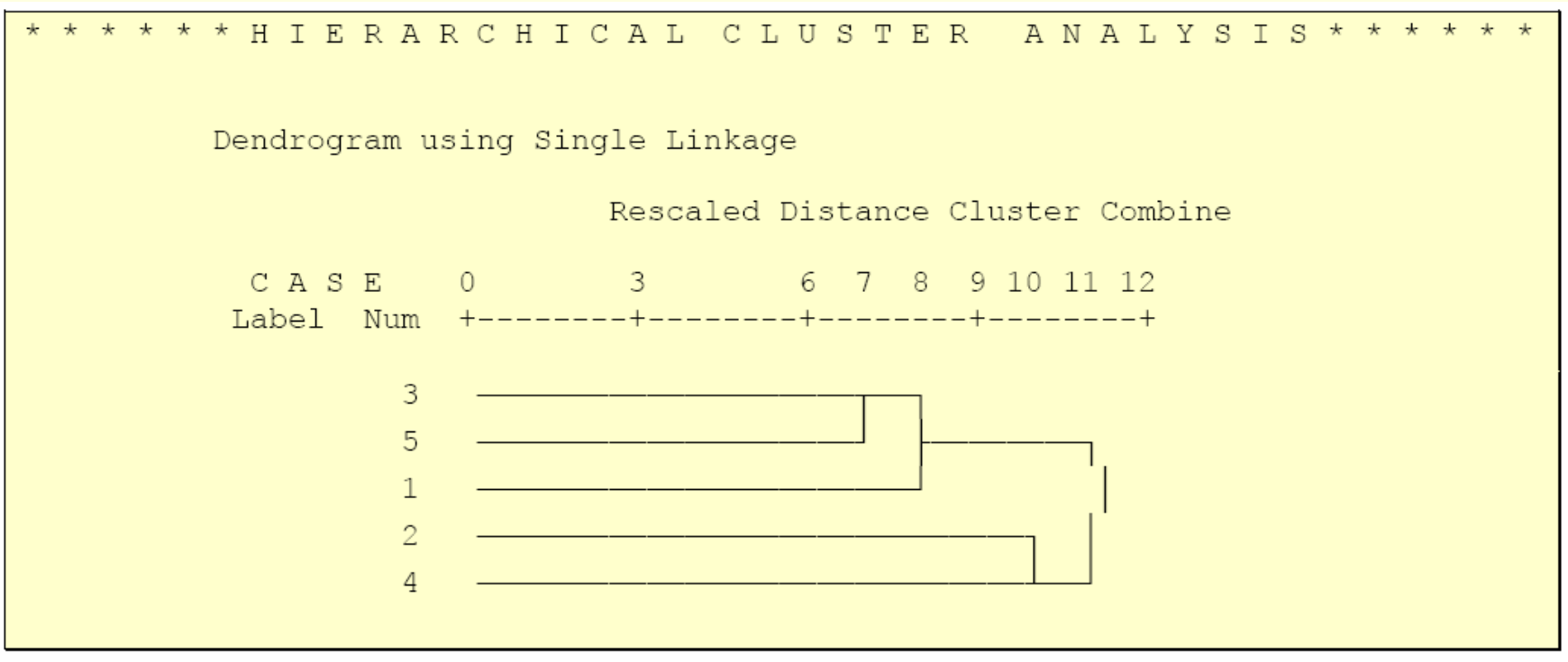
- Ενώνονται τα άτομα 2 και 4 ώστε να προκύψει η ομάδα (2,4).

Ο νέος πίνακας αποστάσεων παίρνει την μορφή

Ομάδα	(1,3,5)	(2,4)
(1,3,5)	0	
(2,4)	11	0

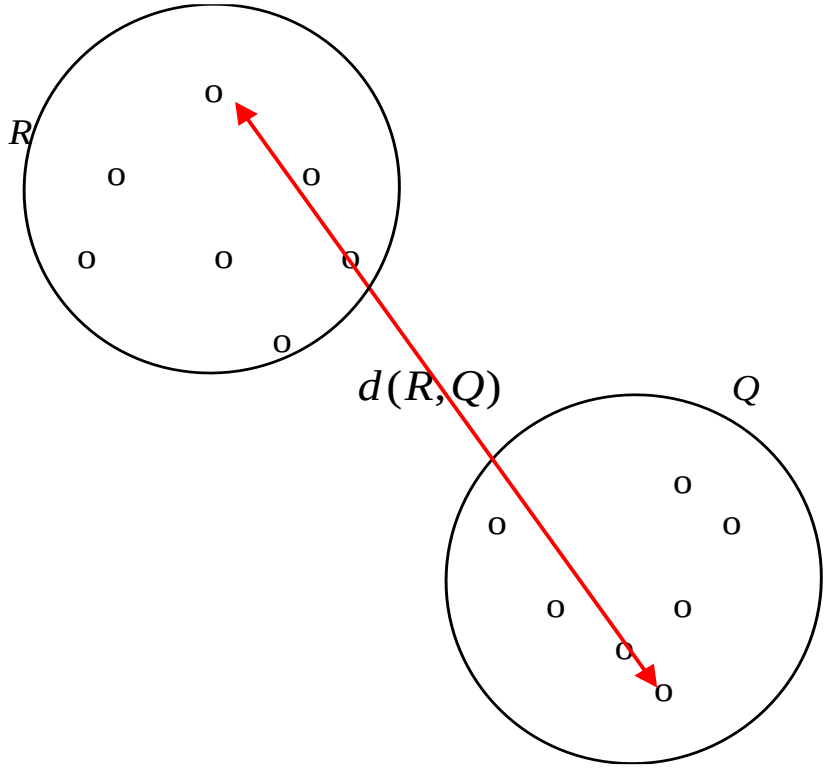
και προχωρώντας στην τελευταία συνένωση (των ομάδων (1,3,5) και (2,4)) φτάνουμε σε μια ομάδα που περιέχει όλα τα άτομα.

Το δενδρόγραμμα



Μέθοδος της πλήρους συνένωσης (Complete Linkage Method)

$$d(R,Q) = \max_{i \in R,\ j \in Q} d_{ij}$$



Το… παρατσούκλι

«**μέθοδος του μακρινότερου γείτονα**» (furthest neighbor method))

Μέθοδος των σταθμισμένων μέσων (Weighted Average Linkage Method)

Η απόσταση μεταξύ ομάδων ορίζεται ως ο μέσος των αποστάσεων όλων των στοιχείων της μιας ομάδας με τα στοιχεία της άλλης.

Αν για παράδειγμα η μια ομάδα περιλαμβάνει τις παρατηρήσεις 1,2,4 και η άλλη τις παρατηρήσεις 3,5 τότε η απόστασή τους είναι ο μέσος των αποστάσεων $d_{13}, d_{15}, d_{23}, d_{25}, d_{43}, d_{45}$, δηλαδή

$$d((124),(35)) = \frac{d_{13} + d_{15} + d_{23} + d_{25} + d_{43} + d_{45}}{6}$$

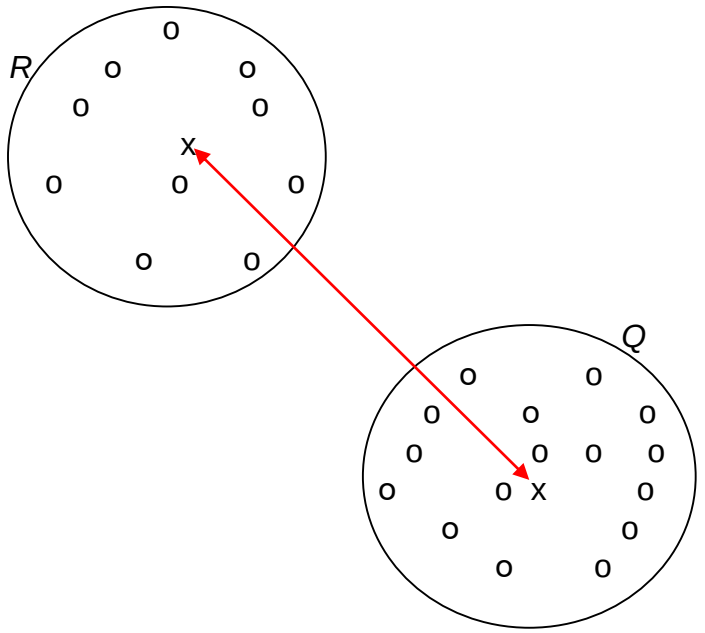
Μέθοδος των των κέντρων βάρους (Centroid method)

Κέντρο βάρους μιας ομάδας *R*:

$$\bar{x}(R) = (\bar{x}_1(R), \bar{x}_2(R), ..., \bar{x}_p(R))$$

όπου

$$\bar{x}_r(R) = \frac{1}{|R|}\sum_{i \in R} x_{ir} \quad \text{για } r = 1, 2, ..., p.$$



Απόσταση δύο ομάδων

$$d(R,Q) = d\left(\bar{x}(R), \bar{x}(Q)\right) = \sqrt{\sum_{r=1}^{p}\left(\bar{x}_r(R) - \bar{x}_r(Q)\right)^2} \ .$$

Μέθοδος του Ward (Ward's method): το άθροισμα των τετραγωνικών αποκλίσεων

Χαρακτηριστική ιδιότητα της μεθόδου	ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες

Αθροίζοντας για όλα τα άτομα μιας ομάδας C παίρνουμε το λεγόμενο ***άθροισμα των τετραγωνικών αποκλίσεων*** της ομάδας C (Error Sum of Squares)

$$ESS(C) = \sum_{i \in C} (d(x_i, \bar{x}(C)))^2$$

το οποίο χρησιμοποιείται ως μέτρο συνεκτικότητας της ομάδας.

Αν κάποια στιγμή υπάρχουν k ομάδες, τότε, προσθέτοντας τα αθροίσματα των τετραγωνικών αποκλίσεων για όλες, προκύπτει το **συνολικό άθροισμα τ.α.**

$$ESS = ESS_1 + ESS_2 + \ldots + ESS_k .$$

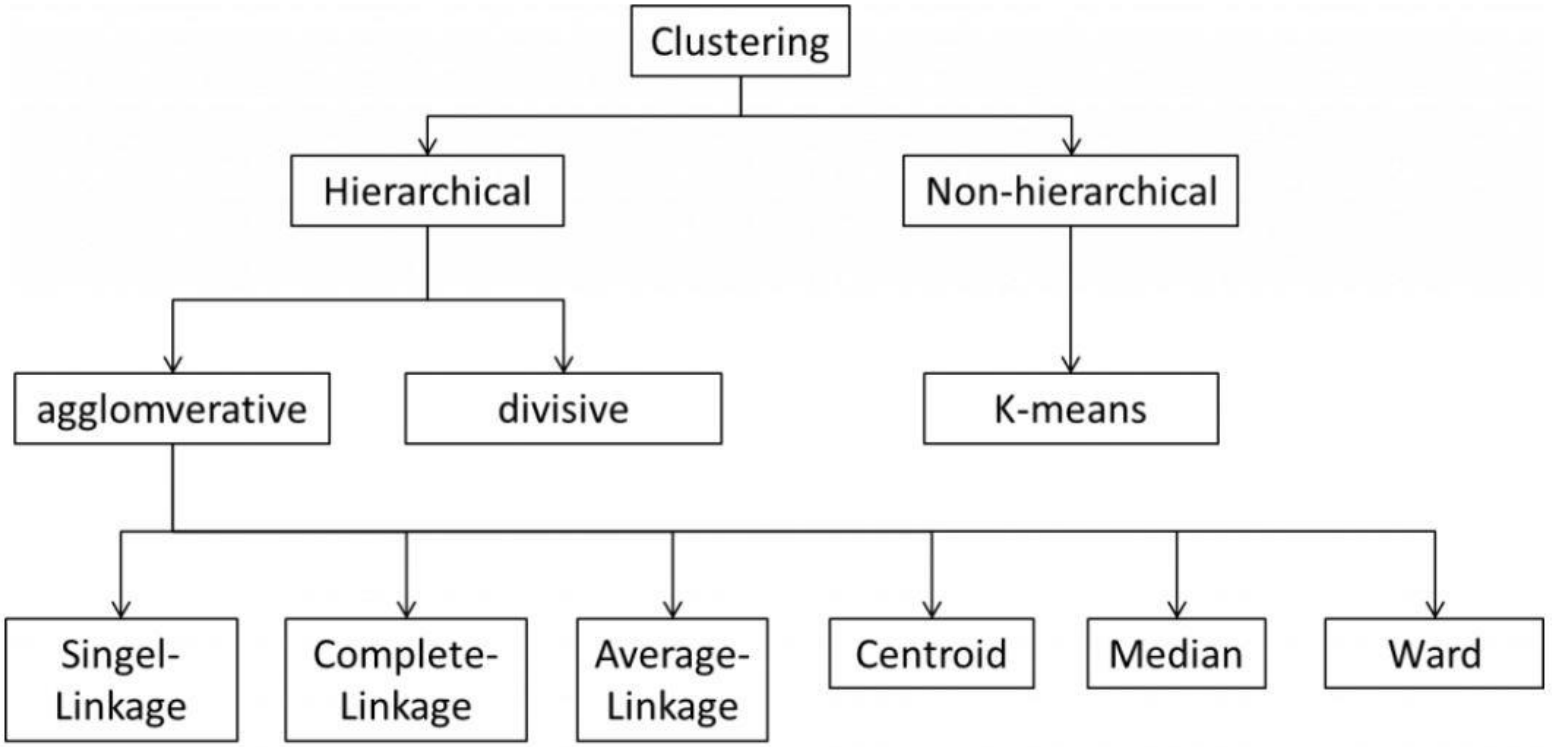
Μέθοδος του Ward το άθροισμα των τετραγωνικών αποκλίσεων

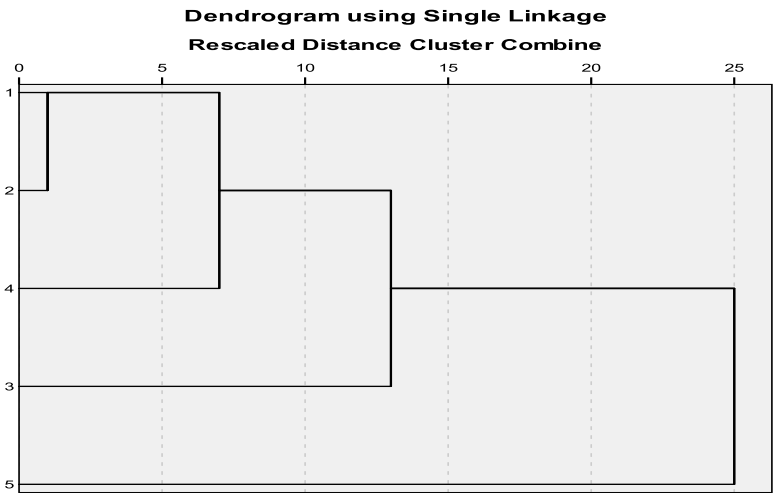
- Αρχικά, ισχύει $ESS_r = 0, r = 1,2,...,n$, οπότε $ESS=0$.
- Στα ενδιάμεσα βήματα 'δοκιμάζονται' όλες οι δυνατές συγχωνεύσεις των ομάδων κατά ζεύγη και τελικά συγχωνεύονται εκείνες οι δύο ομάδες, που η ομαδοποίησή τους οδηγεί στην **μικρότερη αύξηση του συνολικού αθροίσματος τετραγωνικών αποκλίσεων** *ESS* (ελάχιστη απώλεια πληροφορίας).
- Τελικά, όταν και τα *n* άτομα συγχωνευτούν σε μία μόνο ομάδα θα έχουμε

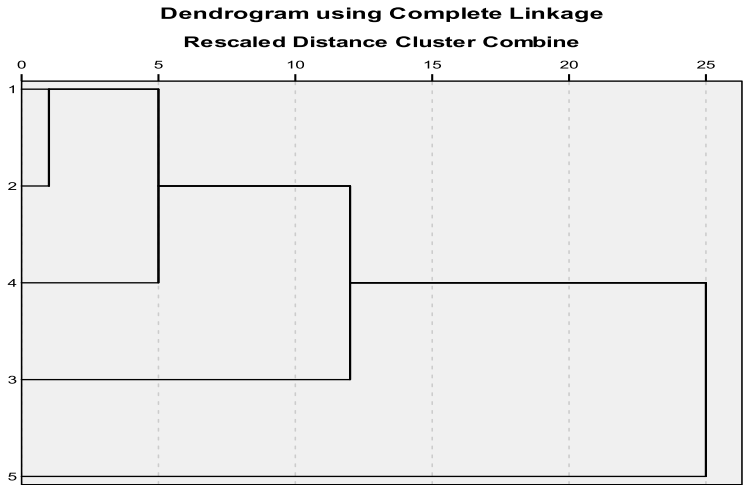
$$ESS = \sum_{j=1}^{N} (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})$$

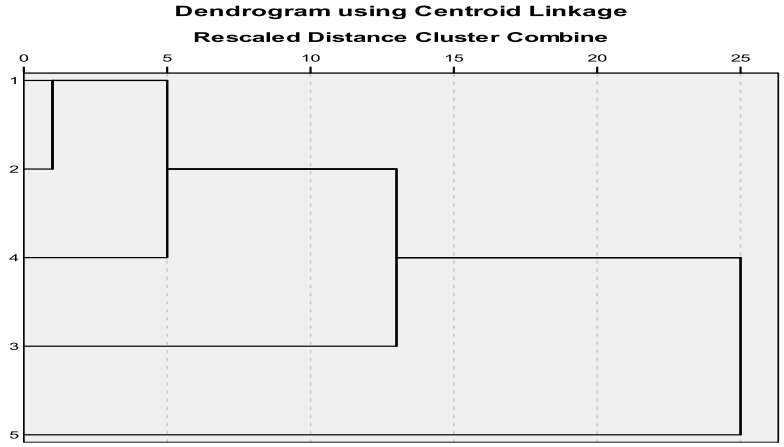
όπου $\mathbf{x}_j$ είναι η πολυδιάστατη μέτρηση του *j*–στου ατόμου και $\bar{\mathbf{x}}$ είναι η μέση τιμή όλων των ατόμων.

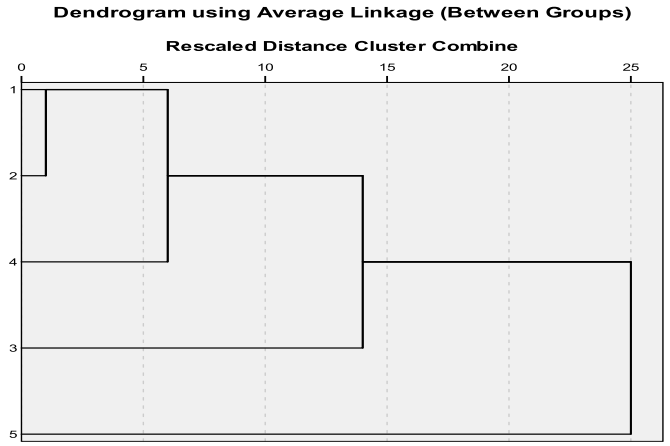
Μέθοδοι ομαδοποίησης

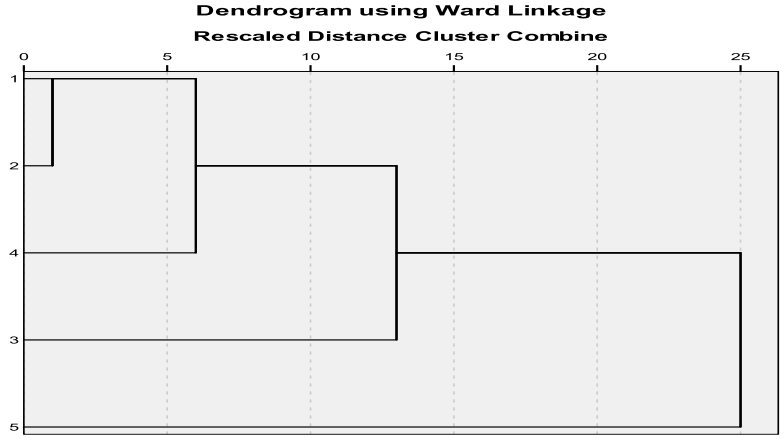


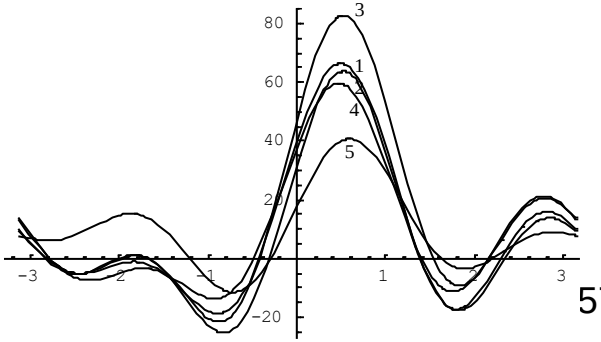

Dendrogram using Single Linkage
Rescaled Distance Cluster Combine
0 5 10 15 20 25
1
2
4
3
5


Dendrogram using Complete Linkage
Rescaled Distance Cluster Combine
0 5 10 15 20 25
1
2
4
3
5


Dendrogram using Centroid Linkage
Rescaled Distance Cluster Combine
0 5 10 15 20 25
1
2
4
3
5


Dendrogram using Average Linkage (Between Groups)
Rescaled Distance Cluster Combine
0 5 10 15 20 25
1
2
4
3
5


Dendrogram using Ward Linkage
Rescaled Distance Cluster Combine
0 5 10 15 20 25
1
2
4
3
5

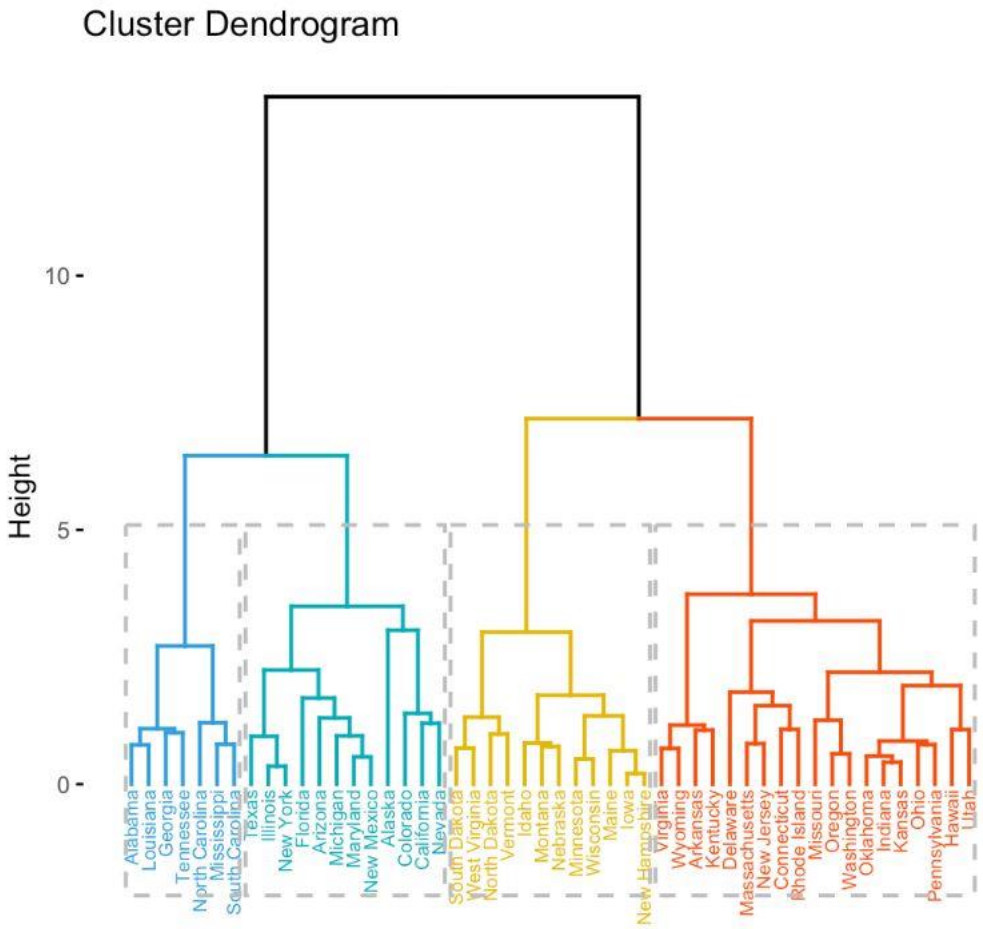

80
60
40
20
-20
-3
-2
-1
1
3
1
2
4
5
3

Επιλογή του πλήθους των ομάδων μέσω δενδρογράμματος

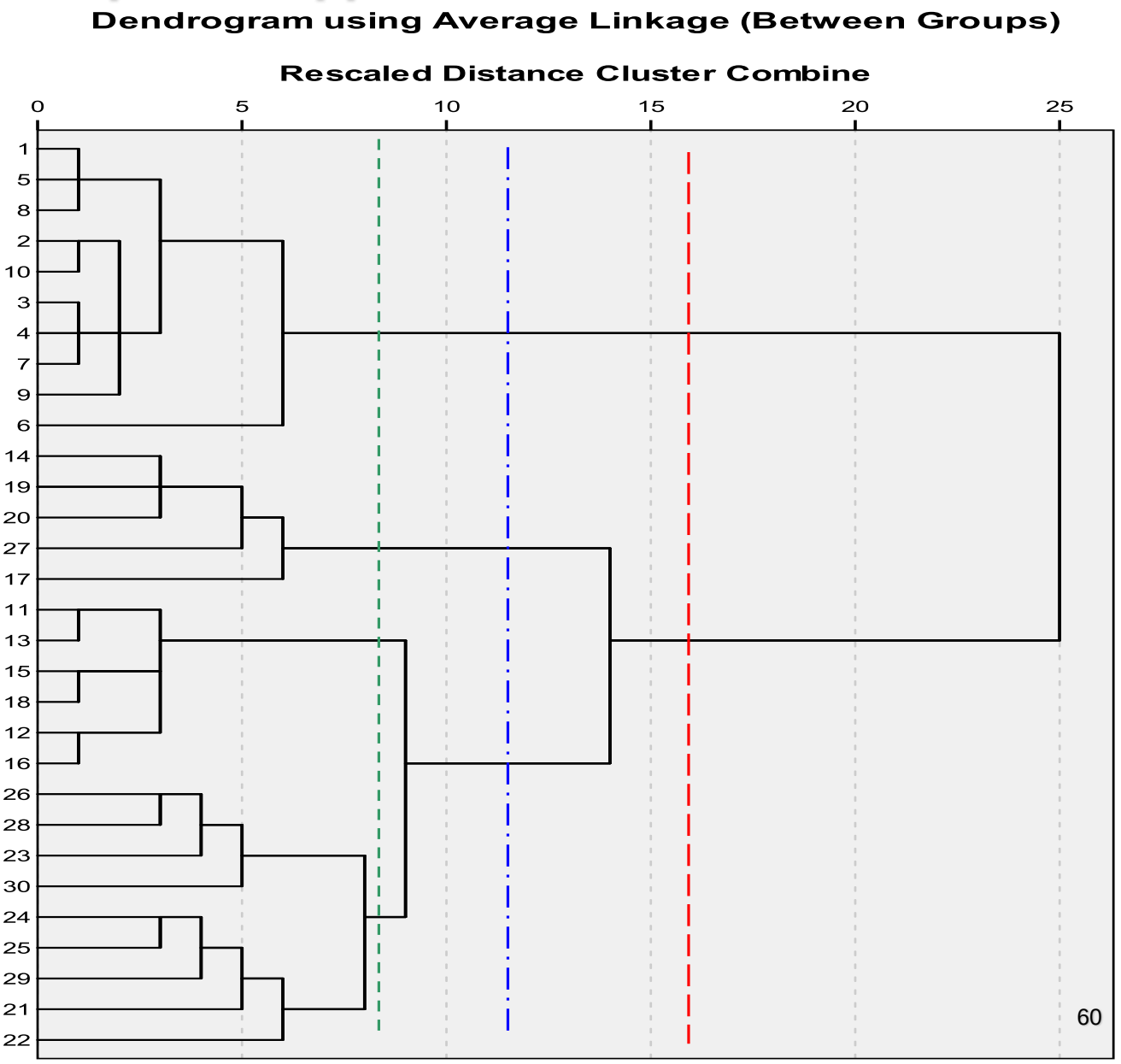
- Σε εκείνο το σημείο του δενδρογράμματος που παρατηρείται η μεγαλύτερη μεταβολή της ποσότητας που καταγράφεται στον οριζόντιο άξονα (απόσταση ή μέτρο ομοιότητας), φέρνουμε μια παράλληλη γραμμή προς τον κατακόρυφο άξονα και να δούμε σε πόσα σημεία τέμνει το δενδρόγραμμα.

- Το πλήθος *k*, για το οποίο παρατηρούμε μεγάλες αποστάσεις συνένωσης σε σχέση με το προηγούμενο (*k*-1 ομάδες) αποτελεί μια λογική τιμή για το βέλτιστο πλήθος των ομάδων.

Επιλογή του πλήθους των ομάδων



Παράδειγμα


Dendrogram using Average Linkage (Between Groups)
Rescaled Distance Cluster Combine
0
5
10
15
20
25
1
5
8
2
10
3
4
7
9
6
14
19
20
27
17
11
13
15
18
12
16
26
28
23
30
24
25
29
21
22

Ανάλυση κατά συστάδες

Iris Data

Fisher Iris Data

Μεταβλητές

SL : Μήκος Σεπάλου
SW: Πλάτος Σεπάλου
PL : Μήκος Πετάλου
PW:Πλάτος Πετάλου

Ποικιλίες

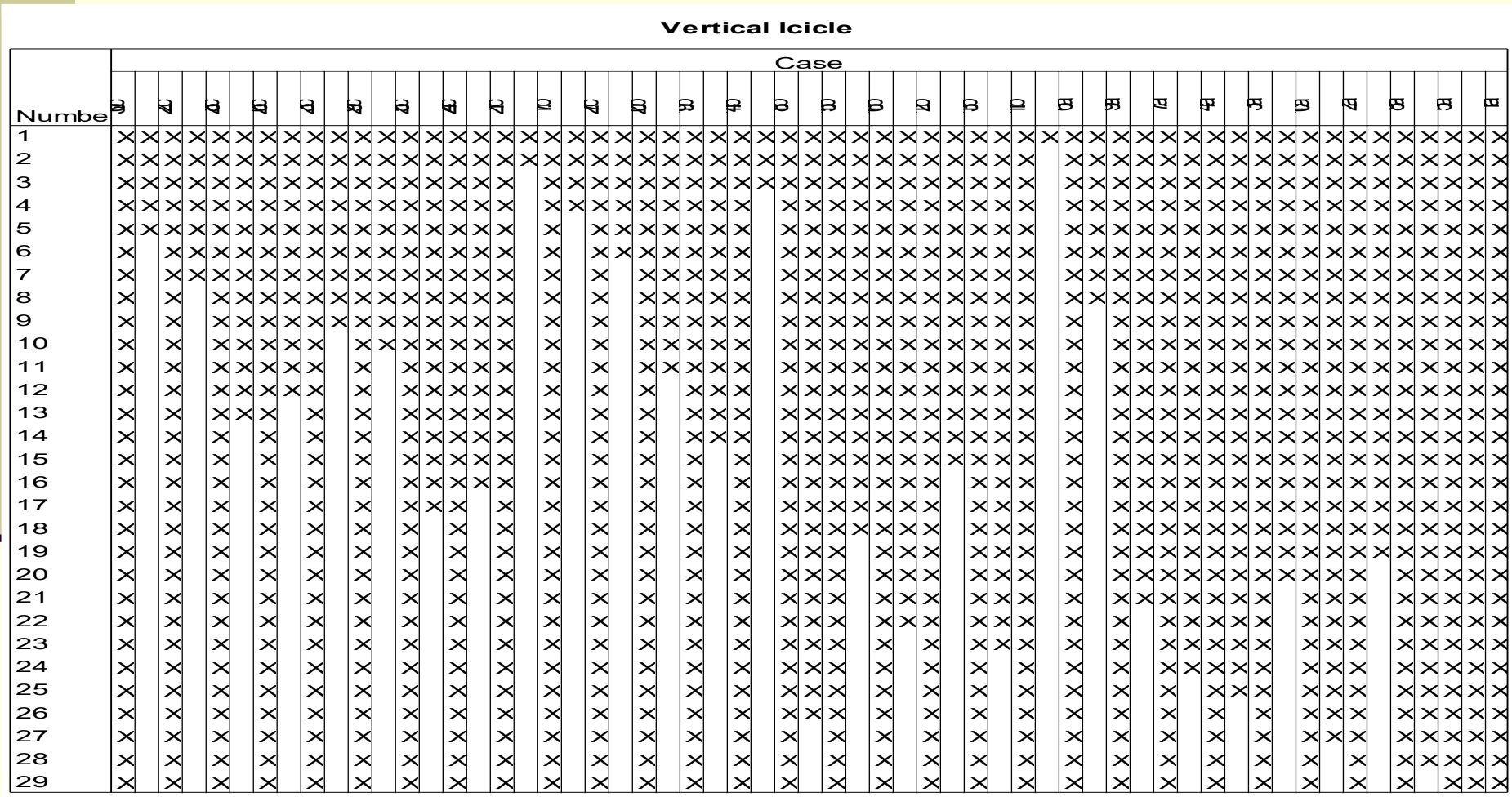
a: Iris Setosa
b: Iris Versicolor
c: Iris Virginica

α/α	Ποικιλία	SL	SW	PL	PW
1	a	5.1	3.5	1.4	.2
2	a	4.9	3.0	1.4	.2
3	a	4.7	3.2	1.3	.2
4	a	4.6	3.1	1.5	.2
5	a	5.0	3.6	1.4	.2
6	a	5.4	3.9	1.7	.4
7	a	4.6	3.4	1.4	.3
8	a	5.0	3.4	1.5	.2
9	a	4.4	2.9	1.4	.2
10	a	4.9	3.1	1.5	.1
11	b	7.0	3.2	4.7	1.4
12	b	6.4	3.2	4.5	1.5
13	b	6.9	3.1	4.9	1.5
14	b	5.5	2.3	4.0	1.3
15	b	6.5	2.8	4.6	1.5
16	b	6.3	3.3	4.7	1.6
17	b	4.9	2.4	3.3	1.0
18	b	6.6	2.9	4.6	1.3
19	b	5.2	2.7	3.9	1.4
20	b	5.7	2.8	4.1	1.3
21	c	6.3	3.3	6.0	2.5
22	c	5.8	2.7	5.1	1.9
23	c	7.1	3.0	5.9	2.1
24	c	6.3	2.9	5.6	1.8
25	c	6.5	3.0	5.8	2.2
26	c	7.6	3.0	6.6	2.1
27	c	4.9	2.5	4.5	1.7
28	c	7.3	2.9	6.3	1.8
29	c	6.7	2.5	5.8	1.8
30	c	7.2	3.6	6.1	1.5

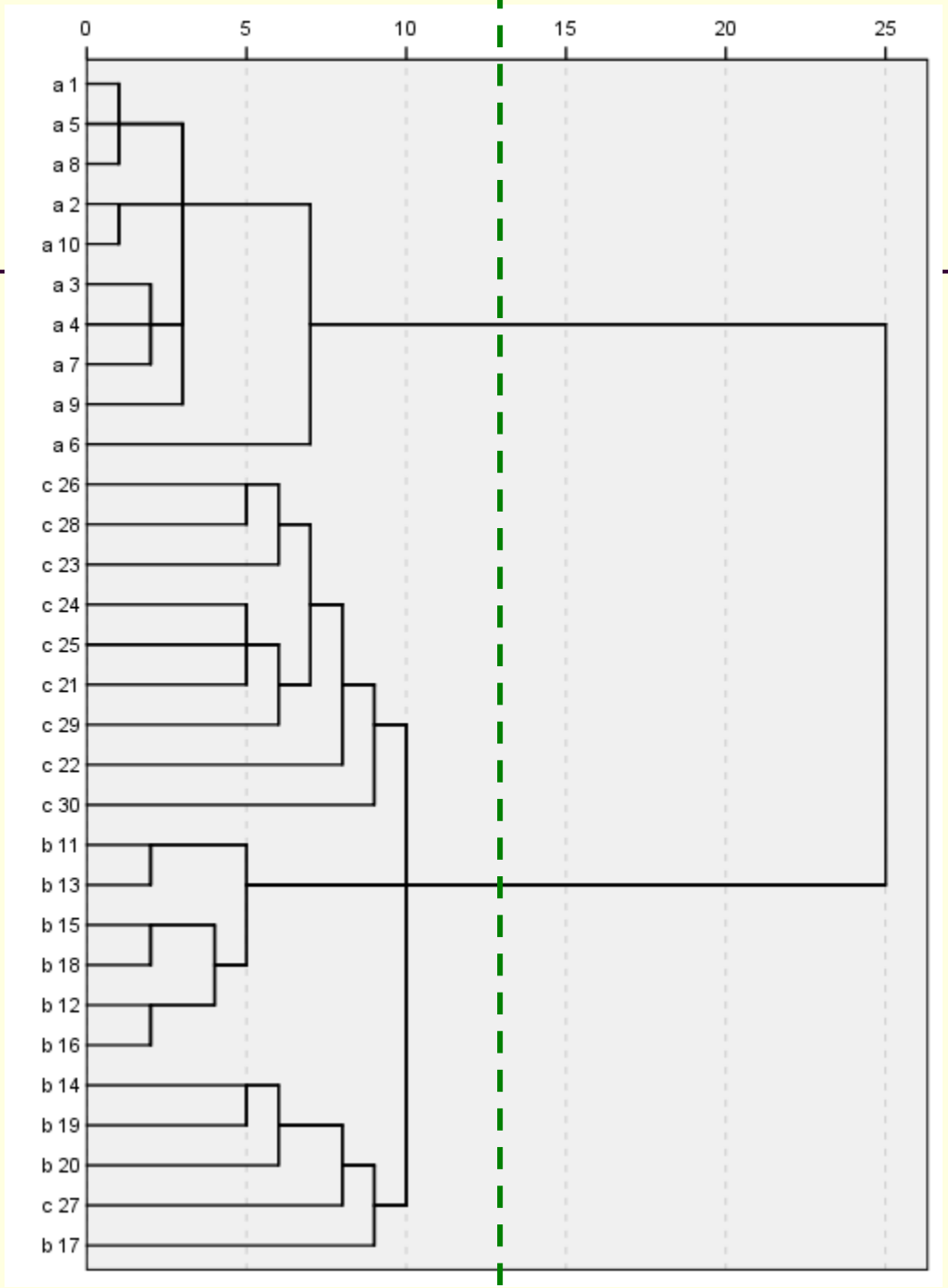
Ιεραρχικές μέθοδοι

Απόσταση: Ευκλείδια Τυποποίηση: Όχι

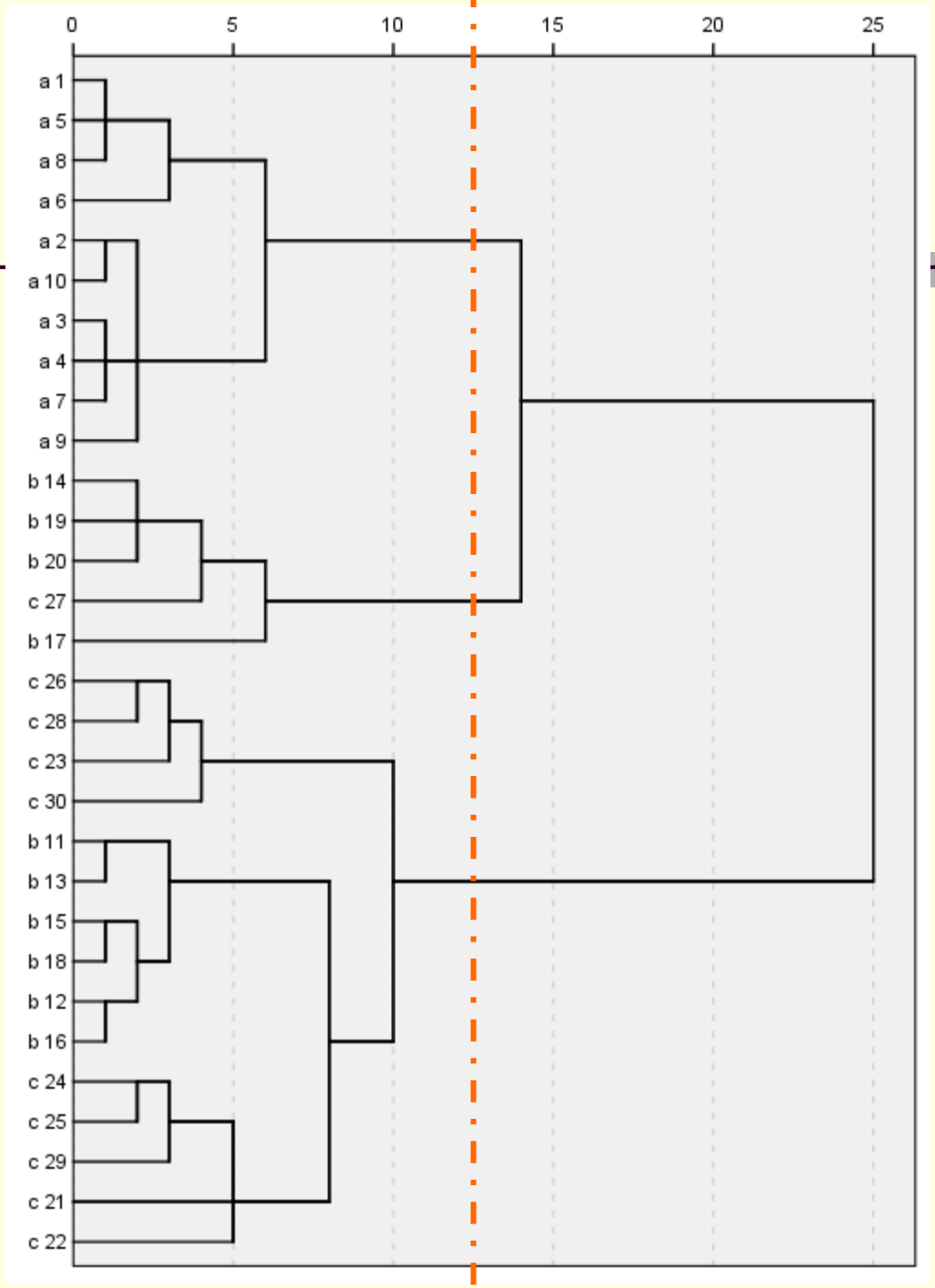
Μέθοδος: Nearest neighbor



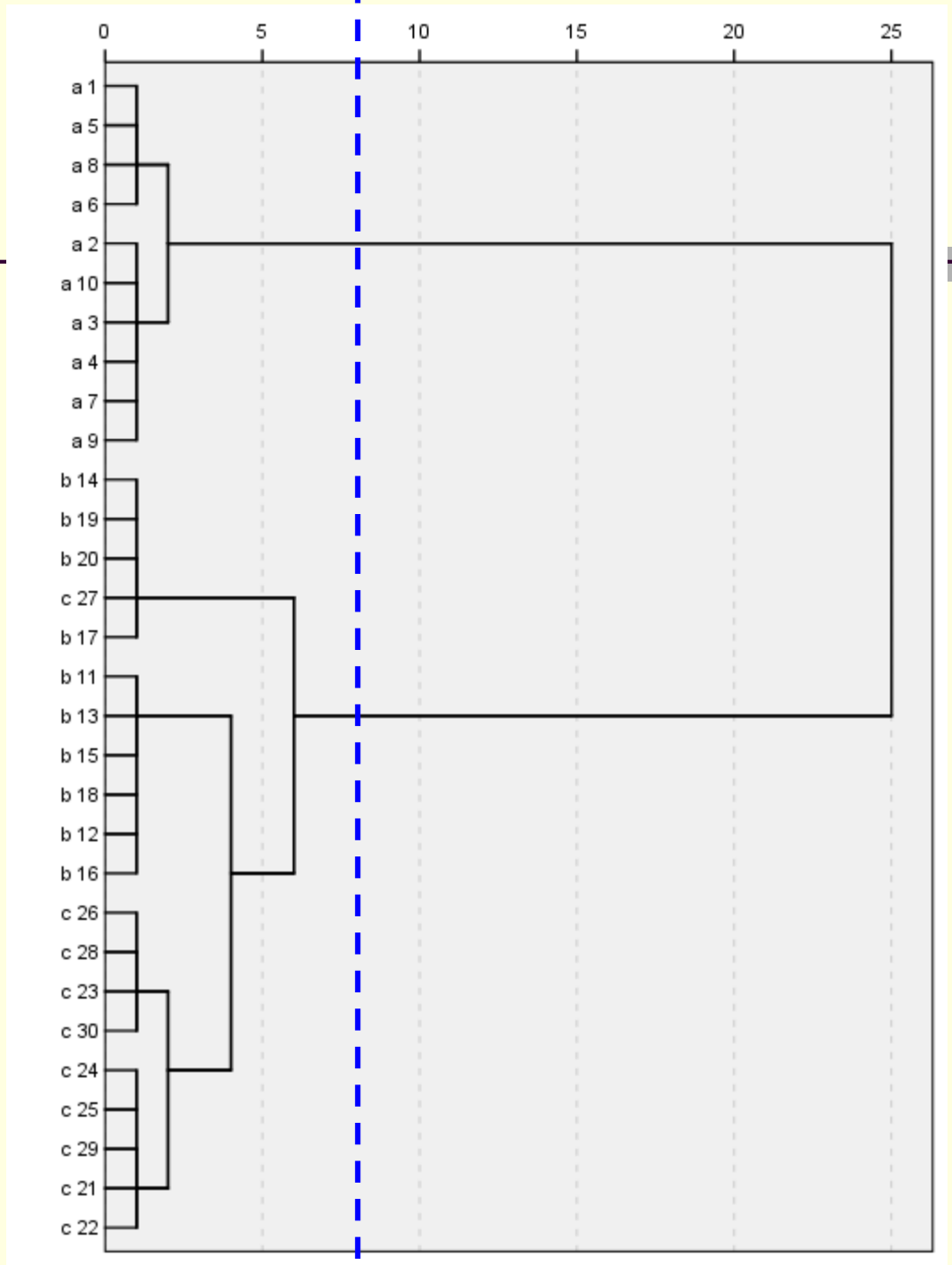
Απόσταση: Ευκλείδια Τυποποίηση: Όχι
Μέθοδος: Nearest neighbor



Απόσταση: Ευκλείδια Τυποποίηση: Όχι
Μέθοδος: Furthest neighbor

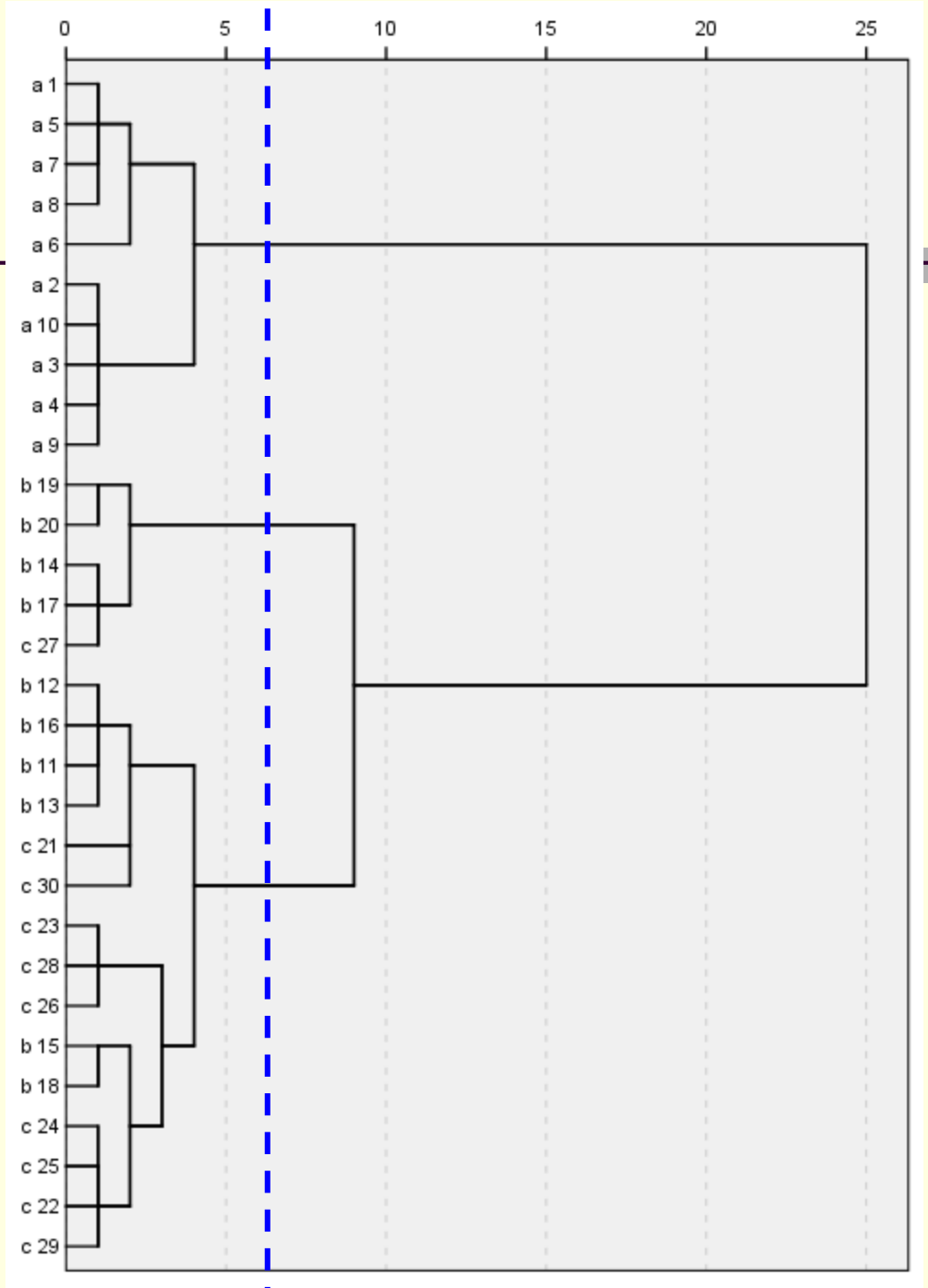


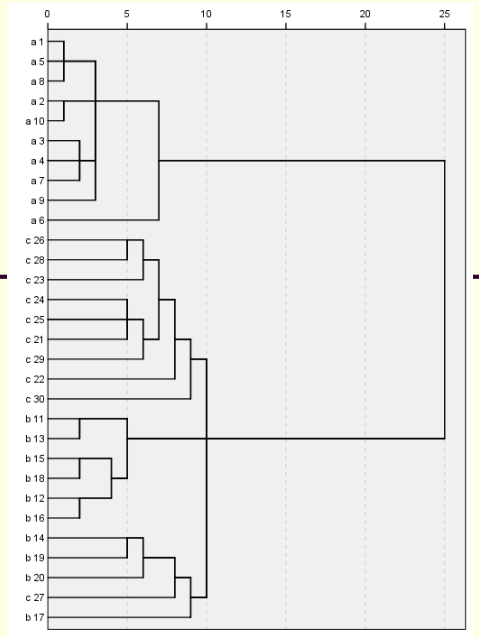
Απόσταση: Ευκλείδια Τυποποίηση: Όχι
Μέθοδος: Ward

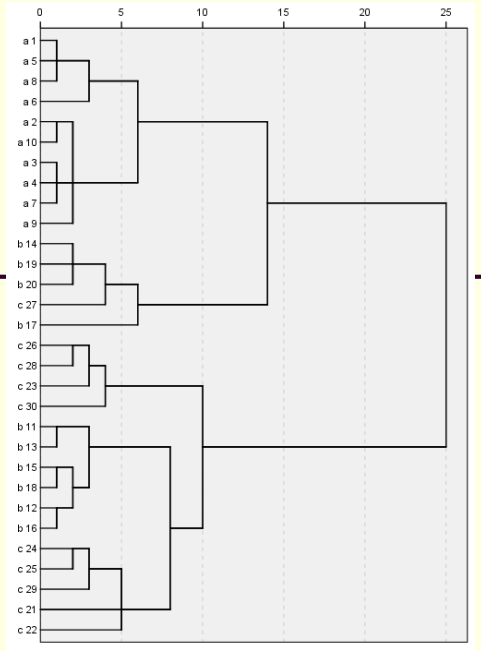


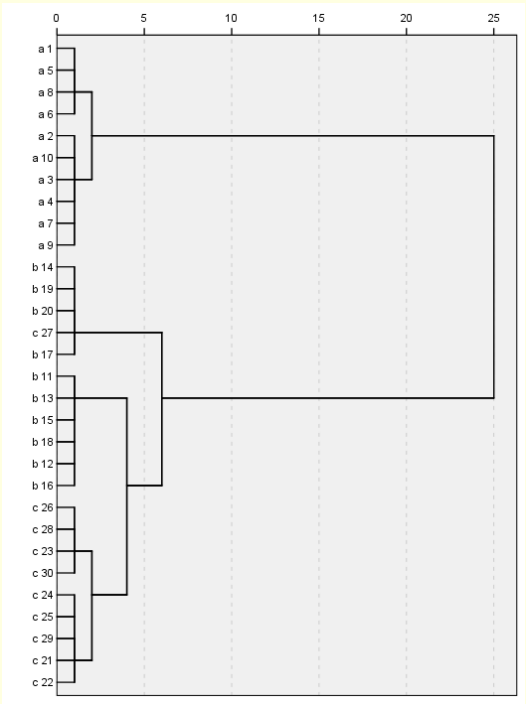
Απόσταση: Ευκλείδια Τυποποίηση: **Ναι**

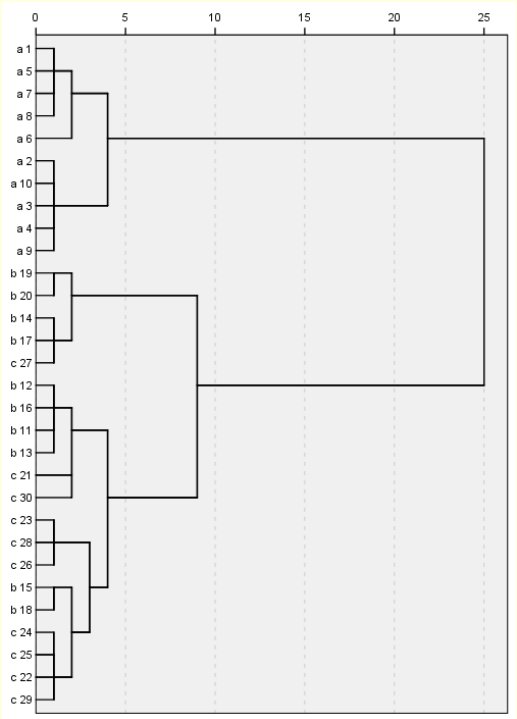
Μέθοδος: Ward




0 5 10 15 20 25
a 1
a 5
a 8
a 2
a 10
a 3
a 4
a 7
a 9
a 6
c 26
c 28
c 23
c 24
c 25
c 21
c 29
c 22
c 30
b 11
b 13
b 15
b 18
b 12
b 16
b 14
b 19
b 20
c 27
b 17


0 5 10 15 20 25
a 1
a 5
a 8
a 6
a 2
a 10
a 3
a 4
a 7
a 9
b 14
b 19
b 20
c 27
b 17
c 26
c 28
c 23
c 30
b 11
b 13
b 15
b 18
b 12
b 16
c 24
c 25
c 29
c 21
c 22


0 5 10 15 20 25
a 1
a 5
a 8
a 6
a 2
a 10
a 3
a 4
a 7
a 9
b 14
b 19
b 20
c 27
b 17
b 11
b 13
b 15
b 18
b 12
b 16
c 26
c 28
c 23
c 30
c 24
c 25
c 29
c 21
c 22


0 5 10 15 20 25
a 1
a 5
a 7
a 8
a 6
a 2
a 10
a 3
a 4
a 9
b 19
b 20
b 14
b 17
c 27
b 12
b 16
b 11
b 13
c 21
c 30
c 23
c 28
c 26
b 15
b 18
c 24
c 25
c 22
c 29