



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ: ΙΕΡΑΡΧΙΚΕΣ ΚΑΙ ΒΑΣΕΙ ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΟΥ ΜΟΝΤΕΛΟΥ ΜΕΘΟΔΟΙ

Μαρία Δ. Ανδρούτσου

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση

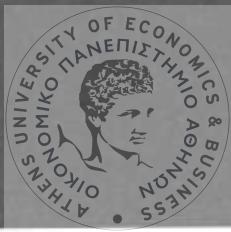
Μεταπτυχιακού Διπλώματος

Συμπληρωματικής Ειδίκευσης στη Στατιστική

Μερικής Παρακολούθησης (Part-time)

Αθήνα
Μάιος 2006

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΚΑΤΑΛΟΓΟΣ





ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΒΙΒΛΙΟΘΗΚΗ
εισ. 79875
Αρ.
ταξ.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ:
ΙΕΡΑΡΧΙΚΕΣ ΚΑΙ ΒΑΣΕΙ
ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΟΥ ΜΟΝΤΕΛΟΥ ΜΕΘΟΔΟΙ

Μαρία Δ. Ανδρούτσου

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος

Συμπληρωματικής Ειδίκευσης στη Στατιστική

Μερικής Παρακολούθησης (Part-time)



Αθήνα
Φεβρουάριος, 2006





ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Εργασία που υποβλήθηκε ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Συμπληρωματικής Ειδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time)

ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ: ΙΕΡΑΡΧΙΚΕΣ ΚΑΙ ΒΑΣΕΙ ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΟΥ ΜΟΝΤΕΛΟΥ ΜΕΘΟΔΟΙ

Μαρία Δ. Ανδρούτσου

Υπεύθυνο μέλος ΔΕΠ:
Ι. Παπαγεωργίου
Λέκτορας

Ο Διευθυντής Μεταπτυχιακών Σπουδών

Μιχαήλ Ζαζάνης
Καθηγητής





ΑΦΙΕΡΩΣΗ

Η διπλωματική εργασία είναι αφιερωμένη στην οικογένειά μου



ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ την κα Παπαγεωργίου Ιουλία (Λέκτορα Καθηγήτρια του Οικονομικού Πανεπιστημίου Αθηνών) για την συνεργασία της και την πολύτιμη βοήθεια της στη διεκπεραίωσης της διπλωματικής εργασίας.





ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Όνομάζομαι Ανδρούτσου Μαρία.

Γεννήθηκα στην Αθήνα στις 3/1/1965.

Το 1987 αποφοίτησα από το Τμήμα Φυσικής του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών.

Εργάστηκα για οκτώ χρόνια σε Φροντιστήρια Μέσης Εκπαίδευσης, για τη διδασκαλία Φυσικής.

Από το 1995 εργάζομαι στη Διεύθυνση Ανθρώπινου Δυναμικού της Εμπορικής Τράπεζας και από το 2000 είμαι Προϊσταμένη Υπηρεσίας.

Τον Οκτώβριο του 2001 εισάγομαι στο Μεταπτυχιακό Τμήμα Συμπληρωματικής Ειδίκευσης στη Στατιστική Μερικής Παρακολούθησης (Part-time), του Τμήματος Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών.

Είμαι έγγαμη και μητέρα δύο παιδιών 11 και 13 ετών.



BIOLOGY FOR ECONOMISTS



ABSTRACT

Maria Androutsou



A review of Cluster Analysis: Model-Based and non model-based techniques

February, 2006

Cluster analysis is a method that classifies points-objects into groups. In this way, objects that are more similar to each other belong to the same cluster, whereas the clusters differ. This thesis deals with two approaches for clustering: the hierarchical methods and the model-based method (mixture models). We describe the two methods analytically and then we perform them in two different data sets. The first one is about economic and demographic indicators of 25 countries (ECON data), whereas the second contains information about the duration of the Old Faithful geyser eruption and the waiting time before the next eruption of the geyser (Old Faithful data).



Επίκληση στην Επιτροπή για την παραγωγή της απόφασης
επενδύσεων στην ανάπτυξη της χώρας και την ανάπτυξη
της οικονομίας της χώρας. Η απόφαση αυτή είναι η πιο
πλήρης απόφαση που έχει γίνει στην Ελλάδα μέχρι σήμερα.
Είναι η πιο πλήρης απόφαση που έχει γίνει στην Ελλάδα μέχρι σήμερα.
Είναι η πιο πλήρης απόφαση που έχει γίνει στην Ελλάδα μέχρι σήμερα.
Είναι η πιο πλήρης απόφαση που έχει γίνει στην Ελλάδα μέχρι σήμερα.
Είναι η πιο πλήρης απόφαση που έχει γίνει στην Ελλάδα μέχρι σήμερα.
Είναι η πιο πλήρης απόφαση που έχει γίνει στην Ελλάδα μέχρι σήμερα.
Είναι η πιο πλήρης απόφαση που έχει γίνει στην Ελλάδα μέχρι σήμερα.
Είναι η πιο πλήρης απόφαση που έχει γίνει στην Ελλάδα μέχρι σήμερα.
Είναι η πιο πλήρης απόφαση που έχει γίνει στην Ελλάδα μέχρι σήμερα.

Επίκληση στην Επιτροπή για την παραγωγή της απόφασης



ΠΕΡΙΛΗΨΗ

Μαρία Ανδρούτσου

Ανάλυση Κατά Συστάδες: Ιεραρχικές και βάσει πιθανοθεωρητικού μοντέλου μέθοδοι

Φεβρουάριος, 2006.

Η ανάλυση κατά συστάδες είναι μια μέθοδος η οποία κατατάσσει, ταξινομεί στοιχεία-αντικείμενα σε ομάδες. Έτσι, αντικείμενα τα οποία είναι περισσότερο όμοια μεταξύ τους ανήκουν στην ίδια ομάδα, ενώ οι ομάδες διαφέρουν μεταξύ τους. Στη διατριβή αυτή θα ασχοληθούμε με δυο μεθόδους εύρεσης των ομάδων-συστάδων: τις ιεραρχικές μεθόδους και τη χρήση πιθανοθεωρητικού μοντέλου (μίξεις κατανομών). Οι δυο μέθοδοι περιγράφονται αναλυτικά και στη συνέχεια εφαρμόζονται σε δυο διαφορετικά σετ δεδομένων. Το ένα αφορά οικονομικούς- δημογραφικούς δείκτες 25 χωρών (ECON data), ενώ το άλλο αφορά τη διάρκεια των εκρήξεων και το χρόνο αναμονής μεταξύ δυο εκρήξεων του θερμοπίδακα Old Faithful (Old Faithful data).





ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

TABLE OF CONTENTS

ΕΙΣΑΓΩΓΗ

1.1	Εισαγωγή	1
1.2	Σύντομη Περιγραφή Διατριβής	2

ΜΕΤΡΑ ΑΠΟΣΤΑΣΗΣ ΚΑΙ ΟΜΟΙΟΤΗΤΑΣ

2.1	Εισαγωγή	5
2.2	Μέτρα Ομοιότητας	5
2.2.1	Μέτρα Απόστασης για Συνεχείς Μεταβλητές	6
2.2.2	Μέτρα Ομοιότητας για Δυαδικές Μεταβλητές	8
2.2.3	Μέτρα Ομοιότητας σε Μεικτού Τύπου Μεταβλητές	11
2.3	Μέτρα Ομοιότητας και Μέτρα Απόστασης Ανάμεσα στις Συστάδες	13

3 ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ ΓΙΑ ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ

3.1	Εισαγωγή	15
3.2	Αλγόριθμοι Συσσωμάτωσης	16
3.2.1	Αλγόριθμος Ατομικής Σύνδεσης	17
3.2.2	Αλγόριθμος Πλήρους Σύνδεσης (Complete Linkage)	20
3.2.3	Αλγόριθμος Group- Average	21
3.2.4	Αλγόριθμος Βασισμένος στα Κέντρα των Συστάδων (Centroid)	22
3.2.5	Αλγόριθμος Βασισμένος στο Διάνυσμα των Μέσων	23
3.2.6	Αλγόριθμος του Ward	24
3.3	Ο Επαναληπτικός Τύπος των Lance και William	25
3.4	Ιδιότητες και Προβλήματα των Ιεραρχικών Τεχνικών	26
3.5	Αλγόριθμοι Διαμέρισης	29
3.5.1	Ο Αλγόριθμος K-Means	29
3.6	Εξακριβώνοντας τη Λύση Στην Ανάλυση Κατά Συστάδες	32



3.7 Εξετάζοντας τον Βέλτιστο Αριθμό Ταξινομήσεων -	33
Συγκρίνοντας Διαφορετικές Ταξινομήσεις	

4 ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ ΜΕ ΧΡΗΣΗ ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

4.1 Εισαγωγή	37
4.2 Εισαγωγικές Έννοιες σε Μίξεις Κατανομών	38
4.3 Εκτίμηση με τη Μέθοδο Μεγίστης Πιθανοφάνειας	39
4.4. Ο EM Αλγόριθμος για Υπολογισμό Παραμέτρων Μίξης	41
4.5 ΕΜ Αλγόριθμος για Ανάλυση κατά Συστάδες	45
4.6 Επιτρέποντας τον Προσανατολισμό και το Μέγεθος να Μεταβάλλεται μεταξύ των Συστάδων	45
4.7 Εκτίμηση Κατάλληλου Αριθμού Συστάδων	47
4.7 Στρατηγική για Ανάλυση σε Συστάδες με Χρήση Μίξης Κανονικών Κατανομών	49
4.8 Μίξεις για Κατηγορικά Δεδομένα	50

5 ΕΦΑΡΜΟΓΗ ΑΝΑΛΥΣΗΣ ΚΑΤΑ ΣΥΣΤΑΔΕΣ ΣΕ ΔΥΟ ΣΕΤ ΔΕΔΟΜΕΝΩΝ – ΣΥΓΚΡΙΣΗ ΙΕΡΑΡΧΙΚΩΝ ΜΕΘΟΔΩΝ ΚΑΙ ΑΝΑΛΥΣΗΣ ΚΑΤΑ ΣΥΑΣΤΑΔΕΣ ΜΕ ΧΡΗΣΗ ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

5.1 Εισαγωγή	53
5.2 Econ Data	53
5.2.1 Ιεραρχικές Μέθοδοι	55
5.2.2 Ανάλυση Κατά Συστάδες με Χρήση Πιθανοθεωρητικού Μοντέλου	56
5.2.3 Σύγκριση των Δυο Διαφορετικών Προσεγγίσεων, Ιεραρχικών Μεθόδων και Πιθανοθεωρητικού Μοντέλου- Τελικά Συμπεράσματα	59
5.3 Old Faithful Data	59

5.3.1 Ανάλυση των Δεδομένων του Old Faithful με Ιεραρχικές Μεθόδους	61
5.3.2 Ανάλυση των Δεδομένων του Old Faithful με Πιθανοθεωρητικό Μοντέλο	62

ΠΑΡΑΡΤΗΜΑ I

S-Plus Λογισμικό	65
------------------	----



ΠΑΡΑΡΤΗΜΑ II

OLD FAITHFUL ΔΕΔΟΜΕΝΑ	67
-----------------------	----

ΒΙΒΛΙΟΓΡΑΦΙΑ	71
---------------------	-----------



ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 2.1: Απεικόνιση δυο ατόμων για δυαδικά δεδομένα	9
Πίνακας 2.2: Συντελεστές ομοιότητας και ανομοιότητας για δυαδικά δεδομένα	9
Πίνακας 3.1: Ο Αλγόριθμος K-Means (σημειώσεις Καρλή)	32
Πίνακας 4.1: Ο αλγόριθμος EM για ανάλυση κατά συστάδες στην περίπτωση μίξης κανονικών κατανομών.	43
Πίνακας 4.2: Παραμετροποίηση του πίνακα συνδιακύμανσης Σ_k στο Γκαουσιανό μοντέλο κι η γεωμετρική ερμηνεία του (Banfield J.D. και Raftery A.E. (1993)).	47
Πίνακας 5.1: Econ Data	54
Πίνακας 5.2: Ιεραρχική ομαδοποίηση σε 3 ομάδες για τα δεδομένα των 25 χωρών	57
Πίνακας 5.3: Αποτελέσματα της ανάλυσης κατά συστάδες στα οικονομικά- δημογραφικά δεδομένα με τη χρήση πιθανοθεωρητικού μοντέλου.	60



ΚΑΤΑΛΟΓΟΣ ΓΡΑΦΗΜΑΤΩΝ

<u>Γράφημα 3.1:</u> Απεικόνιση δενδρογράμματος για αλγορίθμους συσσωμάτωσης και διαιρετότητας.	16
<u>Γράφημα 3.2:</u> Δενδρόγραμμα για τον αλγόριθμο ατομικής σύνδεσης για τον πίνακα αποστάσεων D_1	19
<u>Γράφημα 3.3:</u> Δενδρόγραμμα για τον αλγόριθμο πλήρους σύνδεσης για τον πίνακα αποστάσεων D_1	20
<u>Γράφημα 3.4:</u> Απεικόνιση του αλγόριθμου Group-Average	21
<u>Γράφημα 3.5:</u> Δενδρόγραμμα του αλγορίθμου centroid για τον πίνακα αποστάσεων D_1	24
<u>Γράφημα 3.6:</u> Απεικόνιση του προβλήματος chaining	27
<u>Γράφημα 3.7:</u> Παράδειγμα δεδομένων με δυο ομάδες σε σχήμα ομόκεντρων κύκλων	28
<u>Γράφημα 3.8:</u> Εφαρμογή του αλγορίθμου K-Means	30
<u>Γράφημα 3.9:</u> Εναισθησία του αλγορίθμου K-Means στην επιλογή των αρχικών κέντρων	31
<u>Γράφημα 4.1:</u> Ανάλυση κατά συστάδες με χρήση πιθανοθεωρητικού μοντέλου.	38
<u>Γράφημα 5.1</u> Δενδρογράμματα των μεθόδων ιεραρχικής ομαδοποίησης	58
<u>Γράφημα 5.2</u> Δενδρογράμματα των μεθόδων με τη χρήση πιθανοθεωρητικού μοντέλου	61
<u>Γράφημα 5.3</u> Χρήση ιεραρχικών μεθόδων στα δεδομένα Old Faithful	63
<u>Γράφημα 5.4</u> Χρήση πιθανοθεωρητικών μοντέλων στα δεδομένα Old Faithful	64



ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγή

Η ανάλυση κατά συστάδες είναι μια μέθοδος η οποία κατατάσσει, ταξινομεί στοιχεία-αντικείμενα σε ομάδες. Έτσι, αντικείμενα τα οποία είναι περισσότερο όμοια μεταξύ τους ανήκουν στην ίδια ομάδα, ενώ οι ομάδες διαφέρουν μεταξύ τους. Η ιδέα της ταξινόμησης έχει αποδειχθεί σημαντική και χρησιμοποιείται ευρέως σε πολλές επιστήμες (βιολογία, ψυχολογία, κοινωνικές και οικονομικές επιστήμες, κλπ.). Για παράδειγμα, στη βιολογία η ταξινόμηση των οργανισμών (taxonomy) ήταν η κύρια μέριμνα από τις πρώτες μελέτες των βιολόγων.

Γενικότερα, ένα σχήμα ταξινόμησης μπορεί να αντιπροσωπεύει απλά μια πρόσφορη μέθοδο για να οργανώσει κάποιος ένα τεράστιο σετ δεδομένων έτσι ώστε η εξαγωγή πληροφορίας να γίνεται πιο εύκολα. Με αυτό τον τρόπο πετυχαίνει κανείς μια σύνοψη των δεδομένων, της πληροφορίας που έχει.

Η ιδέα γενικότερα είναι η διαμέριση n αντικειμένων σε ομάδες όπου κάθε ομάδα περιλαμβάνει ένα μόνο αντικείμενο. Τα βασικά δεδομένα στα οποία πραγματοποιείται η μέθοδος της ανάλυσης κατά συστάδες συνήθως παρίστανται με έναν πίνακα X , ο οποίος περιέχει τις τιμές των μεταβλητών,

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Ο σκοπός της ανάλυσης κατά συστάδες είναι να ομαδοποιήσει τα άτομα ή αντικείμενα τα οποία παρίστανται στις n γραμμές του πίνακα X . Φυσικά κάποιος θα μπορούσε να πραγματοποιήσει και ομαδοποίηση στον πίνακα X' , να σχηματίσει ομάδες (συστάδες) μεταξύ των μεταβλητών. Πολλές φορές, μάλιστα, η διαχώριση «αντικειμένου» και «μεταβλητής» δεν είναι και τόσο ξεκάθαρη. Παρόλα αυτά, στο υπόλοιπο της διατριβής θα ασχοληθούμε με την ομαδοποίηση των γραμμών του πίνακα X . Επίσης θα ασχοληθούμε με την ταξινόμηση αντικειμένων τα οποία πριν δεν έχουν ομαδοποιηθεί, δηλαδή θα

θεωρήσουμε ότι ο αριθμός και η σύνθεση των ομάδων δεν μας είναι γνωστός εκ των προτέρων.

1.2 Σύντομη Περιγραφή Διατριβής

Η ανάλυση κατά συστάδες ουσιαστικά χρησιμοποιεί την πληροφορία που υπάρχει σε κάποιες μεταβλητές για να κατατάξει σε ομάδες τα αντικείμενα ενός σετ δεδομένων. Εφόσον η κατάταξη δυο αντικειμένων βασίζεται στο κατά πόσο όμοια ή διαφορετικά είναι θα πρέπει να υπάρχει κάποιος τρόπος για να μπορούμε να μετρήσουμε αυτή την ομοιότητα (ανομοιότητα). Έτσι, στο Κεφάλαιο 2 θα αναφέρουμε μέτρα απόστασης και ομοιότητας για συνεχείς, κατηγορικές, μεταβλητές με κατάταξη κλίμακας και μεικτές μεταβλητές.

Στο Κεφάλαιο 3 θα αναφερθούμε στις ιεραρχικές μεθόδους ταξινόμησης. Στις μεθόδους αυτές η διαμέριση, έως ότου προκύψουν οι τελικές συστάδες, πραγματοποιείται είτε με μια σειρά διαμερισμών, ξεκινώντας με μια συστάδα, η οποία περιέχει όλα τα δεδομένα και σταδιακά διασπάμε (split), είτε με κάθε άτομο να απαρτίζει μόνο του μια συστάδα και σταδιακά ενώνουμε (merge). Οι μέθοδοι- αλγόριθμοι τους οποίους θα περιγράψουμε είναι ο αλγόριθμος ατομικής σύνδεσης, ο πλήρους σύνδεσης, ο group-average, ο centroid, η μέθοδος του Ward και ο αλγόριθμος που βασίζεται στο διάνυσμα των μέσων.

Στο Κεφάλαιο 4 θα ασχοληθούμε με στην ανάλυση κατά συστάδες με τη χρήση πιθανοθεωρητικού μοντέλου και συγκεκριμένα με τη μέθοδο που χρησιμοποιεί μίξεις κατανομών. Οι μέθοδοι, οι οποίες αναπτύσσονται στο Κεφάλαιο 3, δεν βασίζονται σε κάποιο μοντέλο, αλλά σε καθαρά μαθηματικές τεχνικές χωρίς να λαμβάνεται υπόψη η μεταβλητότητα που μπορεί να παίζει ρόλο στα αποτελέσματα. Η βασική διαφορά του πιθανοθεωρητικού μοντέλου είναι ότι κάθε στοιχείο δεν ανήκει σε μια και μοναδική συστάδα, αλλά σε κάθε μια με συγκεκριμένη πιθανότητα.

Τέλος, στο Κεφάλαιο 5 θα εφαρμόσουμε τις μεθόδους- τεχνικές τις οποίες αναπτύξαμε στα Κεφάλαια 3 και 4 σε δυο σετ δεδομένων. Το ένα αφορά σε οικονομικούς- δημογραφικούς δείκτες 25 χωρών. Σκοπός μας είναι

η ομαδοποίηση των χωρών βασιζόμενοι σε αυτούς τους δείκτες. Το δεύτερο σετ δεδομένων αφορά τη διάρκεια των εκρήξεων και το χρόνο αναμονής μεταξύ δυο εκρήξεων του θερμοπίδακα Old Faithful.



ΚΕΦΑΛΑΙΟ 2

ΜΕΤΡΑ ΑΠΟΣΤΑΣΗΣ ΚΑΙ ΟΜΟΙΟΤΗΤΑΣ

2.1 Εισαγωγή

Όπως αναφέραμε σκοπός της ανάλυσης κατά συστάδες είναι η κατάταξη σε ομάδες (συστάδες) ενός σετ δεδομένων χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Αυτό που ουσιαστικά επιδιώκουμε είναι οι παρατηρήσεις που απαρτίζουν μια ομάδα να έχουν όσο το δυνατόν περισσότερα κοινά χαρακτηριστικά, ενώ μεταξύ τους οι ομάδες να είναι όσο το δυνατόν περισσότερο ανομοιογενείς. Οπότε εύκολα διαπιστώνει κανείς ότι δύο από τις πιο σημαντικές έννοιες στην ανάλυση κατά συστάδες είναι οι έννοιες της απόστασης και ομοιότητας. Σκοπός αυτών των εννοιών είναι να μας βοηθήσουν να μετρήσουμε κατά πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους και επομένως να τις κατατάξουμε στην ίδια ομάδα.

Στο κεφάλαιο αυτό θα αναφέρουμε μέτρα απόστασης και ομοιότητας για συνεχείς (Everitt B.S. (1993), Σημειώσεις Καρλή), κατηγορικές, μεταβλητές με κατάταξη κλίμακας (Everitt B.S. (1993), Sneath και Socal (1973)) και μεικτές μεταβλητές (Gower (1971), Jardine και Sibson (1971)). Τέλος θα αναφερθούμε και σε μέτρα απόστασης όχι μόνο ανάμεσα στις παρατηρήσεις που ανήκουν στην ίδια ομάδα, αλλά και μεταξύ των ομάδων (Everitt B.S. (1993), Σημειώσεις Καρλή).

2.2 Μέτρα Ομοιότητας

Ο συντελεστής ομοιότητας δείχνει το μέγεθος της σχέσης μεταξύ δύο παρατηρήσεων δεδομένης της ύπαρξης p μεταβλητών κοινών και στα δύο αντικείμενα. Η ομοιότητα ανάμεσα σε δύο αντικείμενα, έστω i και j , είναι μια συνάρτηση της μορφής

$$s_{ij} = f(\mathbf{x}_i, \mathbf{x}_j) \quad (2.1)$$

όπου $\mathbf{x}'_i = \{x_{i1}, \dots, x_{ip}\}$ και $\mathbf{x}'_j = \{x_{j1}, \dots, x_{jp}\}$ είναι οι τιμές των p μεταβλητών για τα δύο αντικείμενα. Πολλές συναρτήσεις έχουν προταθεί κατά καιρούς ανάλογα εν μέρει με το είδος της μεταβλητής, αλλά και με το είδος του αντικειμένου.

Τα περισσότερα μέτρα ομοιότητας είναι σχεδιασμένα ώστε να ισχύει η συμμετρία $s_{ij} = s_{ji}$ (Constantine και Gower (1978)). Οι περισσότεροι συντελεστές είναι μη αρνητικοί και έχουν σαν μέγιστη τιμή τη μονάδα. Παρόλα αυτά υπάρχουν και μέτρα για τα οποία ισχύει $-1 \leq s_{ij} \leq 1$.

Τα μέτρα ομοιότητας συνδέονται άμεσα με τα μέτρα ανομοιότητας. Η σχέση των δύο μέτρων, για εκείνα τα μέτρα ομοιότητας που είναι φραγμένα στο μηδέν και έχουν μέγιστη τιμή τη μονάδα, είναι $d_{ij} = 1 - s_{ij}$. Τα μέτρα ανομοιότητας είναι και αυτά συμμετρικά και μη αρνητικά. Ο βαθμός ομοιότητας ανάμεσα σε δύο αντικείμενα αυξάνει με το s_{ij} και μειώνεται καθώς αυξάνει το d_{ij} . Είναι λογικό, λοιπόν, ένα αντικείμενο να έχει $s_{ii} = 1$ και $d_{ii} = 0$ με τον εαυτό του.

2.2.1 Μέτρα Απόστασης για Συνεχείς Μεταβλητές

Τα συνεχή δεδομένα είναι η πιο απλή περίπτωση μιας και αρκετά μέτρα απόστασης μπορούν να χρησιμοποιηθούν (τα μέτρα απόστασης συχνά συνδέονται με τα μέτρα ανομοιότητας μιας και όσο μικρότερη η απόσταση ανάμεσα σε δύο άτομα τόσο μεγαλύτερη η ομοιότητά τους). Εδώ θα περιγράψουμε μερικά από τα πιο δημοφιλή (Everitt B.S. (1993), Σημειώσεις Καρλή).

Ευκλείδεια Απόσταση:

Έστω δυο διανύσματα \mathbf{x}, \mathbf{y} . Τότε η ευκλείδεια απόσταση ορίζεται ως

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2} \quad (2.2)$$

Η απόσταση αυτή είναι η πιο απλή και γνωστή απόσταση. Παρόλα αυτά έχει αρκετά μειονεκτήματα. Τα πιο σημαντικά είναι ότι εξαρτάται από την κλίμακα μέτρησης των μεταβλητών και επίσης μεταβλητές με μεγάλες

απόλυτες τιμές έχουν πολύ μεγαλύτερο βάρος. Επίσης δεν έχει κάποιο στατιστικό υπόβαθρο, αγνοεί τη μεταβλητότητα των μεταβλητών και επιπλέον ακραίες παρατηρήσεις έχουν πολύ μεγάλη επίδραση.

City-block (Manhattan) Απόσταση:

Η απόσταση αυτή διαφέρει από την ευκλείδεια στο ότι δεν χρησιμοποιεί τις τετραγωνικές αποκλίσεις, αλλά τις απόλυτες αποκλίσεις, δηλαδή

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^p |x_i - y_i| \quad (2.3)$$

Το μέτρο αυτό, λόγω της ομοιότητας του με το προηγούμενο εμφανίζει τα ίδια μειονεκτήματα με αυτό, εκτός του ότι είναι πιο ανθεκτικό στις ακραίες παρατηρήσεις επειδή τους δίνει μικρότερο βάρος.

Απόσταση Minkowski (L_qnorm):

Η απόσταση Minkowski γενικεύει τις δύο προηγούμενες αποστάσεις και ορίζεται ως

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^p |x_i - y_i|^q \right]^{1/q} \quad (2.4)$$

Η τιμή της παραμέτρου q καθορίζεται κάθε φορά από τον ερευνητή και δίνει ιδιαίτερο βάρος σε κάποιες παρατηρήσεις. Είναι προφανές ότι για $q=1$ προκύπτει η απόσταση Manhattan, ενώ για $q=2$ η ευκλείδεια. Γενίκευση αποτελεί η Power Distance με τύπο

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^p |x_i - y_i|^r \right]^{1/r} \quad (2.5)$$

όπου τα q, r ορίζονται από τον ερευνητή.

Απόσταση Chebyshev:

Η απόσταση αυτή δεν χρησιμοποιεί όλες τις αποκλίσεις, αλλά μόνο τις μεγαλύτερες από αυτές, δηλαδή

$$d(\mathbf{x}, \mathbf{y}) = \max \{(x_i - y_i), i=1, \dots, p\} \quad (2.6)$$

Η πιο πάνω απόσταση εξαρτάται από τις διαφορές στην κλίμακα και επομένως αν οι κλίμακες είναι διαφορετικές ουσιαστικά θα αντικατοπτρίζει τη διαφορά στη μεταβλητή με τη μεγαλύτερη κλίμακα

Απόσταση Mahalanobis:

Όλες οι παραπάνω αποστάσεις έχουν το μειονέκτημα ότι δεν λαμβάνουν υπόψη τους τις συσχετίσεις μεταξύ των μεταβλητών. Ένα μέτρο το οποίο βασίζεται σε στατιστικές έννοιες και λαμβάνει υπόψη του διακυμάνσεις και συνδιακυμάνσεις είναι η απόσταση Mahalanobis

$$d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})' \mathbf{S}^{-1} (\mathbf{x} - \mathbf{y}) \quad (2.7)$$

όπου \mathbf{S} ο δειγματικός πίνακας συνδιακύμανσης.

Τέλος, θα μπορούσαμε να χρησιμοποιήσουμε σαν μέτρο το γνωστό μας συντελεστή συσχέτισης με τη διαφορά ότι δεν αθροίζουμε ως προς όλες τις μεταβλητές, αλλά ως προς τις παρατηρήσεις. Πρέπει όμως να τονίσουμε ότι ο συντελεστής συσχέτισης ως απόσταση δεν ικανοποιεί τις ιδιότητες της απόστασης και η έννοια της συσχέτισης είναι διαφορετική από την έννοια της ομοιότητας. Για παράδειγμα αν δύο παρατηρήσεις διαφέρουν κατά 1 μονάδα σε κάθε μεταβλητή τότε η τιμή του συντελεστή είναι η μεγαλύτερη δυνατή και ίση με 1.

Οι αποστάσεις που περιγράψαμε μέχρι τώρα έχουν το μειονέκτημα της εξάρτησής τους από την κλίμακα των μεταβλητών. Ένας τρόπος για να επιλυθεί το πρόβλημα αυτό είναι η τυποποίηση των μεταβλητών. Το μειονέκτημα της συγκεκριμένης μεθόδου είναι ότι με αυτόν τον τρόπο αγνοούνται οι τυχών συσχετίσεις μεταξύ των μεταβλητών.

2.2.2 Μέτρα Ομοιότητας για Δυαδικές Μεταβλητές

Οι απλούστεροι και πιο συνηθισμένοι συντελεστές ομοιότητας είναι αυτοί για δίτιμες μεταβλητές όπου κάθε μεταβλητή έχει μόνο δύο τιμές. Τέτοια δεδομένα για δύο αντικείμενα, ας πούμε i και j συχνά απεικονίζονται με έναν 2×2 πίνακα (Πίνακας 2.1). Εδώ $p=a+b+c+d$. Πρέπει όμως να σημειώσουμε ότι η πιο πάνω απεικόνιση δεν πρέπει να συγχέεται με τον 2×2 πίνακα συνάφειας.

Πολλοί συντελεστές ομοιότητας έχουν προταθεί. Στον Πίνακα 2.1 δίνουμε τους πιο δημοφιλείς. Επίσης αναγράφουμε και τους συντελεστές ανομοιότητας, οι οποίοι απλά προκύπτουν από τη σχέση (1 -συντελεστής ομοιότητας). Από τους συντελεστές αυτούς οι πιο κοινοί στην χρήση είναι οι

συντελεστής ταιριάσματος (matching coefficient) και ο συντελεστής του Jaccard. Ο πρώτος είναι απλά ο λόγος του ολικού αριθμού μεταβλητών που είναι ίδιοι για τα δύο άτομα ως προς το συνολικό αριθμό των μεταβλητών, ενώ ο δεύτερος είναι ο αντίστοιχος λόγος όταν τα «αρνητικά» ταιριάσματα (d) αγνοούνται. Το πρόβλημα του αν πρέπει να συμπεριλαμβάνονται τα αρνητικά ταιριάσματα ή όχι αφορά μόνο την περίπτωση που η μεταβλητή είναι του τύπου παρουσία ή απουσία ενός χαρακτηριστικού.

Πίνακας 2.1 Απεικόνιση δυο ατόμων για δυαδικά δεδομένα

		Άτομο i	
		1	2
Άτομο j	1	a	b
	2	c	d
Σύνολο		$a+c$	$b+d$
		p	

Πίνακας 2.2: Συντελεστές ομοιότητας και ανομοιότητας για δυαδικά δεδομένα

		$s(x, y)$	$d(x, y)$
Αρνητικά και αρνητικά ταιριάσματα	Matching Coefficient	$\frac{a+d}{p}$	$\frac{b+c}{p}$
	Rogers και Tarimoto (1960)	$\frac{a+d}{(a+d)+2(b+c)}$	$\frac{2(b+c)}{(a+d)+2(b+c)}$
	Socal και Sneath (1963)	$\frac{2(a+d)}{2(a+d)+(b+c)}$	$\frac{(b+c)}{2(a+d)+(b+c)}$
Μόνο θετικά ταιριάσματα	Jaccard (1908)	$\frac{a}{a+b+c}$	$\frac{b+c}{a+b+c}$
	Dice (1945), Sorensen (1948)	$\frac{2a}{2a+b+c}$	$\frac{b+c}{2a+b+c}$
	Sokal και Sneath (1963)	$\frac{a}{a+2(b+c)}$	$\frac{2(b+c)}{a+2(b+c)}$

Οι συντελεστές ομοιότητας του Πίνακα 2.2 μπορεί να έχουν πολύ διαφορετικές τιμές για το ίδιο σετ δεδομένων. Ας υποθέσουμε, για

παράδειγμα, δύο άτομα που έχουν τα παρακάτω σκορ σε δέκα δυαδικές μεταβλητές

	Μεταβλητή									
	1	2	3	4	5	6	7	8	9	10
Άτομο 1	1	0	0	0	1	1	0	0	1	0
Άτομο2	0	0	0	0	1	0	0	1	1	0

Τότε ο αντίστοιχος 2×2 πίνακας είναι

		Άτομο 1		
		1	0	
Άτομο 2	1	1	2	1
	0	0	2	5
Σύνολο		4	6	10

Στο παράδειγμα αυτό ο συντελεστής matching ισούται με 0.70, ο συντελεστής του Jaccard με 0.40, των Dice - Sorensen με 0.57, των Socal και Sneath (αρνητικά και θετικά ταιριάσματα) με 0.82 και των Socal και Sneath μόνο με αρνητικά ταιριάσματα με 0.25. Αν οι παραπάνω συντελεστές έπαιρναν διαφορετικές τιμές για το ίδιο ζεύγος ατόμων με τρόπο από κοινού μονοτονικό (δηλαδή για όλα τα διαφορετικά ζεύγη ατόμων είτε αυξάνονταν είτε μειώνονταν όλοι) δεν θα υπήρχε ιδιαίτερο πρόβλημα. Κάτι τέτοιο όμως δεν ισχύει. Ας θεωρήσουμε και ένα τρίτο άτομο με τις πιο κάτω τιμές

	Μεταβλητή									
	1	2	3	4	5	6	7	8	9	10
Άτομο 3	0	0	0	0	0	0	0	1	0	0

Οι τιμές για τους συντελεστές matching και Jaccard είναι οι παρακάτω:

Συντελεστής Matching	Συντελεστής Jaccard
$s_{12} = 0.70$	$s_{12} = 0.40$
$s_{13} = 0.50$	$s_{13} = 0.00$
$s_{23} = 0.80$	$s_{23} = 0.33$



Όπως παρατηρούμε οι συντελεστές δεν είναι από κοινού μονοτονικοί.

Κατηγορικά δεδομένα με περισσότερες από δυο κατηγορίες (δεδομένα σε ονομαστική κλίμακα), για παράδειγμα χρώμα ματιών, μπορούν να αντιμετωπιστούν με τρόπο παρόμοιο των δυαδικών, μετατρέποντας τους σε μια σειρά δυαδικών μεταβλητών, π.χ. μια μεταβλητή με τιμές 1 αν τα άτομα έχουν γαλάζια μάτια 0 διαφορετικά, μια άλλη με τιμές 1 για τα άτομα με καστανά μάτια ο διαφορετικά, κλπ. Αυτός ο τρόπος, όμως, δεν είναι ιδιαίτερα ελκυστικός και αυτό διότι τα αρνητικά ταιριάσματα (0) είναι μοιραία παρόντα. Μια καλύτερη μέθοδος είναι να προσδιοριστεί ένα σκορ s_{ijk} με τιμές 1 ή 0, σε κάθε μεταβλητή k , ανάλογα με το αν τα δυο άτομα i και j έχουν την ίδια τιμή στην μεταβλητή. Τα σκορ για όλα τα άτομα αθροίζονται και τελικά προκύπτει οι συντελεστές ομοιότητας και ανομοιότητας αντίστοιχα

$$s_{ij} = \frac{\sum_{k=1}^p s_{ijk}}{p} \quad d_{ij} = \frac{p - \sum_{k=1}^p s_{ijk}}{p} \quad (2.8)$$

Στην περίπτωση κατηγορικών μεταβλητών σε κλίμακα διάταξης συνηθίζεται να τις θεωρούμε συνεχείς και χρησιμοποιούμε μια κατάλληλη απόσταση. Σε αυτές τις περιπτώσεις χρειάζεται προσοχή ώστε να χρησιμοποιείται η ίδια κλίμακα. Εναλλακτικός τρόπος είναι να μετασχηματιστούν σε δίτιμες.

2.2.3 Μέτρα Ομοιότητας σε Μεικτού Τύπου Μεταβλητές

Τα περισσότερα προβλήματα στην ανάλυση κατά συστάδες αφορούν σετ δεδομένων τα οποία δεν περιέχουν μόνο συνεχείς ή μόνο κατηγορικές μεταβλητές, αλλά αποτελούνται από συνεχείς και κατηγορικές (δίτιμες, σε ονομαστική κλίμακα ή κλίμακα κατάταξης). Στην περίπτωση αυτή πρέπει να υπάρχει ένα μέτρο ομοιότητας, το οποίο να εκφράζει την ομοιότητα ανάμεσα στα ζεύγη ατόμων συνυπολογίζοντας κάθε τύπο μεταβλητής. Ένας τέτοιος συντελεστής προτάθηκε από τον Gower (1971) και ορίζεται ως

$$s_{ij} = \frac{\sum_{k=1}^p w_{ijk} s_{ijk}}{\sum_{k=1}^p w_{ijk}} \quad (2.9)$$



Στον τύπο (2.9) το s_{ijk} είναι η ομοιότητα ανάμεσα στα i και j άτομα όπως μετρούνται στην μεταβλητή k και w_{ijk} είναι 1 ή 0 ανάλογα από το αν η σύγκριση θεωρείται απαραίτητη για την μεταβλητή k . Βάρη ίσα με το 0 θεωρούνται όταν η μεταβλητή k είναι άγνωστη για το ένα ή και τα δύο άτομα ή σε δίτιμες μεταβλητές όταν απαιτείται να εξαιρεθούν τα αρνητικά ταιριάσματα. Για τα κατηγορικά δεδομένα οι συνιστώσες ομοιότητας s_{ijk} παίρνουν την τιμή 1 όταν και τα δύο άτομα έχουν την ίδια τιμή και 0 διαφορετικά. Για ποσοτικές μεταβλητές η ομοιότητα υπολογίζεται από τον τύπο

$$s_{ijk} = 1 - \left| x_{ik} - x_{jk} \right| / R_k \quad (2.10)$$

όπου x_{ik} και x_{jk} είναι οι τιμές για τις δύο μεταβλητές για την μεταβλητή k και R_k είναι το εύρος της μεταβλητής K , συνήθως υπολογισμένο για τον αριθμό των ατόμων που απαρτίζουν τη συστάδα.

Για να γίνει περισσότερο κατανοητή η χρήση του συντελεστή ας θεωρήσουμε το πιο κάτω παράδειγμα πέντε ψυχικά νοσούντων ατόμων.

	Βάρος (pounds)	Επίπεδο Άγχους	Κατάθλιψη	Παραισθήσεις	Ομάδα Ηλικίας
Ασθενής 1	120	Ήπιο	Όχι	Όχι	Νεαρή
Ασθενής 2	150	Μέτριο	Ναι	Όχι	Μεσαία
Ασθενής 3	110	Ισχυρό	Ναι	Ναι	Ηλικιωμένη
Ασθενής 4	145	Ήπιο	Όχι	Ναι	Ηλικιωμένη
Ασθενής 5	120	Ήπιο	Όχι	Ναι	Νεαρή

Επίσης ας θεωρήσουμε ότι ο ερευνητής επιθυμεί να εξαιρέσει τα αρνητικά ταιριάσματα στην μεταβλητή της κατάθλιψης και των παραισθήσεων από τον υπολογισμό του συντελεστή ομοιότητας μεταξύ δύο ατόμων. Ο συντελεστής του Gower για το ζεύγος ατόμων 1 και 2 ισούται με

$$s_{12} = \frac{1 \times \left(1 - \frac{30}{40} \right) + 1 \times 0 + 1 \times 0 + 0 \times 1 + 1 \times 0}{1 + 1 + 1 + 0 + 1} = 0.0625$$

Οι τιμές για όλα τα ζεύγη των ασθενών παρουσιάζονται στον πίνακα ομοιότητας S

$$\mathbf{S} = \begin{pmatrix} 1 & & & & \\ 0.062 & 1 & & & \\ 0.150 & 0.200 & 1 & & \\ 0.344 & 0.175 & 0.425 & 1 & \\ 0.750 & 0.005 & 0.350 & 0.475 & 1 \end{pmatrix}$$

2.3 Μέτρα Ομοιότητας και Μέτρα Απόστασης Ανάμεσα στις Συστάδες

Στις προηγούμενες παραγράφους οι μέθοδοι που περιγράψαμε αφορούσαν ομοιότητα και αποστάσεις για τα άτομα εντός των συστάδων. Στις εφαρμογές της ανάλυσης κατά συστάδες είναι συχνά απαραίτητο να οριστούν μέτρα για τη σχέση μεταξύ των συστάδων. Προβλήματα που απαντώνται περιλαμβάνουν

1. Την επιλογή ενός στατιστικού μέτρου για κάθε μεταβλητή. Λογικές επιλογές θα μπορούσαν να είναι αναλογίες για ποιοτικές μεταβλητές και μέσες τιμές για ποσοτικές μεταβλητές
2. Μέτρα για την εντός των ομάδων μεταβλητότητα
3. Κατασκευή ενός μέτρου ομοιότητας ή απόστασης βασισμένο στο 1. και που πιθανόν να επιτρέπει το 2. Επιτρέποντας όμως τον υπολογισμό της εντός των ομάδων μεταβλητότητας μπορεί να είναι ιδιαίτερα περίπλοκο αν αυτή δεν είναι σταθερή από την μια ομάδα στην άλλη.

Μια προφανή μέθοδος για κατασκευή των μεταξύ των ομάδων μέτρων απόστασης είναι η αντικατάσταση των μέσων των ομάδων για την μεταβλητή p στον τύπο της Ευκλείδειας ή κάποιας άλλης από τις αποστάσεις που αναφέραμε πιο πάνω. Έτσι, αν για παράδειγμα το διάνυσμα των μέσων για την ομάδα A είναι $\bar{\mathbf{x}}'_A = [\bar{x}_{A1}, \bar{x}_{A2}, \dots, \bar{x}_{Ap}]$ και την ομάδα B είναι $\bar{\mathbf{x}}'_B = [\bar{x}_{B1}, \bar{x}_{B2}, \dots, \bar{x}_{Bp}]$ τότε η Ευκλείδεια απόσταση ορίζεται ως

$$d_{AB} = \sqrt{\sum_{i=1}^p (\bar{x}_{Ai} - \bar{x}_{Bi})^2} \quad (2.11)$$

Επίσης μπορεί να χρησιμοποιηθεί και η απόσταση Mahalanobis

$$D^2 = (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)' \mathbf{W}^{-1} (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B) \quad (2.12)$$

όπου \mathbf{W} είναι ένας $p \times p$ πίνακας των της διασποράς εντός των ομάδας για τις δύο ομάδες. Όταν οι συσχετίσεις ανάμεσα στις μεταβλητές είναι μικρές το D^2 είναι ίδιο με την τετραγωνισμένη Ευκλείδεια απόσταση υπολογισμένη στα τυποποιημένα δεδομένα.

Η χρήση του D^2 υποννοεί ότι ο ερευνητής είναι πρόθυμος να υποθέσει ότι οι διασπορές των μεταβλητών είναι τουλάχιστον προσεγγιστικά ίδιες και στις δύο ομάδες. Όταν αυτό δεν ισχύει, το D^2 είναι ακατάλληλο μέτρο και σε αυτές τις περιπτώσεις χρησιμοποιείται το μέτρο των Jardine και Sibson (1971)

$$R_{AB} = \log \left[\frac{|1/2(\mathbf{W}_B)|}{\sqrt{|\mathbf{W}_A||\mathbf{W}_B|}} \right] + \frac{1}{2} \log_2 \left(1 + \frac{1}{4} D_{AB}^2 \right) \quad (2.13)$$

όπου

$$D_{AB}^2 = (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)' \left[\frac{1}{2} (\mathbf{W}_A + \mathbf{W}_B) \right]^{-1} (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B) \quad (2.14)$$

Όταν $\mathbf{W}_A = \mathbf{W}_B$ προκύπτει η απόσταση Mahalanobis.

Υπάρχουν αρκετά ακόμη μέτρα για την μεταξύ των ομάδων απόσταση τα οποία και θα εξετάσουμε στο επόμενο κεφάλαιο όπου θα αναφέρουμε τις ιεραρχικές μεθόδους για την εύρεση των ομάδων.

ΚΕΦΑΛΑΙΟ 3

ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ ΓΙΑ ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ

3.1 Εισαγωγή

Στις ιεραρχικές μεθόδους ταξινόμησης η διαμέριση των δεδομένων δεν πραγματοποιείται σε ένα μοναδικό βήμα. Αντίθετα, η διαμέριση, έως ότου προκύψουν οι τελικές συστάδες, πραγματοποιείται με μια σειρά διαμερισμάν, οι οποίες μπορεί να προκύπτουν είτε ξεκινώντας με μια συστάδα και σταδιακά διασπάμε (split), η οποία περιέχει όλα τα δεδομένα, είτε με κάθε άτομο να απαρτίζει μόνο του μια συστάδα και σταδιακά ενώνουμε (merge).

Οι ιεραρχικές μέθοδοι χωρίζονται σε δυο κατηγορίες: στις μεθόδους συσσωμάτωσης (agglomerative), οι οποίες ξεκινούν με κάθε άτομο (στοιχείο) να απαρτίζει μια ομάδα και σε κάθε βήμα τα άτομα ενώνονται δημιουργώντας λιγότερες ομάδες, και στις μεθόδους διαιρετότητας (divisible), όπου όλα τα δεδομένα αποτελούν μια συστάδα και σε κάθε βήμα αυτή σπάει σε μικρότερες ομάδες.

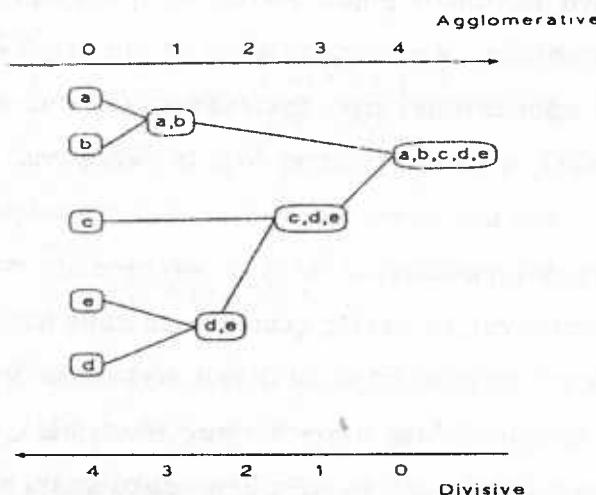
Και οι δυο μέθοδοι έχουν ως κοινό ότι από τη στιγμή που μια ένωση ή μια διαίρεση έχει συμβεί, αυτή είναι αμετάκλητη. Έτσι, σε έναν αλγόριθμο συσσωμάτωσης από τη στιγμή που δυο άτομα ενώνονται δεν μπορούν σε κάποιο άλλο βήμα να διαχωριστούν ξανά. Επίσης εφόσον ένας αλγόριθμος συσσωμάτωσης τελικά θα ενώσει όλα τα άτομα σε μια συστάδα και ένας αλγόριθμος διαιρετότητας εν τέλει θα δημιουργήσει τόσες συστάδες όσα είναι και τα άτομα στο σετ δεδομένων, ο ερευνητής καλείται να βρει εκείνη τη λύση που υποδεικνύει τον βέλτιστο αριθμό συστάδων Οι ιεραρχικές ταξινομήσεις συχνά αναπαρίστανται με ένα διάγραμμα δύο διαστάσεων γνωστό ως δενδρόγραμμα, το οποίο παριστά τους διαχωρισμούς ή ενώσεις των δεδομένων σε κάθε βήμα του αλγόριθμου. Στο Γράφημα 3.1 απεικονίζεται ένα δενδρόγραμμα.

Στο κεφάλαιο που ακολουθεί θα αναφέρουμε τους πιο γνωστούς αλγόριθμους για ιεραρχική ανάλυση κατά συστάδες, συσσωμάτωσης και διαιρετότητας (Everitt B.S. (1993), σημειώσεις Καρλή, Sneath (1957),



Johnson (1967), Ward (1963), Gower (1967), Lance και William (1967), Jain A.K., Murty M.N. και Flynn P.J. (1999)). Επίσης παραθέτουμε κάποιες μεθόδους που έχουν προταθεί για εύρεση του βέλτιστου αριθμού συστάδων (Everitt B.S. (1993), σημειώσεις Καρλή).

Γράφημα 3.1 Απεικόνιση δενδρογράμματος για αλγορίθμους συσσωμάτωσης και διαιρετότητας.



3.2 Αλγόριθμοι Συσσωμάτωσης

Ένα αλγόριθμος συσσωμάτωσης δημιουργεί μια σειρά διαμερίσεων των δεδομένων P_n, P_{n-1}, \dots, P_1 . Η πρώτη P_n αποτελείται από n συστάδες, με ένα άτομο στη κάθε μια, και η τελευταία P_1 από μια μοναδική ομάδα που περιλαμβάνει n άτομα. Η βασική οργάνωση του αλγορίθμου δίνεται πιο κάτω

ΑΡΧΗ: Συστάδες C_1, C_2, \dots, C_n κάθε μια να περιέχει ένα μόνο άτομο.

1. Βρες το κοντινότερο ζευγάρι συστάδων, έστω C_i, C_j , ένωσε τις C_i, C_j , διέγραψε την C_j και μείωσε των αριθμών των ομάδων κατά ένα

Αν ο αριθμός των ομάδων (συστάδων) ισούται με ένα σταμάτα, διαφορετικά επέστρεψε στο βήμα 1.

Σε κάθε βήμα οι μέθοδοι ενώνουν (συγκολλούν) άτομα ή ομάδες ατόμων τα οποία είναι τα κοντινότερα (ή πιο όμοια). Διαφορές μεταξύ των μεθόδων προκύπτουν από τον ορισμό της απόστασης ή ομοιότητας μεταξύ των ατόμων. Στις παραγράφους που ακολουθούν θα αναπτύξουμε τους αλγορίθμους ατομικής σύνδεσης (single linkage), πλήρους σύνδεσης (complete linkage), group-average, centroid, τη μέθοδο με χρήση του μέσου και τη μέθοδο του Ward.

3.2.1 Αλγόριθμος Ατομικής Σύνδεσης

Ένας από τους πιο απλούς αλγορίθμους συσσωμάτωσης είναι αλγόριθμος ατομικής σύνδεσης (single linkage), συχνά ονομαζόμενος και ως τεχνικής του κοντινότερου γείτονα (nearest neighbour). Πρώτος τον περιέγραψε o Florek et al (1951) και ακολούθησαν οι Sneath (1957) και Johnson (1967). Το καθοριστικό χαρακτηριστικό του αλγόριθμου αυτού είναι ότι η απόσταση ανάμεσα στις ομάδες ορίζεται από το ζευγάρι των ατόμων που μοιάζουν πιο πολύ (έχουν τη μικρότερη απόσταση ή μεγαλύτερη ομοιότητα) σε σχέση με τα υπόλοιπα ζεύγη.

Προκειμένου να γίνει πιο κατανοητή η τεχνική του αλγορίθμου θα την εφαρμόσουμε στον παρακάτω πίνακα αποστάσεων

$$D_1 = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0.0 & & & & \\ 2 & 2.0 & 0.0 & & & \\ 3 & 6.0 & 5.0 & 0.0 & & \\ 4 & 10.0 & 9.0 & 4.0 & 0.0 & \\ 5 & 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}$$

Το μικρότερο νούμερο στον πίνακα είναι το 2.0 το οποίο αντιστοιχεί στο ζευγάρι (1,2) και επομένως τα δύο αυτά άτομα ενώνονται και σχηματίζουν μια συστάδα 2 ατόμων. Οι αποστάσεις ανάμεσα στην σχηματισμένη συστάδα και τα υπόλοιπα μέλη, υπολογίζονται ως εξής:

$$\begin{aligned} d_{(12)3} &= \min[d_{13}, d_{23}] = d_{23} = 5.0 \\ d_{(12)4} &= \min[d_{14}, d_{24}] = d_{24} = 9.0 \\ d_{(12)5} &= \min[d_{15}, d_{25}] = d_{25} = 8.0 \end{aligned} \quad (3.1)$$

Έτσι, τώρα προκύπτει ένας καινούριος πίνακας αποστάσεων

$$D_2 = \begin{pmatrix} (12) & 3 & 4 & 5 \\ (12) & 0.0 & & \\ 3 & 5.0 & 0.0 & \\ 4 & 9.0 & 4.0 & 0.0 \\ 5 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}$$

Το μικρότερο νούμερο στον πίνακα αυτό είναι το 3 μεταξύ των ατόμων 4,5. Οι νέες αποστάσεις τώρα υπολογίζονται ίσες με

$$\begin{aligned} d_{(12)3} &= 5.0 \\ d_{(12)(45)} &= \min[d_{14}, d_{15}, d_{24}, d_{25}] = d_{25} = 8.0 \\ d_{(45)3} &= \min[d_{34}, d_{35}] = d_{34} = 4.0 \end{aligned} \quad (3.2)$$

Αυτές οδηγούν στον πίνακα

$$D_3 = \begin{pmatrix} (12) & 3 & (45) \\ (12) & 0.0 & & \\ 3 & 5.0 & 0.0 & \\ (45) & 8.0 & 4.0 & 0.0 \end{pmatrix}$$

Η μικρότερη απόσταση είναι η $d_{(45)3}$ και έτσι το άτομο 3 προστίθεται στη συστάδα, η οποία περιλαμβάνει τα άτομα 4 και 5. Τέλος οι ομάδες των 1,2 και 3,4,5 ενώνονται για να σχηματίσουν μια τελική ομάδα. Οι διαμερίσεις σε κάθε βήμα είναι:

	Στάδιο	Ομάδες
P ₅		[1],[2],[3],[4],[5]
P ₄		[1,2],[3],[4],[5]
P ₃		[1,2],[3],[4,5]
P ₂		[1,2],[3,4,5]
P ₁		[1,2,3,4,5]

Το αντίστοιχο δενδρόγραμμα απεικονίζεται στο Γράφημα 3.2.

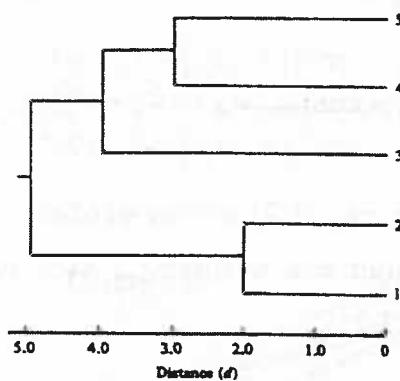
Επίσης πρέπει να τονίσουμε ότι ο αλγόριθμος προχωρά ιεραρχικά με την έννοια ότι για παράδειγμα στο τέταρτο στάδιο δεν θα μπορούσαμε να έχουμε τις ομάδες (1,2,4) και (3,5) εφόσον καμία δεν προκύπτει από ένωση ατόμων παρόντων σε προηγούμενο βήμα.

Δώσαμε ένα παράδειγμα σε ένα σετ δεδομένων το οποίο είναι μικρό. Όταν το σετ δεδομένων μας είναι πολύ μεγάλο τότε τα πράγματα είναι περισσότερο πολύπλοκα.

Το δενδρόγραμμα για τον αλγόριθμο αυτό μπορεί να βρεθεί πολύ εύκολα:

- (1) Βρες την μικρότερη απόσταση- d_{12} - και διέγραψέ την από τον πίνακα
- (2) Βρες την αμέσως μικρότερη- d_{45} - και διέγραψέ την
- (3) Κοίτα την επόμενη μικρότερη απόσταση- d_{34} -ένωσε το άτομο 3 με τα 4 και 5 και διέγραψε την απόσταση
- (4) Μόνο το άτομο 5 έχει απομείνει, οπότε ένωσέ τα όλα μαζί

Γράφημα 3.2 Δενδρόγραμμα για τον αλγόριθμο ατομικής σύνδεσης για τον πίνακα αποστάσεων D_1



Σε γενικές γραμμές τα βήματα που ακολουθεί ο αλγόριθμος ατομικής σύνδεσης είναι τα ακόλουθα:

-
1. Υπολόγισε τον πίνακα αποστάσεων (ομοιότητας), ο οποίος περιλαμβάνει την απόσταση μεταξύ κάθε ζεύγους ατόμων. Μεταχειρίσου κάθε άτομο σαν μια ομάδα
 2. Βρες τις πιο όμοιες ομάδες χρησιμοποιώντας τον πίνακα αποστάσεων (ομοιότητας) και ένωσε αυτές. Υπολόγισε τον καινούριο πίνακα αποστάσεων.
 3. Αν όλα τα άτομα απαρτίζουν μια ομάδα, σταμάτα. Διαφορετικά πήγαινε στο βήμα 2.
-

3.2.2 Αλγόριθμος Πλήρους Σύνδεσης (Complete Linkage)

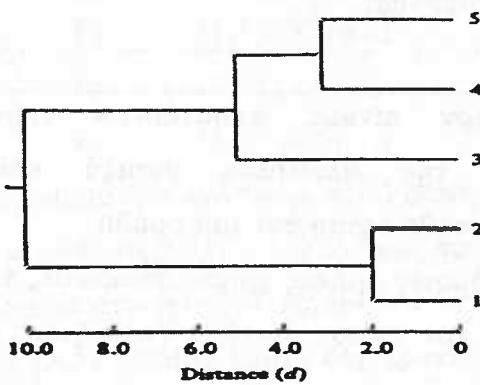
Ο αλγόριθμος πλήρους σύνδεσης ή του μακρινότερου γείτονα είναι ο αντίθετος του αλγόριθμου ατομικής σύνδεσης με την έννοια ότι οι αποστάσεις μεταξύ των ομάδων δεν είναι πλέον οι κοντινότερες, αλλά οι πιο μεγάλες (Everitt B.S. (1993)).

Για να γίνει το παραπάνω κατανοητό θα θεωρήσουμε τον πίνακα αποστάσεων D_1 της παραγράφου 3.2.1. Το πρώτο στάδιο παραμένει ίδιο με αυτό της μεθόδου ατομικής σύνδεσης με τα άτομα 1 και 2 να απαρτίζουν την πρώτη ομάδα. Ο αλγόριθμός τώρα ψάχνει τη μεγαλύτερη απόσταση από τις παρατηρήσεις της ομάδας (άτομο 1 ή άτομο 2) με τα υπόλοιπα άτομα. Έτσι, οι αποστάσεις υπολογίζονται ως

$$\begin{aligned} d_{(12)3} &= \min[d_{13}, d_{23}] = d_{13} = 6.0 \\ d_{(12)4} &= \min[d_{14}, d_{24}] = d_{14} = 10.0 \\ d_{(12)5} &= \min[d_{15}, d_{25}] = d_{15} = 9.0 \end{aligned} \quad (3.3)$$

δηλαδή οι αποστάσεις ανάμεσα στα (1,2) και τα υπόλοιπα άτομα τώρα είναι οι μεγαλύτερες. Το δενδρόγραμμα που αντιστοιχεί στον αλγόριθμο πλήρους σύνδεσης δίνεται στο Γράφημα 3.3.

Γράφημα 3.3 Δενδρόγραμμα για τον αλγόριθμο πλήρους σύνδεσης για τον πίνακα αποστάσεων D_1



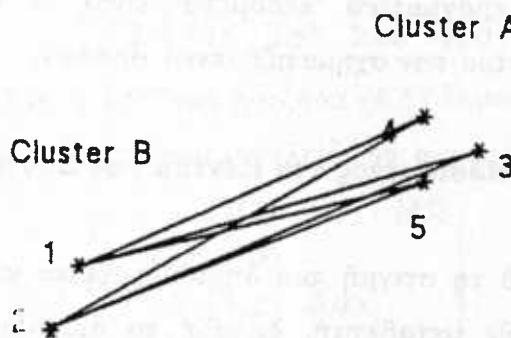
Ένα σημαντικό σημείο των δύο αυτών αλγορίθμων είναι ότι και οι δύο εξαρτώνται από την κατάταξη των αποστάσεων. Αν αλλάζαμε όλες τις αποστάσεις χωρίς να αλλάξουμε την κατάταξή τους, αν για παράδειγμα παίρναμε τις τετραγωνισμένες τιμές τους, δεν θα υπήρχε καμία διαφορά στο

σχηματισμό των ομάδων. Αυτό είναι ένα σημαντικό πλεονέκτημα ιδίως στις εφαρμογές στην κοινωνική επιστήμη όπου οι αποστάσεις ορίζονται συχνά υποκειμενικά και δεν θα ήταν δικαιολογημένο να τις μεταχειριζόμαστε σαν μετρικές.

3.2.3 Αλγόριθμος Group-Average

Και ο αλγόριθμος Group-Average χρησιμοποιεί των πίνακα αποστάσεων (Everitt B.S. (1993)). Η διαφορά τώρα είναι ότι η απόσταση μεταξύ ομάδων βρίσκεται από την μέση απόσταση όλων των ζευγών των παρατηρήσεων που απαρτίζουν την ομάδα. Γραφική απεικόνιση του μέτρου αυτού δίνει το Γράφημα 3.4.

Γράφημα 3.4 Απεικόνιση του αλγόριθμου Group-Average



Εφαρμόζοντας τη μέθοδο στον πίνακα αποστάσεων D_1 της παραγράφου 3.2.3 το πρώτο στάδιο είναι ίδιο με αυτό των δυο προηγούμενων μεθόδων. Σχηματίζεται μια ομάδα από τις παρατηρήσεις 1 και 2. Το νέο σετ αποστάσεων είναι τώρα

$$\begin{aligned} d_{(12)3} &= \frac{1}{2}[d_{13} + d_{23}] = 5.5 \\ d_{(12)4} &= \frac{1}{2}[d_{14} + d_{24}] = 9.5 \\ d_{(12)5} &= \frac{1}{2}[d_{15} + d_{25}] = 8.5 \end{aligned} \quad (3.4)$$

Ο πίνακας αποστάσεων D_2 υπολογίζεται ως

$$D_2 = \begin{pmatrix} (12) & 3 & 4 & 5 \\ (12) & 0.0 & & \\ 3 & 5.5 & 0.0 & \\ 4 & 9.5 & 4.0 & 0.0 \\ 5 & 8.5 & 5.0 & 3.0 & 0.0 \end{pmatrix}$$

Η μικρότερη απόσταση είναι η d_{45} και τώρα η δεύτερη ομάδα που σχηματίζεται είναι η (4,5). Η μέση απόσταση ανάμεσα στις δύο διμελείς ομάδες είναι

$$d_{(12)(45)} = \frac{1}{4} [d_{14} + d_{15} + d_{24} + d_{25}] = 9.0 \quad (3.5)$$

και η διαδικασία συνεχίζει όπως περιγράψαμε στις προηγούμενες παραγράφους.

Μέχρι τώρα ασχοληθήκαμε με αλγόριθμους που βασίζονται στον πίνακα αποστάσεων και όχι με τα πραγματικά δεδομένα. Ένας αλγόριθμος, ο οποίος χρησιμοποιεί τα πραγματικά δεδομένα είναι ο centroid, αλγόριθμος βασισμένος στα κέντρα των σχηματιζόμενων ομάδων.

3.2.4 Αλγόριθμος Βασισμένος στα Κέντρα των Συστάδων (Centroid)

Οι ομάδες από τη στιγμή που δημιουργούνται αναπαρίστανται από τη μέση τιμή για κάθε μεταβλητή, δηλαδή το διάνυσμα του μέσου, και η απόσταση καθορίζεται από το διάνυσμα των μέσων (η απόσταση υπολογίζεται από τα κέντρα των ομάδων). Φυσικά αυτή η μέθοδος υπονοεί ότι οι μεταβλητές πρέπει να είναι τουλάχιστον σε κλίμακα κατάταξης. Παρόλα αυτά η μέθοδος χρησιμοποιείται και σε άλλες κατηγορίες μεταβλητών (Everitt B.S. (1993)).

Θεωρούμε το παράδειγμα για τα παρακάτω άτομα και τις μεταβλητές 1 και 2

Στάδιο	Μεταβλητή 1	Μεταβλητή 2
1	1.0	1.0
2	1.0	2.0
3	6.0	3.0
4	8.0	2.0
5	8.0	0.0

το οποίο δίνει τον παρακάτω πίνακα αποστάσεων βασισμένο στις ευκλείδειες αποστάσεις

$$D_1 = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0.00 & & & & \\ 2 & 1.00 & 0.00 & & & \\ 3 & 5.39 & 5.10 & 0.00 & & \\ 4 & 7.07 & 7.00 & 2.24 & 0.00 & \\ 5 & 7.07 & 7.28 & 3.61 & 2.00 & 0.00 \end{pmatrix}$$

Ο πίνακας έχει σαν μικρότερη τιμή την 1.00 μεταξύ των ατόμων 1 και 2, τα οποία και σχηματίζουν μια ομάδα. Το διάνυσμα των μέσων υπολογίζεται για αυτά τα άτομα ίσο με (1.0, 1.5) και ο νέος πίνακας προκύπτει

$$D_2 = \begin{pmatrix} (12) & 3 & 4 & 5 \\ 3 & 0.00 & & & \\ 4 & 5.22 & 0.00 & & \\ 5 & 7.02 & 2.24 & 0.00 & \\ 5 & 7.16 & 3.61 & 2.00 & 0.00 \end{pmatrix}$$

Η μικρότερη τιμή είναι η 2.00 και η ομάδα (4,5) δημιουργείται με διάνυσμα μέσων το (8.0, 1.0). Ο νέος πίνακας ευκλείδειων αποστάσεων είναι

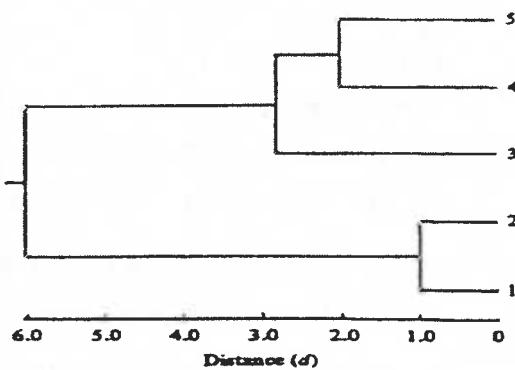
$$D_3 = \begin{pmatrix} (12) & 3 & (45) \\ 3 & 0.00 & & \\ 45 & 5.22 & 0.00 & \\ 45 & 7.02 & 2.83 & 0.00 \end{pmatrix}$$

Ο πίνακας D_3 έχει μικρότερη τιμή για το ζεύγος ([4,5],3) και τώρα τα άτομα 3,4,5 σχηματίζουν μια ομάδα. Το τελικό στάδιο αποτελείται από την τελική ένωση των δυο εναπομενόντων ομάδων. Το δενδρόγραμμα της μεθόδου αυτής είναι στο Γράφημα 3.5.

3.2.5 Αλγόριθμος Βασισμένος στο Διάνυσμα των Διαμέσων

Η μέθοδος αυτή, η οποία προτάθηκε από τον Gower (1967) μοιάζει με την πιο πάνω με τη διαφορά ότι τώρα υπολογίζουμε το διάνυσμα της διαμέσου των μεταβλητών.

Γράφημα 3.5 Δενδρόγραμμα του αλγορίθμου centroid για τον πίνακα αποστάσεων D_1



Η μέθοδος αυτή προτάθηκε λόγω κάποιων μειονεκτημάτων του αλγόριθμου που βασίζεται στο διάνυσμα των μέσων. Το πρώτο παρουσιάζεται όταν τα μεγέθη των ομάδων που πρόκειται να ενωθούν είναι ανόμοια, οπότε το κέντρο της νέας ομάδας θα είναι πολύ κοντά στη μεγαλύτερη ομάδα και μπορεί να παραμείνει μέσα στην ομάδα. Επίσης αν τα κέντρα των ομάδων που ενώνονται είναι τα (α) και (β), τότε η απόσταση από το κέντρο της τρίτης ομάδας (γ) που δημιουργήθηκε από την ένωση των (α) και (β) κείτεται κατά μήκος της διαμέσου του τριγώνου που σχηματίζουν τα κέντρα (α), (β) και (γ).

3.2.6 Αλγόριθμος του Ward

O Ward το 1963 πρότεινε μια μέθοδο αναζητώντας μια διαμέριση P_n, P_{n-1}, \dots, P_1 η οποία να ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες. Σε κάθε βήμα της ανάλυσης υπολογίζουμε την απόσταση της κάθε παρατήρησης από το κέντρο της ομάδας. Αν αθροίσουμε για όλες τις ομάδες έχουμε μια τιμή που είναι το συνολικό άθροισμα. Στο πρώτο βήμα το άθροισμα είναι 0, αφού κάθε παρατήρηση είναι και μια ομάδα. Σε κάθε βήμα ενώνουμε τις ομάδες οι οποίες αν ενωθούν οδηγούν στη μικρότερη αύξηση του συνολικού άθροισματος αποστάσεων.

Η μέθοδος του Ward χρησιμοποιεί το άθροισμα τετραγώνων των καταλοίπων ESS (Error Sum-of Squares)

$$ESS = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.6)$$

Ας θεωρήσουμε τη μονομεταβλητή περίπτωση 10 παρατηρήσεων $(2, 6, 5, 6, 2, 2, 2, 0, 0, 0)$. Το $ESS = (2 - 2.5)^2 + (6 - 2.5)^2 + \dots + (0 - 2.5)^2 = 50.5$

Αν όμως οι 10 παρατηρήσεις ταξινομηθούν σύμφωνα με τα σκορ τους σε 4 σετ

$$\{0, 0, 0\} \quad \{2, 2, 2, 2\} \quad \{5\} \quad \{6, 6\}$$

το ESS μπορεί να υπολογιστεί σαν το άθροισμα των 4 διαφορετικών αθροισμάτων τετραγώνων των λαθών

$$ESS_{\text{ολικό}} = ESS_{\text{Ομάδα 1}} + ESS_{\text{Ομάδα 2}} + ESS_{\text{Ομάδα 3}} + ESS_{\text{Ομάδα 4}} = 0$$

Ο αλγόριθμος λοιπόν ψάχνει να κατατάξει μέλη στις ομάδες που οδηγούν στο μικρότερο ESS μέσα στις ομάδες, όπως στη δεύτερη περίπτωση

3.3 Ο Επαναληπτικός Τύπος των Lance και William

Οι Lance και William (1967) παρήγαγαν έναν επαναληπτικό τύπο, ο οποίος δίνει την απόσταση μεταξύ μιας ομάδας k και μιας ομάδας (ij) σχηματισμένη από την ένωση των ομάδων i και j ως

$$d_{k(ij)} = a_i d_{ki} + a_j d_{kj} + \beta d_{ij} + \gamma |d_{ki} - d_{kj}| \quad (3.7)$$

όπου d_{ij} είναι η απόσταση μεταξύ των ομάδων i και j . Η απόσταση μεταξύ των ατόμων της ομάδας που χρησιμοποιείται στις τεχνικές ιεραρχικών αλγορίθμων τις οποίες περιγράψαμε πιο πάνω προκύπτει από τον τύπο (3.7) με κατάλληλη επιλογή των παραμέτρων:

- Ο αλγόριθμος ατομικής σύνδεσης αντιστοιχεί στις ακόλουθες τιμές: $a_i = a_j = 1/2, \beta = 0, \gamma = -1/2$. Αντικαθιστώντας τις τιμές αυτές στον τύπο (3.7) οδηγούμαστε στη σχέση

$$d_{k(ij)} = \frac{1}{2} d_{ki} + \frac{1}{2} d_{kj} - \frac{1}{2} |d_{ki} - d_{kj}| \quad (3.8)$$

Αν $d_{ki} > d_{kj}$ τότε $|d_{ki} - d_{kj}| = d_{ki} - d_{kj}$ και από (3.8) προκύπτει

$$d_{k(ij)} = d_{kj} \quad (3.9)$$

Αντίστοιχα αν $d_{ki} < d_{kj}$ τότε $|d_{ki} - d_{kj}| = d_{kj} - d_{ki}$ και επομένως

$$d_{k(ij)} = \min[d_{ki}, d_{kj}] \quad (3.10)$$

- Για τον αλγόριθμο πλήρους σύνδεσης οι τιμές είναι

$$a_i = a_j = 1/2, \beta = 0, \gamma = 1/2 \quad (3.11)$$

- Για τον αλγόριθμο Group-Average οι παράμετροι αντιστοιχούν στις τιμές

$$a_i = \frac{n_i}{n_i + n_j}, a_j = \frac{n_j}{n_i + n_j}, \beta = 0, \gamma = 0 \quad (3.12)$$

Για τις 3 μεθόδους τα d_{ij} μπορεί να είναι είτε μέτρα αποστάσεων είτε μέτρα ομοιότητας

- Στην περίπτωση του αλγορίθμου centroid και για ευκλείδειες αποστάσεις έχουμε

$$a_i = \frac{n_i}{n_i + n_j}, a_j = \frac{n_j}{n_i + n_j}, \beta = -a_i a_j, \gamma = 0 \quad (3.13)$$

- Τέλος για την μέθοδο του Ward και στην περίπτωση που τα d_{ij} αντιστοιχούν σε τετραγωνισμένες ευκλείδειες αποστάσεις οι αντίστοιχοι παράμετροι είναι οι

$$a_i = \frac{n_k + n_i}{n_k + n_i + n_j}, a_j = \frac{n_k + n_j}{n_k + n_i + n_j}, \beta = -\frac{n_k}{n_k + n_i + n_j}, \gamma = 0 \quad (3.14)$$

3.4 Ιδιότητες και Προβλήματα των Ιεραρχικών Τεχνικών

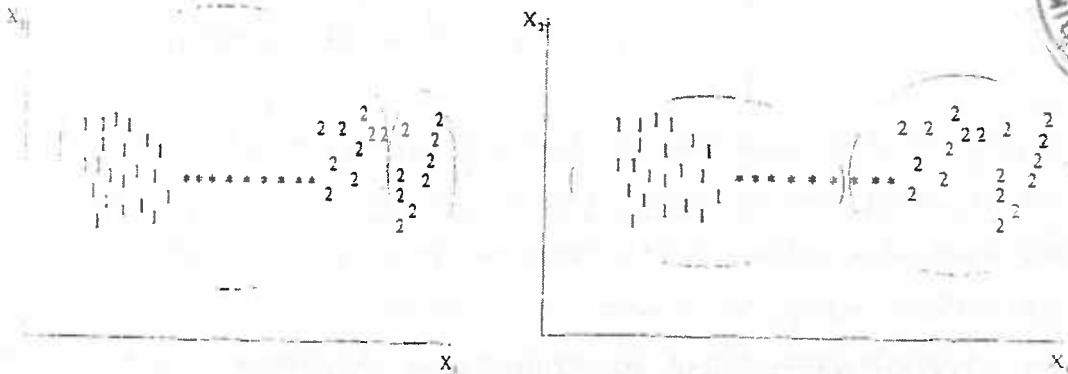
Στις προηγούμενες παραγράφους αναφερθήκαμε σε μερικούς από τους πιο γνωστούς ιεραρχικούς αλγόριθμους για ανάλυση κατά συστάδες. Από αυτούς οι πιο δημοφιλείς είναι ο αλγόριθμος ατομικής σύνδεσης, πλήρους σύνδεσης και η μέθοδος του Ward. Παρόλα αυτά, οι μέθοδοι δεν έχουν την ίδια ικανότητα να βρίσκουν πάντα τις πραγματικές (κατάλληλες) ομάδες. Φαίνεται ότι όλοι οι αλγόριθμοι δεν είναι το ίδιο αποδοτικοί όταν εφαρμόζονται στα ίδια δεδομένα (Jain A.K., Murty M.N. και Flynn P.J. (1999)).

Ο αλγόριθμος ατομικής σύνδεσης έχει την τάση να δημιουργεί ομάδες που είναι επιμήκεις. Ένα τέτοιο παράδειγμα απεικονίζεται στο Γράφημα 3.6. Εδώ οι πραγματικές ομάδες είναι οι 1 και 2, ενώ τα σημεία με τον αστερίσκο



είναι θόρυβος. Βλέπουμε ότι ο αλγόριθμος ατομικής σύνδεσης δεν έχει τη διακριτική ικανότητα να ξεχωρίσει το θόρυβο και να βρει τις πραγματικές ομάδες. Το φαινόμενο αυτό συχνά ονομάζεται σαν πρόβλημα αλυσίδας (chaining). Η ομάδα των σημείων 1 είναι επιμήκης εξαιτίας του θορύβου (*). Αντίθετα παρατηρούμε ότι ο αλγόριθμος πλήρους σύνδεσης είναι πιο αποδοτικός και μπορεί να δει την πραγματική δομή των δεδομένων. Αυτό συμβαίνει διότι ο τελευταίος έχει την τάση να δημιουργεί πιο συμμιγείς ομάδες. Διαφορετικά ο αλγόριθμος ατομικής σύνδεσης είναι πιο ευπροσάρμοστος από τον πλήρους σύνδεσης. Για παράδειγμα ο πρώτος μπορεί να εξάγει τις ομόκεντρες ομάδες του Γραφήματος 3.7. Παρόλα αυτά, από εφαρμογές έχει προκύψει ότι ο αλγόριθμος πλήρους σύνδεσης δίνει καλύτερα αποτελέσματα από τον ατομικής σύνδεσης.

Γράφημα 3.6: Απεικόνιση του προβλήματος chaining



α. Εφαρμογή του αλγορίθμου ατομικής σύνδεσης. Παρατηρούμε την αποτυχία του αλγορίθμου να αναγνωρίσει το

θόρυβο “*”

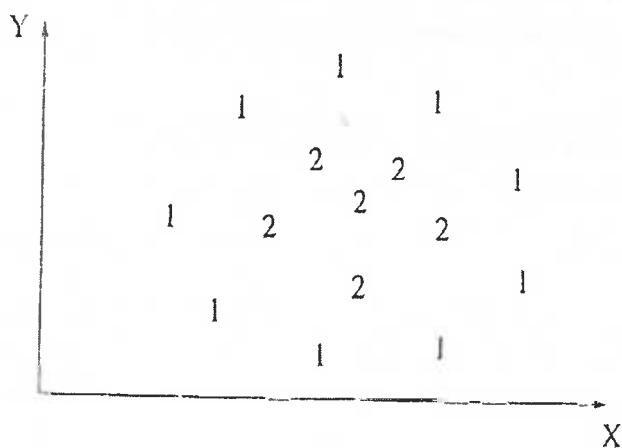
β. Εφαρμογή του αλγορίθμου πλήρους σύνδεσης. Παρατηρούμε την καλύτερη διακριτική ικανότητα του αλγορίθμου σε σχέση με τον ατομικής σύνδεσης.

Από την άλλη, οι δυο παραπάνω αλγόριθμοι έχουν το πλεονέκτημα, όπως αναφέραμε σε προηγούμενη παράγραφο, ότι παραμένουν αμετάβλητοι κάτω από μονοτονικό μετασχηματισμό του πίνακα αποστάσεων (ομοιότητας). Επίσης και οι δυο μπορούν να χρησιμοποιηθούν σε μεγάλα σετ δεδομένων. Επιπλέον, έχουν προταθεί αρκετές μέθοδοι προκειμένου ο αλγόριθμος ατομικής σύνδεσης να ξεπεράσει το πρόβλημα της αλυσίδας.



Στην περίπτωση της ύπαρξης ακραίων παρατηρήσεων ο αλγόριθμος ατομικής σύνδεσης, ο centroid και ο αλγόριθμος που βασίζεται στη διάμεσο παραμένουν ανεπηρέαστοι. Από την άλλη, η μέθοδος του Ward και ο group average δίνουν φτωχά αποτελέσματα. Επίσης, όταν τα δεδομένα είναι τέτοια ώστε να περιέχουν πραγματικές ομάδες που επικαλύπτονται από θόρυβο ο αλγόριθμος ατομικής σύνδεσης, ο αλγόριθμος της διαμέσου και ο centroid δεν δίνουν καλά αποτελέσματα, ενώ η μέθοδος του Ward και ο group average δίνουν ποιοτικότερα αποτελέσματα.

Γράφημα 3.7: Παράδειγμα δεδομένων με δυο ομάδες σε σχήμα ομόκεντρων κύκλων



Τέλος, αρκετοί από τους ιεραρχικούς αλγορίθμους που περιγράψαμε έχουν την τάση να βρίσκουν σφαιρικές ομάδες ακόμα και αν τα δεδομένα αποτελούνται από ομάδες με διαφορετικά σχήματα. Γενικά αποτυγχάνουν να βρουν ομάδες με περίεργα σχήματα.

Συγκρίνοντας τις μεθόδους μεταξύ τους θα πρέπει να γνωρίζουμε ότι από πειράματα προσομοίωσης οι μέθοδοι με την καλύτερη επίδοση είναι του Ward και του group average. Η μέθοδος ατομικής σύνδεσης είναι αυτή με τη χειρότερη επίδοση. Παρόλα αυτά, σε πολλά προβλήματα δεν είναι ζεκάθαρο ποια μέθοδος είναι προτιμότερη. Δεν φαίνεται να υπάρχει μια μοναδική μέθοδος που να είναι ανώτερη σε όλα τα είδη δεδομένων. Αυτό που πρέπει πάντα να έχει ο ερευνητής στο μυαλό του είναι πως αν οι ομάδες είναι αρκετά διαφορετικές μεταξύ τους κάθε μέθοδος θα βρει τη σωστή ομαδοποίηση.

Επίσης θα πρέπει κανείς να γνωρίζει πως κάθε μέθοδος δουλεύει καλύτερα με συγκεκριμένη μορφή δεδομένων.

3.5 Αλγόριθμοι Διαμέρισης

Οι αλγόριθμοι διαμέρισης είναι ουσιαστικά δυο τύπων, οι *μονοθετικοί* που διαμερίζουν τα δεδομένα με βάση το διάνυσμα ενός μόνο ατόμου $\mathbf{x} = (x_1, \dots, x_p)$, με p το συνολικό αριθμό παραμέτρων, και οι *πολυθετικοί* όπου οι διαιρέσεις βασίζονται στις τιμές που παίρνουν τα διανύσματα όλων των ατόμων. Αυτός ο τύπος ανάλυσης κατά συστάδες είναι λιγότερο δημοφιλής και γι' αυτό θα ασχοληθούμε μόνο με τον πιο δημοφιλή πολυθετικό αλγόριθμο, τον K-Means.

3.5.1 Ο Αλγόριθμος K-Means

Η μέθοδος αυτή θεωρεί ότι ο αριθμός των συστάδων είναι γνωστός εκ των προτέρων. Επειδή το τελευταίο δεν είναι γνωστό συνήθως χρειάζεται είτε να τρέξουμε τον αλγόριθμο με διαφορετικό αριθμό ομάδων κάθε φορά είτε πρέπει με κάποιον άλλο τρόπο να έχουμε επιλέξει τον αριθμό ομάδων (π.χ. συνηθίζεται να εφαρμόζεται μια ανάλυση σε κυρίες συνιστώσες για μια πρώτη ιδέα του αριθμού των ομάδων).

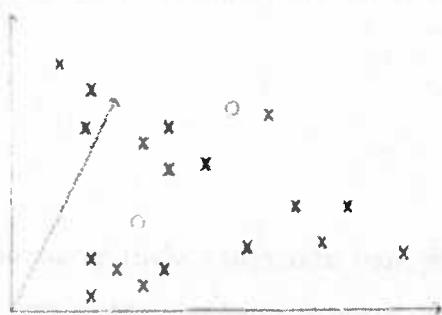
Ο K-Means (Πίνακας 3.1) δουλεύει επαναληπτικά (σημειώσεις Καρλή). Χρησιμοποιεί την έννοια του κέντρου της ομάδας (*centroid*) και στη συνέχεια κατατάσσει τις παρατηρήσεις ανάλογα με την απόστασή τους από τα κέντρα των άλλων ομάδων. Το κέντρο της ομάδας κι εδώ είναι το διάνυσμα των μέσων για κάθε μεταβλητή.

Στη συνέχεια για κάθε μεταβλητή υπολογίζουμε την ευκλείδεια απόστασή της από τα κέντρα των ομάδων που έχουμε και κατατάσσουμε την παρατήρηση στην ομάδα που είναι πιο κοντά (στην ομάδα με κέντρο πιο κοντά στην παρατήρηση). Αφού κατατάξουμε όλες τις παρατηρήσεις τότε υπολογίζουμε εκ νέου τα κέντρα, απλά ως τα διανύσματα των μέσων για τις παρατηρήσεις που ανήκουν στην κάθε ομάδα. Η διαδικασία επαναλαμβάνεται

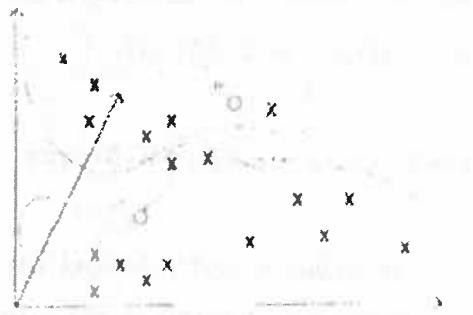
έως ότου δεν υπάρχουν διαφορές ανάμεσα σε δυο διαδοχικές επαναλήψεις. Αν κάποιος θέλει να χρησιμοποιήσει διαφορετική απόσταση από την ευκλείδεια θα πρέπει να κάνει ειδικούς μετασχηματισμούς στα δεδομένα πριν τη χρησιμοποιήσει.

Το Γράφημα 3.8 δείχνει πώς δουλεύει ο αλγόριθμος. Οι παρατηρήσεις συμβολίζονται με “x” και τα αρχικά κέντρα με “o” στην εικόνα α. Στην εικόνα β σχηματίζονται 2 νέφη τα οποία προκύπτουν μετρώντας την απόσταση από κάθε κέντρο και κατατάσσοντάς την στην ομάδα με το πλησιέστερο κέντρο. Στη εικόνα γ βλέπουμε τα νέα κέντρα που σχηματίζονται με βάση τα προηγούμενα νέφη, κλπ.

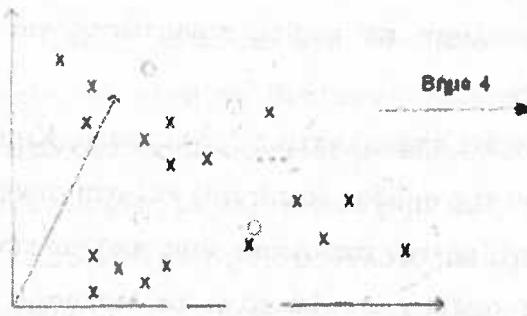
Γράφημα 3.8: Εφαρμογή του αλγορίθμου K-Means



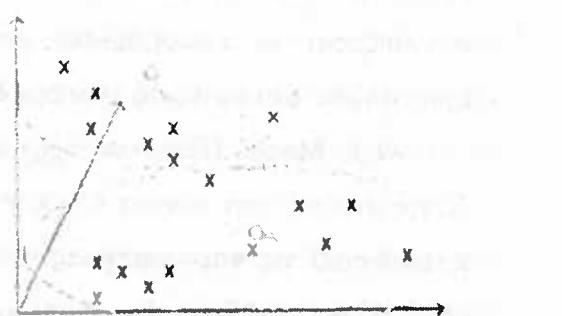
Βήμα 1: Επιλογή αρχικών κέντρων



Βήμα 2: Κατάταξη παρατηρήσεων σε ομάδες



Βήμα 3: Υπολογισμός νέων κέντρων

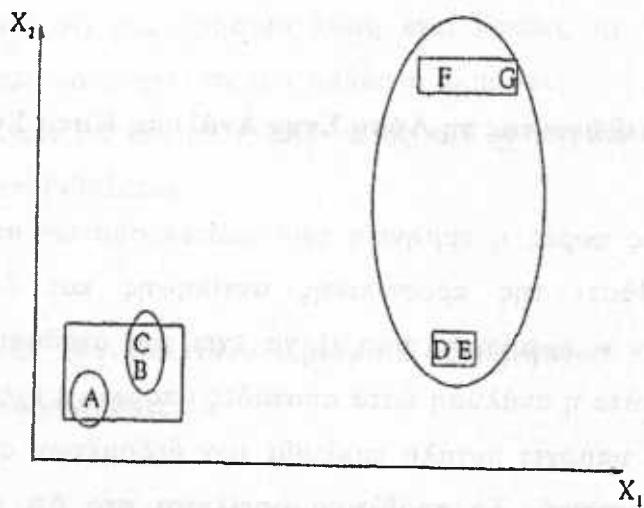


Βήμα 4: Επαν-καταταξη παρατηρήσεων σε ομάδες

Ο αλγόριθμος K-Means δουλεύει αρκετά καλά για μεγάλα σετ δεδομένων και μάλιστα πιο γρήγορα από τους αλγόριθμους ιεραρχικής ομαδοποίησης και χρειάζεται σχετικά λίγες επαναλήψεις. Ο αλγόριθμος ελαχιστοποιεί το άθροισμα των τετραγωνικών αποστάσεων των παρατηρήσεων από τα κέντρα των ομάδων που ανήκουν. Συνήθως η λύση περιέχει ομάδες με περίπου ίσο αριθμό παρατηρήσεων.

Το μεγάλο μειονέκτημα του αλγορίθμου είναι ότι πρέπει να γνωρίζουμε εκ των προτέρων τον αριθμό των ομάδων. Πρακτικά αυτό σημαίνει ότι πρέπει να τρέξουμε τον αλγόριθμο αρκετές φορές με διαφορετικό αριθμό ομάδων προκειμένου να επιλέξουμε τη βέλτιστη λύση (βλέπε παράγραφο 3.6), διότι λάθος διαμέριση πολλές φορές οδηγεί τον αλγόριθμο σε κάποιο τοπικό παρά στο ολικό ελάχιστο. Για να γίνει το πιο πάνω κατανοητό ας θεωρήσουμε το παράδειγμα των 7 παρατηρήσεων του Γραφήματος 3.9, τις οποίες θέλουμε να κατατάξουμε σε 3 ομάδες (Jain A.K., Murty M.N. και Flynn P.J. (1999)). Αρχικά χρησιμοποιούμε τις παρατηρήσεις A, B, C σαν αρχικές τιμές οπότε προκύπτουν οι ομάδες {A}, {B,C}, {E,D,G,F}. Αν Όμως ξεκινήσουμε από τα σημεία A, D ,F προκύπτουν οι ομάδες (συμβολίζονται στο Γράφημα με το τετράγωνο) {A,B,C}, {E,D}και {G,F}. Στην πρώτη περίπτωση οι διαφορές μέσα στις ομάδες είναι πολύ μεγαλύτερες από τη δεύτερη.

Γράφημα 3.9: Ενασθησία του αλγορίθμου K-Means στην επιλογή των αρχικών κέντρων



Τέλος θα πρέπει κάποιος να έχει υπόψη του ότι συνήθως η μέθοδος χρησιμοποιεί την ευκλείδεια απόσταση, αλλά μπορεί να χρησιμοποιηθεί οποιαδήποτε απόσταση. Στην περίπτωση των μη συνεχών δεδομένων υπάρχει το πρόβλημα ότι δεν μπορούμε να ορίσουμε το μέσο της ομάδας, αλλά σε αυτή την περίπτωση μπορούμε να χρησιμοποιήσουμε αντίστοιχα μέτρα.

Πίνακας 3.1: Ο Αλγόριθμος K-Means (σημειώσεις Καρλή)

-
- **Βήμα 1:** Βρες τα αρχικά κέντρα
 - **Βήμα 2:** Κατέταξε κάθε παρατήρηση στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση
 - **Βήμα 3:** Από τις παρατηρήσεις που είναι μέσα στην ομάδα υπολόγισε τα νέα κέντρα
 - **Βήμα 4:** Αν τα νέα κέντρα δεν διαφέρουν από τα παλιά σταμάτα, διαφορετικά πήγαινε στο βήμα 2
-

Στην περίπτωση των κατηγορικών δεδομένων με κατάταξη (ordinal data) μπορούμε να χρησιμοποιήσουμε το διάνυσμα των διαμέσων (medoid) ή για ονομαστικά δεδομένα την κορυφή, την τιμή με τη μεγαλύτερη συχνότητα. Φυσικά αυτές οι επιλογές είναι κατά πολύ κατώτερες λόγω των ιδιοτήτων τους, αλλά μας προσφέρουν τη δυνατότητα χρήσης του αλγορίθμου σε κάθε μορφής δεδομένα. Στην περίπτωση μεικτού τύπου δεδομένων το κέντρο κάθε ομάδας μπορεί να αποτελείται από τις κορυφές των κατηγορικών μεταβλητών και τους μέσους των συνεχών.

3.6 Εξακριβώνοντας τη Λύση Στην Ανάλυση Κατά Συστάδες

Πολλές φορές η ερμηνεία των αποτελεσμάτων από έναν αλγόριθμο προκύπτει βάσει της προσωπικής αντίληψης και διορατικότητας του ερευνητή. Αν ο ερευνητής μπορεί να έχει μια αίσθηση των ομάδων που παράγονται τότε η ανάλυση κατά συστάδες μπορεί να έχει επιτυχία. Από την άλλη αν δεν υπάρχει μεγάλη εμπειρία των δεδομένων αυτή μπορεί να μην είναι ικανοποιητική. Το πρόβλημα οφείλεται στο ότι εφαρμόζοντας έναν αλγόριθμο σε τυχαία δεδομένα αυτός θα δώσει κάποια ομαδοποίηση ακόμη κι αν αυτό δεν είναι αληθές.

Αρκετά τεστ έχουν προταθεί κατά καιρούς τα οποία σκοπό έχουν να εξετάσουν αν πραγματικά τα δεδομένα φανερώνουν ότι υπάρχει κάποια ομαδοποίηση (Everitt B.S. (1993)). Η μηδενική υπόθεση συνήθως παίρνει κάποια από τις ακόλουθες μορφές.

H_0 : Όλοι οι $n \times n$ πίνακες ομοιότητες (απόστασης) είναι όμοιοι



H_0 : Όλα τα σετ των n θέσεων σε ένα χώρο p διαστάσεων είναι ισοδύναμα

H_0 : Όλες οι διατάξεις των n ατόμων είναι ισοδύναμες.

Δεν θα επεκταθούμε περισσότερο στην παράθεση αυτών των δοκιμασιών. Πρέπει όμως να τονίσουμε ότι τα τεστ αυτά δεν χρησιμοποιούνται ιδιαίτερα κυρίως διότι δεν είναι διαθέσιμα στα στατιστικά πακέτα που χρησιμοποιούνται για την ανάλυση κατά συστάδες.

Η σταθερότητα της λύσης της ομαδοποίησης μπορεί να εξεταστεί διαιρώντας τυχαία τα δεδομένα σε δύο υπό-σετ και εφαρμόζοντας την ανάλυση σε κάθε σετ ξεχωριστά. Παρόμοιες λύσεις πρέπει να προκύψουν και από τα δυο σετ όταν τα δεδομένα είναι ξεκάθαρα δομημένα. Παρομοίως, η ανάλυση μπορεί να επαναληφθεί χρησιμοποιώντας κάποιες και όχι όλες τις μεταβλητές. Διαγραφή μικρού αριθμού μεταβλητών από την ανάλυση δεν πρέπει λογικά να επηρεάζει τις ομάδες που βρέθηκαν αρχικά, αν αυτές είναι οι πραγματικές. Μια άλλη λύση είναι μια επαναληπτική διαδικασία η οποία να συγκρίνει τις ομάδες που βρέθηκαν με τη χρήση μεταβλητών άλλων από αυτές που αρχικά χρησιμοποιήθηκαν. Αν οι διαφορές στις ομάδες παραμένουν αυτό στοιχειοθετεί ότι μια χρήσιμη λύση έχει βρεθεί, με την έννοια ότι θέτοντας ένα άτομο να ανήκει σε μια ομάδα η πληροφόρηση που προέρχεται από τη χρήση άλλων μεταβλητών, από αυτές που χρησιμοποιήθηκαν αρχικά στην ανάλυση, μεταβιβάζεται.

3.7 Εξετάζοντας τον Βέλτιστο Αριθμό Ταξινομήσεων - Συγκρίνοντας Διαφορετικές Ταξινομήσεις

Όπως αναφέραμε και προηγουμένως ένα από τα προβλήματα που έχει να αντιμετωπίσει ο ερευνητής είναι η επιλογή του βέλτιστου αριθμού ομάδων. Κάποιες τεχνικές (κριτήρια), οι οποίες έχουν σκοπό να εξετάσουν ποιος είναι ο πιο κατάλληλος αριθμός ομάδων (Everitt B.S. (1993), σημειώσεις Καρλή), είναι:

- Υπολόγισε το λόγο $E(2)/E(1)$, όπου $E(2)$ είναι το άθροισμα τετραγώνων των λαθών μέσα σε κάθε ομάδα όταν τα δεδομένα διαμερίζονται σε δυο ομάδες και $E(1)$ δίνει το άθροισμα τετραγώνων





των λαθών όταν υπάρχει μόνο μια ομάδα. Η υπόθεση ύπαρξης μιας μοναδικής ομάδας απορρίπτεται αν ο λόγος είναι μικρότερος από μια συγκεκριμένη τιμή.

- Υπολόγισε το $\frac{\text{trace}(\mathbf{B})/(g-1)}{\text{trace}(\mathbf{W})/(n-g)}$, όπου \mathbf{B} και \mathbf{W} είναι πίνακες που αντιπροσωπεύουν το άθροισμα τετραγώνων μεταξύ των ομάδων (between clusters sum of squares) και εντός των ομάδων (within clusters sum of squares). Η μέγιστη τιμή του παραπάνω δείκτη υποδεικνύει το σωστό αριθμό των ομάδων.

Συνήθως αναλύοντας ένα σετ δεδομένων χρησιμοποιούνται διάφορες τεχνικές για ανάλυση κατά συστάδες. Επομένως χρειάζεται να γίνει σύγκριση μεταξύ των διαφόρων τεχνικών. Συχνά πάλι συνηθίζεται ακόμα και αν εφαρμόζεται μια τεχνική να χρησιμοποιούνται διαφορετικοί πίνακες ομοιότητας (αποστάσεων), οπότε μια σύγκριση μεταξύ τους είναι απαραίτητη. Αν το σετ δεδομένων δεν είναι μεγάλο αυτές οι συγκρίσεις μπορούν να πραγματοποιηθούν άτυπα, απλά εξετάζοντας τις διαφορετικές ομάδες ή τα δενδρογράμματα που μοιάζουν και που διαφέρουν. Σε πολλές εφαρμογές, όμως, η προσέγγιση αυτή είναι επίπονη και χρονοβόρα. Κατά συνέπεια αρκετοί ερευνητές πρότειναν πιο τυποκρατικές διαδικασίες για τη σύγκριση των ταξινομήσεων. Εδώ θα αναφερθούμε στην πιο συχνά χρησιμοποιούμενη.

Ας θεωρήσουμε ότι έχουμε n άτομα για ομαδοποίηση. Τότε για ένα συγκεκριμένο αριθμό ομάδων, g , ορίζουμε το δείκτη R_g ως το λόγο του αθροίσματος των συνολικών ζευγαριών ατόμων τα οποία ομαδοποιούνται μαζί στις δυο διαφορετικές τεχνικές ανάλυσης κατά συστάδες που πρόκειται να συγκριθούν και του αριθμού των ζευγών των ατόμων τα οποία πέφτουν σε διαφορετικές ομάδες στις δυο τεχνικές προς τον συνολικό αριθμό ζευγών $\binom{n}{2}$. Επομένως ο R_g μπορεί να μεταφραστεί σαν την πιθανότητα δυο άτομα να χρησιμοποιηθούν με τον ίδιο τρόπο στις δυο τεχνικές. Οπότε,

$$R_g = \left[T_g - \frac{1}{2} P_g - \frac{1}{2} Q_g + \binom{n}{2} \right] / \binom{n}{2} \quad (3.15)$$

όπου



$$T_g = \sum_{i=1}^g \sum_{j=1}^g m_{ij}^2 - n \quad (3.16)$$

$$P_g = \sum_{i=1}^g m_i^2 - n \quad (3.17)$$

$$Q_g = \sum_{i=1}^g m_j^2 - n \quad (3.18)$$

και η ποσότητα m_{ij} είναι ο αριθμός των ατόμων κοινά στην ομάδα i της πρώτης λύσης και της j ομάδας στη δεύτερη. Οι όροι m_j και m_i είναι οι περιθώριες του πίνακα τιμών m_{ij} . Ο R_g παίρνει τιμές στο διάστημα $[0,1]$ και παίρνει τη μέγιστη τιμή όταν υπάρχει πλήρης συμφωνία μεταξύ των δύο διαφορετικών ταξινομήσεων.

Ένας διαφορετικός δείκτης που συγκρίνει δυο διαφορετικές ιεραρχικές μεθόδους και ο οποίος μοιάζει με τον R_g είναι ο

$$B_g = T_g / \sqrt{P_g Q_g} \quad (3.19)$$

Ο δείκτης αυτός χρησιμοποιείται κάνοντας τη γραφική παράσταση των (g, B_g) , $g = 2, \dots, n-1$ για κάθε ζεύγος διαμερίσεων που προκύπτει από τα δύο δενδρογράμματα.

ΚΕΦΑΛΑΙΟ 4

ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ ΜΕ ΧΡΗΣΗ ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

4.1 Εισαγωγή

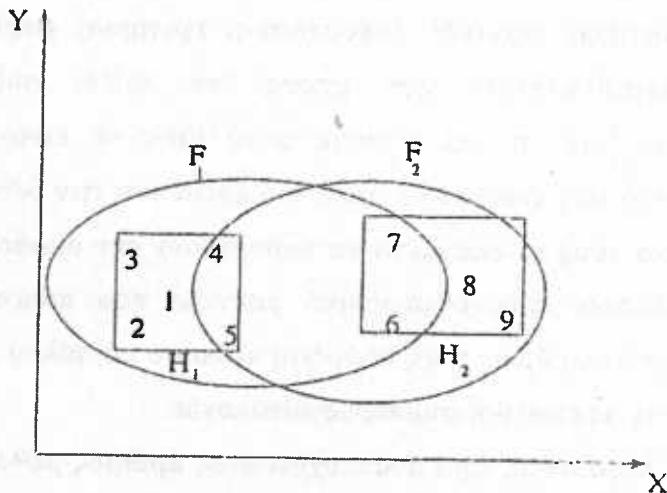
Οι μέθοδοι για ανάλυση κατά συστάδες που περιγράψαμε στο προηγούμενο κεφάλαιο έχουν ως κοινό σημείο ότι δεν έχουν βασιστεί σε κάποιο μοντέλο, το οποίο είναι το κύριο εργαλείο των μεθόδων ανάλυσης στην στατιστική επιστήμη. Οι ιεραρχικές μέθοδοι βασίζονται καθαρά σε χρήση μαθηματικών τεχνικών (αποστάσεις, κριτήρια) χωρίς να λαμβάνεται υπόψη η μεταβλητότητα που μπορεί να παίζει σοβαρό ρόλο στα αποτελέσματα. Από τη μια πλευρά, αυτό ίσως να είναι επιθυμητό (π.χ. περιπτώσεις που μας ενδιαφέρει απλά η διερεύνηση των δεδομένων) μιας και αφήνουμε μόνα τους τα δεδομένα να καθορίσουν την ομαδοποίηση, χωρίς να υποθέτουμε κάποιο πιθανοθεωρητικό μοντέλο που πιθανόν να είναι και λάθος, από την άλλη όμως η μη θεώρηση κάποιου μοντέλου δεν μας επιτρέπει να προβούμε σε στατιστική συμπερασματολογία.

Για το λόγο αυτό, έχει αναπτυχθεί ένας αριθμός μεθοδολογιών για την προσέγγιση του προβλήματος της ανάλυσης κατά συστάδες, οι οποίες έχουν σαν κοινή βάση την χρήση/ υπόθεση στατιστικού μοντέλου (π.χ. Aitkin, Anderson και Hinde (1981), Symons (1981), McLachlan (1982), McLachlan και Basford (1988)). Η βασική διαφορά με τις μέχρι τώρα αναφερθείσες ιεραρχικές μεθόδους είναι ότι τώρα κάθε στοιχείο δεν ανήκει σε μια και μοναδική συστάδα, αλλά ανήκει σε κάθε συστάδα με συγκεκριμένη πιθανότητα (γνωστό ως fuzzy clustering) (Γράφημα 4.1). Στο κεφάλαιο αυτό θα αναπτύξουμε τη μέθοδο που χρησιμοποιεί **μίξεις κατανομών** (mixture models) και ιδιαίτερα μίξεις κανονικών κατανομών (Dempster et al (1977), Fraley C. και Raftery A.E. (1998), Banfield J.D. και Raftery A.E. (1993), McLachlan G.J., Peel D. και Basford K.E. (2002), σημειώσεις Καρλή).

Θα αναφερθούμε στις μίξεις κατανομών και κυρίως κανονικών κατανομών, θα περιγράψουμε τον EM αλγόριθμο (Estimation Maximization

algorithm) για την εκτίμηση των παραμέτρων των μελών της μίξης καθώς και μεθόδους προκειμένου επιλέξουμε τον κατάλληλο αριθμό των μελών της μίξης (Wolfe (1971) και Fraley C. και Raftery A.E. (1998)). Τέλος θα περιγράψουμε την στρατηγική εφαρμογής της μίξης κατανομών και θα περιγράψουμε την χρήση πιθανοθεωρητικού μοντέλου στην περίπτωση διμεταβλητών δεδομένων (McLachlan και Basford (1988), Everitt B.S. (1993)).

Γράφημα 4.1: Ανάλυση κατά συστάδες με χρήση πιθανοθεωρητικού μοντέλου. Οι ομάδες H_1, H_2 έχουν προκύψει χωρίς χρήση πιθανοθεωρητικού μοντέλου και κάθε άτομο ανήκει αποκλειστικά σε κάποια ομάδα, ενώ οι F_1 και F_2 με χρήση πιθανοθεωρητικού μοντέλου.



4.2 Εισαγωγικές Έννοιες σε Μίξεις Κατανομών

Έστω ότι ένας πληθυσμός αποτελείται από k υποπληθυσμούς. Κάθε υποπληθυσμός ακολουθεί συγκεκριμένη κατανομή, f_j , $j=1, \dots, k$. Θεωρούμε επίσης ότι όλοι οι υποπληθυσμοί ακολουθούν την ίδια κατανομή, π.χ. την κανονική, οπότε συμβολίζουμε την κατανομή του j υποπληθυσμού με $f(\mathbf{x}|\theta_j)$, όπου θ_j το διάνυσμα των παραμέτρων για την κατανομή και \mathbf{x} το διάνυσμα των παρατηρήσεων. Αν πάρουμε ένα άτομο από τον πληθυσμό

τυχαία και δεν γνωρίζουμε από ποιον υποπληθυσμό προέρχεται τότε από το θεώρημα ολικής πιθανότητας η κατανομή του θα είναι

$$f(x) = \sum_{j=1}^k p_j f(x|\theta_j) \quad (4.1)$$

όπου $0 < p_j < 1$, $\sum_{j=1}^k p_j = 1$ η πιθανότητα ένα τυχαίο άτομο να ανήκει στον υποπληθυσμό j (σημειώσεις Καρλή).

Για να γίνει το πιο πάνω κατανοητό ας θεωρήσουμε σαν παράδειγμα ένα δείγμα ατόμων όπου έχει καταγραφεί το ύψος τους. Το δείγμα θα περιέχει το ύψος αντρών και γυναικών, το οποίο είναι γνωστό ότι διαφέρει. Αν έχει καταγραφεί το φύλο του δείγματος, τότε η εκτίμηση της μέσης τιμής και της διακύμανσης για τα ύψη των αντρών και γυναικών είναι υπόθεση ρουτίνας. Τι γίνεται όμως στην περίπτωση όπου το φύλο δεν είναι γνωστό; Τώρα η πυκνότητα πιθανότητας του ύψους θα έχει τη μορφή

$$h(\text{ύψος}) = p(\text{θήλυ}) h_1(\text{ύψος}; \text{θήλυ}) + p(\text{άρρεν}) h_2(\text{ύψος}; \text{άρρεν})$$

όπου $p(\text{θήλυ})$ και $p(\text{άρρεν})$ η πιθανότητα ότι ένα μέλος του πληθυσμού είναι αντίστοιχα γυναίκα ή άντρας και h_1 και h_2 οι συναρτήσεις πυκνότητας του ύψους για τις γυναίκες και τους άντρες.

Στην περίπτωση αυτή, όπου οι δεσμευμένες πιθανότητες δεν είναι γνωστές, η εκτίμηση των πιθανοτήτων του να είναι κάποιος άντρας ή γυναίκα καθώς και η εκτίμηση των παραμέτρων των δεσμευμένων κατανομών μπορεί να είναι πολύ δύσκολη. Παρόλα αυτά, ο Karl Pearson (1984) υποθέτοντας ότι οι h_1 και h_2 ακολουθούν κανονικές κατανομές μπόρεσε με τη μέθοδο των ροπών να εκτιμήσει τις παραπάνω παραμέτρους.

Στην ενότητα που ακολουθεί θα περιγράψουμε την εκτίμηση των παραμέτρων μίξης κανονικών κατανομών (και μάλιστα πολυμεταβλητών κανονικών κατανομών) με τη μέθοδο της μεγιστοποίησης της πιθανοφάνειας (maximum likelihood estimation).

4.3 Εκτίμηση με τη Μέθοδο Μεγίστης Πιθανοφάνειας

Θεωρούμε την (4.1) να περιλαμβάνει περισσότερες από δύο παραμέτρους, με κάθε διάνυσμα παραμέτρων να συσχετίζεται με ένα μέλος

της μίξης, και επίσης να περιλαμβάνει περισσότερες από μια μεταβλητές. Θεωρώντας επιπλέον ότι κάθε παράμετρος ακολουθεί πολυμεταβλητή κανονική κατανομή η (4.1) παίρνει τη μορφή

$$f(\mathbf{x}) = \sum_{j=1}^k p_j f(\mathbf{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad (4.2)$$

με

$$f(\mathbf{x}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{p/2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right] \quad (4.3)$$

όπου p ο αριθμός των μεταβλητών, k ο αριθμός των μελών και $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ το διανύσμα του μέσου και ο πίνακας συνδιακύμανσης του i μέλουν.

Οι συστάδες τώρα μπορούν να σχηματιστούν με βάση τις μέγιστες τιμές των εκτιμώμενων εκ των υστέρων πιθανοτήτων

$$\hat{p}(s/x) = \frac{\hat{p}_s f(x, \hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\Sigma}}_s)}{\sum_{j=1}^k \hat{p}_j f(x, \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)} \quad (4.4)$$

,όπου το $\hat{p}(s/x)$ είναι η εκτιμώμενη πιθανότητα ότι η παρατήρηση x προέρχεται από τον πληθυσμό s .

Πριν αναφερθούμε στην εκτίμηση των παραπάνω πιθανοτήτων θα πρέπει να βρεθεί ένας τρόπος εκτίμησης των παραμέτρων της πυκνότητας (4.2). Για το σκοπό αυτό θα γράψουμε την λογαριθμοποιημένη πιθανοφάνεια της (4.2) και στη συνέχεια θα την μεγιστοποιήσουμε παίρνοντας την πρώτη παράγωγο, ως προς κάθε παράμετρο, ίση με το μηδέν.

Η λογαριθμοποιημένη πιθανοφάνεια έχει την παρακάτω μορφή

$$L = \sum_{i=1}^n \ln \left[\sum_{j=1}^k p_j f(\mathbf{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right] \quad (4.5)$$

και οι εξισώσεις της μεγιστοποιημένης πιθανοφάνειας είναι οι

$$\hat{p}_i = \frac{1}{n} \sum_{j=1}^n \hat{p}(i/x_j), \quad i = 1, 2, \dots, k-1 \quad (4.6)$$

$$\boldsymbol{\mu}_i = \frac{1}{n \hat{p}_i} \sum_{j=1}^n \hat{p}(i/x_j) \mathbf{x}_j, \quad i = 1, 2, \dots, k \quad (4.7)$$

$$\boldsymbol{\Sigma}_i = \frac{1}{n \hat{p}_i} \sum_{j=1}^n \hat{p}(i/x_j) (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i) (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i)', \quad i = 1, 2, \dots, k \quad (4.8)$$

Γραμμένες σε αυτή τη μορφή οι εξισώσεις (4.6)-(4.8) φαίνονται να είναι ανάλογες με εκείνες για την εκτίμηση των παραμέτρων μιας απλής κανονικής

κατανομής εκτός του ότι εδώ κάθε στοιχείο του δείγματος επιβαρύνεται με την εκ των υστέρων πιθανότητα (4.4). Οι παραπάνω εξισώσεις δεν δίνουν τις ακριβείς εκτιμήσεις των παραμέτρων. Ένας τρόπος εύρεσης προσεγγιστικής λύσης είναι η χρήση μιας επαναληπτικής διαδικασίας, γνωστή σαν EM αλγόριθμος (Estimation Maximization Algorithm).

4.4. Ο EM Αλγόριθμος για Υπολογισμό Παραμέτρων Μίξης

Ο EM αλγόριθμος εισήχθηκε από τους Dempster et al (1977) και είναι μια επαναληπτική διαδικασία για την εκτίμηση των παραμέτρων της (4.2). Τα «ολοκληρωμένα» δεδομένα δεν είναι το διάνυσμα των δεδομένων $\mathbf{y}_{obs} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T$, αλλά το διάνυσμα $\mathbf{x}_i = (\mathbf{y}_i, \mathbf{z}_i)$ (McLachlan G.J., Peel D., Basford K.E. και Adams P. (2002)), όπου \mathbf{z}_i ορίζεται ως

$$z_{ik} = \begin{cases} 1 & \text{αν } \mathbf{y}_i \text{ ανήκει στον υποπληθυσμό } k \\ 0 & \text{διαφορετικά} \end{cases} \quad (4.9)$$

Επιπλέον θεωρούμε ότι τα $x_c = (x_1^T, x_2^T, \dots, x_n^T)^T$, όπου $x_i^T = (y_i^T, z_i^T)^T$, $\dots, x_n^T = (y_n^T, z_n^T)^T$, είναι ανεξάρτητα και ισόνομα κατανεμημένα με τα z_1, \dots, z_n να είναι ανεξάρτητες πραγματώσεις μιας πολυωνυμικής κατανομής με αντίστοιχες πιθανότητες (ή πραγματοποιήσεις) p_1, \dots, p_k , δηλαδή

$$z_1, \dots, z_n \stackrel{iid}{\sim} Mult_k(1, \mathbf{p}) \quad (4.10)$$

όπου $\mathbf{p} = (p_1, \dots, p_k)^T$. Με τους παραπάνω προσδιορισμούς η λογαριθμοποιημένη πιθανοφάνεια ισούται με

$$\ln L = \sum_{i=1}^k \sum_{j=1}^n z_{ij} \ln \left\{ p_i f(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right\} \quad (4.11)$$

Ο αλγόριθμος EM (Πίνακας 4.1) είναι εύκολος να προγραμματιστεί και προχωρά επαναληπτικά σε 2 στάδια: E (για εύρεση της μαθηματικής ελπίδας) και M (για τη μεγιστοποίηση). Στην (K+1) επανάληψη το βήμα E απαιτεί τον υπολογισμό του



$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} (\ln L(\Psi) | \mathbf{y}_{obs}), \quad (4.12)$$

όπου $\Psi = (p_1, \dots, p_{k-1}, \boldsymbol{\mu}_i^T, \boldsymbol{\Sigma}_i^T)^T$, δηλαδή τη δεσμευμένη μέση τιμή της λογαριθμοποιημένης πιθανοφάνειας των «ολοκληρωμένων» δεδομένων δοθέντος των παρατηρήσεων \mathbf{y}_{obs} , χρησιμοποιώντας την $\Psi^{(k)}$ για Ψ . Εφόσον η λογαριθμοποιημένη πιθανοφάνεια είναι γραμμική συνάρτηση της μεταβλητής z_{ik} τότε η αλλαγή στο βήμα Ε προκύπτει αντικαθιστώντας απλά το z_{ik} με την δεσμευμένη μέση τιμή της δοθέντος του y_j , χρησιμοποιώντας την $\Psi^{(k)}$ για Ψ . Με άλλα λόγια, το z_{ik} αντικαθίσταται από το

$$\begin{aligned} \tau_i(\mathbf{y}_j; \Psi^{(k)}) &= E_{\Psi^{(k)}}(Z_{ij} | \mathbf{y}_j) = pr_{\Psi^{(k)}}(Z_{ij} = 1 | \mathbf{y}_j) \\ &= \frac{p_i f(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{h=1}^n p_h f(\mathbf{y}_j; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)} \quad i=1, \dots, k \text{ και } j=1, \dots, n \end{aligned}$$

όπου $\tau_i(\mathbf{y}_j; \Psi^{(k)})$ είναι η εκτίμηση στο βήμα αυτό της εκ των υστέρων πιθανότητας ότι η j οντότητα με διάνυσμα \mathbf{y}_j ανήκει στην i συνιστώσα.

Στο βήμα Μ της (K+1) επανάληψης σκοπός είναι να επιλεγεί εκείνη η τιμή του Ψ , έστω $\Psi^{(k+1)}$, που μεγιστοποιεί το $Q(\Psi; \Psi^{(k)})$. Ακολουθεί ότι στο βήμα Μ για την (K+1) επανάληψη, η τρέχουσα προσαρμογή των πιθανοτήτων των μίξεων, οι μέσοι των συνιστωσών και οι πίνακες συνδιακύμανσης δίδονται από τις παρακάτω εξισώσεις

$$p_i^{(k+1)} = \sum_{j=1}^n \tau_i(\mathbf{y}_j; \Psi^{(k)}) / n \quad (4.13)$$

$$\boldsymbol{\mu}_i^{(k+1)} = \sum_{j=1}^n \tau_i(\mathbf{y}_j; \Psi^{(k)}) \mathbf{y}_j / \sum_{i=1}^n \tau_i(\Psi^{(k)}) \quad (4.14)$$

$$\boldsymbol{\Sigma}_i^{(k+1)} = \sum_{j=1}^n \tau_i(\mathbf{y}_j; \Psi^{(k)}) (\mathbf{y}_j - \boldsymbol{\mu}_i^{(k+1)}) (\mathbf{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T / \sum_{i=1}^n \tau_i(\mathbf{y}_j; \Psi^{(k)}) \quad (4.15)$$

για $i = 1, \dots, k$.

Ο αλγόριθμος σταματά την επαναληπτική διαδικασία όταν ικανοποιηθεί το παρακάτω κριτήριο

$$L(\Psi^{(k+1)}) \geq L(\Psi^{(k)}) \quad (4.16)$$

το οποίο υποδηλώνει ότι η $L(\Psi^{(k)})$ συγκλίνει σε κάποιο L^* για μια σειρά τιμών πιθανοφάνειας φραγμένες άνω. Τα βήματα Ε και Μ εναλλάσσονται έως ότου η πιθανοφάνεια (ή οι εκτιμήσεις των παραμέτρων) αλλάζουν κατά μια πολύ μικρή ποσότητα (συνήθως προκαθορισμένη) στην περίπτωση της σύγκλισης.

Πίνακας 4.1: Ο αλγόριθμος EM για ανάλυση κατά συστάδες στην περίπτωση μίξης κανονικών κατανομών.

Θέτουμε αρχικές για το z_{ik} (βλέπε παράγραφο 4.7)

Επανέλαβε

Μ-βήμα: Μεγιστοποίησε την (4.11) δοθέντος του \hat{z}_{ik}

$$n_k \leftarrow \sum_{i=1}^n \hat{z}_{ik}$$

$$\hat{p}_k \leftarrow \frac{n_k}{n}$$

$$\hat{\mu}_k \leftarrow \frac{\sum_{i=1}^n z_{ik} \mathbf{y}_i}{n_k}$$

$\hat{\Sigma}_k$: εξαρτάται από το μοντέλο

(βλέπε παράγραφο 4.5)

Ε-βήμα: Υπολόγισε το \hat{z}_{ik} δοθέντων των τιμών που εκτιμήθηκαν στο βήμα Μ

$$z_{ik} \leftarrow \frac{\hat{p}_i f(\mathbf{y}_j; \hat{\mu}_i, \hat{\Sigma}_i)}{\sum_{h=1}^k \hat{p}_h f(\mathbf{y}_j; \hat{\mu}_h, \hat{\Sigma}_h)}$$

Μέχρι να ικανοποιηθεί το κριτήριο σύγκλισης (4.16)

Ο αλγόριθμος EM για την ανάλυση κατά συστάδες έχει μια σειρά από περιορισμούς. Πρώτον, ο ρυθμός σύγκλισης μπορεί να είναι πολύ αργός. Το τελευταίο, όμως, δεν φαίνεται να είναι πρόβλημα στην πράξη για μίξεις που διαχωρίζονται καλά όταν κάποιος ξεκινήσει με λογικές τιμές. Δεύτερον, ο αριθμός των δεσμευμένων πιθανοτήτων για κάθε παρατήρηση είναι ίσος με τον αριθμό συνιστώσων στην μίξη πράγμα το οποίο κάνει τον αλγόριθμο μη

πρακτικό για μοντέλα με μεγάλο αριθμό συνιστωσών. Επίσης η συνάρτηση πιθανοφάνειας τείνει να έχει πολλά τοπικά μέγιστα για τις μίξεις με κανονικές κατανομές, με αποτέλεσμα ο αλγόριθμος να συγκλίνει σε κάποιο τοπικό και όχι στο ολικό μέγιστο. Τέλος, ο EM καταρρέει όταν ο πίνακας συνδιακυμάνσεων για μια ή περισσότερες συνιστώσες γίνει μοναδιαίος (*singular*) ή σχεδόν μοναδιαίος. Γενικά η διαδικασία δεν μπορεί να προχωρήσει αν οι συνιστώσες περιέχουν μόνο λίγες παρατηρήσεις ή αν οι παρατηρήσεις που περιέχουν είναι συγγραμμικές. Αν ο EM, για ένα μοντέλο που έχει ένα συγκεκριμένο αριθμό συνιστωσών, χρησιμοποιηθεί για μια μίξη στην οποία στην πραγματικότητα υπάρχουν λιγότερες ομάδες, τότε μπορεί να καταρρεύσει λόγω του ότι ο πίνακας συνδιακύμανσης μπορεί να γίνει μοναδιαίος.

Για να αποφευχθούν τα πιο πάνω προβλήματα υπάρχει αρχικά ανάγκη να ελέγχουμε το σχετικό μέγεθος των προσαρμοσμένων ποσοστών μίξης (*fitted mixing proportions*) και τον πίνακα διακυμάνσεων τόσο των συνιστωσών για τις μονομεταβλητές παρατηρήσεις όσο και των γενικευμένων πινάκων διακύμανσης των συνιστωσών για να εξακριβώσουμε τα ψεύτικα τοπικά μέγιστα. Επίσης υπάρχει ανάγκη να ελέγχουμε τις ευκλείδειες αποστάσεις ανάμεσα στους μέσους των προσαρμοσμένων συνιστωσών, προκειμένου να δούμε αν οι συστάδες που προσαρμόσαμε αντιπροσωπεύουν μια πραγματική διαχώριση ανάμεσα στους μέσους ή αν προκύπτουν επειδή μια ή περισσότερες συστάδες πέφτουν σε έναν υποχώρο του αρχικού χώρου.

Τέλος θα θέλαμε να αναφέρουμε ότι έχουν προταθεί αρκετές παραλλαγές του EM αλγορίθμου. Αυτές περιλαμβάνουν τους *στοχαστικούς* (stochastic) EM ή SEM αλγόριθμους, όπου το \hat{z}_{ik} προσομοιώνεται παρά εκτιμάται στο βήμα E (Broniatowski, Celeux (1984) και Celeux και Diebolt (1985)), και τους αλγόριθμους *ταξινόμησης* (classification), EM ή CEM, (Celeux και Govaert (1992)) όπου μετατρέπουν το \hat{z}_{ik} από το βήμα E σε μια διακριτή κατάταξη πριν εφαρμόσουν το βήμα M. Ο αλγόριθμος k-Means μπορεί να αποδειχθεί ότι είναι μια ειδική κατηγορία του αλγορίθμου CEM που αφορά ένα ομοιόμορφο σφαιρικό Γκαουσιανό μοντέλο με $\Sigma_k = \lambda I$.

4.5 ΕΜ Αλγόριθμος για Ανάλυση κατά Συστάδες

Στην προηγούμενη παράγραφο περιγράψαμε την χρήση του ΕΜ αλγόριθμου στην περίπτωση της μίξης κατανομών. Ο λόγος της αναλυτικής περιγραφής είναι ότι η ίδια τεχνική μπορεί να χρησιμοποιηθεί και στην ανάλυση κατά συστάδες. Δηλαδή, ένας τρόπος για να δούμε το πρόβλημα της μίξης κατανομών πιθανοθεωριτικά είναι να θεωρήσουμε την κατανομή του δείγματος μια κατανομή μίξης, όπου οι συνιστώσες της μίξης δεν αντιπροσωπεύουν τίποτα άλλο παρά τις συστάδες που ψάχνουμε, ενώ p_i εκφράζει την πιθανότητα η παρατήρηση να ανήκει στην συστάδα i (τώρα η κάθε παρατήρηση δεν ανήκει αυστηρά σε μια μοναδική συστάδα, αλλά σε κάθε συστάδα με πιθανότητα p_i) (Fraley C. και Raftery A.E. (1998), Banfield J.D. και Raftery A.E. (1993), McLachlan G.J., Peel D. και Basford K.E. (2002)).

4.6 Επιτρέποντας τον Προσανατολισμό και το Μέγεθος να Μεταβάλλεται μεταξύ των Συστάδων

Οι συστάδες που προκύπτουν από τη χρήση μίξεων κανονικών κατανομών είναι ελλειψοειδείς με κέντρο τους μέσους μ_k . Ο πίνακας συνδιακύμανσης Σ_k καθορίζει τα υπόλοιπα γεωμετρικά χαρακτηριστικά των συστάδων.

Οι Banfield και Raftery (1993) ανέπτυξαν ένα πλαίσιο για ανάλυση κατά συστάδες με χρήση μοντέλου, παραμετροποιώντας τον πίνακα συνδιακύμανσης με βάση τη φασματική του ανάλυση (eigenvalue decomposition)

$$\Sigma_k = D_k \Lambda_k D_k^T \quad (4.17)$$

όπου D_k είναι ο ορθογώνιος πίνακας ιδιοδιανυσμάτων, Λ_k ο διαγώνιος πίνακας με τις ιδιοτιμές του Σ_k στην κυρία διαγώνιο. Ο προσανατολισμός των κυρίων συνιστώσων του Σ_k καθορίζεται από τον D_k , ενώ ο Λ_k καθορίζει το μέγεθος και το σχήμα των πυκνοτήτων των ισούψών καμπυλών (contours).

Συνήθως γράφουμε τον Λ_k ως $\Lambda_k = \lambda_k A_k$, με λ_k την πρώτη ιδιοτιμή του πίνακα Σ_k , $A_k = \text{diag}\{a_{1k} \dots a_{pk}\}$ και $1 = a_{1k} \geq \dots \geq a_{pk} > 0$. Οπότε, ο D_k καθορίζει τον προσανατολισμό του χώρου p διαστάσεων παρά τον αριθμού των στοιχείων που περιέχει. Αν τα a_{jk} είναι όμοιου μεγέθους τότε η k -οστή συστάδα θα τείνει να είναι σφαιρική στις p διαστάσεις, ενώ αν το $a_{2k} \ll 1$ θα είναι συγκεντρωμένη σε μια γραμμή και αν $a_{2k} \approx 1$ και $a_{3k} \ll 1$ θα είναι συγκεντρωμένη σε ένα διδιάστατο επίπεδο (plane) στον χώρο p -διαστάσεων, και ου το καθεξής.

Με βάση την (4.17) μπορούμε να επιτρέψουμε τις συστάδες να διαφέρουν σε μέγεθος και σχήμα κάνοντας κάποιες υποθέσεις για τον πίνακα συνδιακυμάνσεων. Ο Πίνακας 4.2 δείχνει αναλυτικά τα γεωμετρικά χαρακτηριστικά των συστάδων για διάφορες παραμετροποιήσεις του πίνακα συνδιακυμάνσεων. Στην περίπτωση όπου ο πίνακας Σ_k ισούται με λI οδηγεί στο κριτήριο των ελαχίστων τετραγώνων και οι συστάδες προκύπτουν σφαιρικές και έχουν ίδιους όγκους. Όταν ο Σ_k ισούται με (κοινό για κάθε k) λDAD οι συστάδες έχουν το ίδιο σχήμα, όγκο και προσανατολισμό, ενώ όταν ο Σ_k ισούται με $\lambda D_k A D_k$ το μόνο που διαφέρει είναι ο προσανατολισμός.

Τελειώνοντας θα θέλαμε να αναφέρουμε ότι η παραμετροποίηση του πίνακα συνδιακυμάνσεων παίζει σημαντικό ρόλο στην ανάλυση κατά συστάδες με χρήση πιθανοθεωρητικού μοντέλου. Αυτό συμβαίνει διότι, όπως αναφέραμε στην προηγούμενη ενότητα, όταν ο πίνακας συνδιακύμανσης δεν έχει κανέναν περιορισμό (unconstrained) τότε συχνά εμφανίζεται το πρόβλημα του μοναδιαίου πίνακα συνδιακύμανσης. Προκειμένου να αποφευχθεί το πρόβλημα χρειάζεται να κάνουμε υποθέσεις για την μορφή του πίνακα συνδιακύμανσης, το οποίο οδηγεί σε περιορισμούς στα γεωμετρικά χαρακτηριστικά των συστάδων. Αναλυτικά όλοι οι δυνατοί περιορισμοί δίνονται στον Πίνακα 4.2.

Πίνακας 4.2: Παραμετροποίηση του πίνακα συνδιακύμανσης Σ_k στο Γκαουσιανό μοντέλο κι η γεωμετρική ερμηνεία του (Banfield J.D. και Raftery A.E. (1993)).

Σ_k	Κατανομή	Όγκος	Σχήμα	Προσανατολισμός
λI	Σφαιρική	Ίδιος	Ίδιο	NA
$\lambda_k I$	Σφαιρική	Ευμετάβλητος	Ίδιο	NA
λDAD	Ελλειψοειδής	Ίδιος	Ίδιο	Ίδιος
$\lambda_k D_k A_k D_k$	Ελλειψοειδής	Ευμετάβλητος	Ευμετάβλητο	Ευμετάβλητος
$\lambda D_k AD_k$	Ελλειψοειδής	Ίδιος	Ίδιο	Ευμετάβλητος
$\lambda_k D_k AD_k$	Ελλειψοειδής	Ευμετάβλητος	Ίδιο	Ευμετάβλητος

4.7 Εκτίμηση Κατάλληλου Αριθμού Συστάδων

Ως τώρα ασχοληθήκαμε με την εκτίμηση των παραμέτρων των κανονικών κατανομών θεωρώντας ότι γνωρίζουμε τον αριθμό των συστάδων. Το ερώτημα που τίθεται τώρα είναι το πώς θα μπορούσε κάποιος να εξετάσει ότι ο αριθμός των συστάδων που προσάρμοσε είναι πράγματι η καλύτερη λύση. Μήπως ένα μοντέλο με λιγότερες ή περισσότερες συστάδες προσαρμόζει καλύτερα τα δεδομένα; Το πιο πάνω πρόβλημα, δηλαδή του να εξεταστεί ο κατάλληλος αριθμός συστάδων, είναι αρκετά δύσκολο και δεν έχει επιλυθεί πλήρως. Η στατιστική δοκιμασία που βασίζεται στο λόγο των πιθανοφανειών δεν είναι κατάλληλη σε αυτή την περίπτωση, διότι το $-2 \ln \lambda$ δεν ακολουθεί τη συνήθη κατανομή χ^2 , με βαθμούς ελευθερίας ίσους με τη διαφορά των παραμέτρων στις δύο υποθέσεις. Παρόλα αυτά, θα παρουσιάσουμε δύο διαφορετικούς τρόπους, οι οποίοι προσπαθούν να δώσουν λύση.

Ο πρώτος τρόπος αφορά μια διαφοροποίηση της παραπάνω δοκιμασίας του λόγου των πιθανοφανειών, η οποία προτάθηκε από τον Wolfe (1971). Ο τελευταίος πρότεινε ότι η κατανομή κάτω από τη μηδενική υπόθεση του $-2 \ln \lambda$, η οποία διερευνά την υπόθεση ότι τα δεδομένα προέρχονται από μια

μίξη κατανομών με g_1 συνιστώσες έναντι της εναλλακτικής ότι αυτά προέρχονται από μια μίξη με g_2 συνιστώσες, μπορεί να προσεγγιστεί από την

$$-2c \ln \lambda \sim \chi_d^2 \quad (4.18)$$

όπου οι βαθμοί ελευθερίας, d , είναι ίσοι με το διπλάσιο της διαφοράς στον αριθμό των παραμέτρων στις δύο υποθέσεις, με τα ποσοστά μίξης να μην καταμετρούνται στις παραμέτρους. Ο Wolfe πρότεινε την ακόλουθη τιμή για το c

$$c = (n - 1 - p - \frac{1}{2} g_2) / n \quad (4.19)$$

Ο δεύτερος τρόπος αφορά μια Μπεϋζιανή προσέγγιση με χρήση του κριτηρίου BIC (Bayesian Information Criterion), ο οποίος προτάθηκε από τους Fraley και Raftery (1998). Το κριτήριο υπολογίζει το διπλάσιο του παράγοντα Bayes –τις εκ των υστέρων συμπληρωματικές πιθανότητες (odds) ενός μοντέλου έναντι κάποιου άλλου, υποθέτοντας ότι κανένα δεν προσαρμόζει καλύτερα τα δεδομένα από το άλλο εκ των προτέρων. Δηλαδή,

$$2 \ln p(x|M) + \text{σταθερά} \approx 2L_M(x, \hat{\theta}) - m_M \ln(n) \equiv BIC \quad (4.20)$$

όπου $p(x|M)$ είναι η πιθανοφάνεια των δεδομένων για το μοντέλο M , $L_M(x, \hat{\theta})$ είναι η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια της μίξης για το μοντέλο και m_M είναι ο αριθμός των ανεξάρτητων παραμέτρων προς εκτίμηση στο μοντέλο. Ο αριθμός των συστάδων δεν θεωρείται ανεξάρτητη παράμετρος για τον υπολογισμό του κριτηρίου. Αν τα μοντέλα είναι ισοδύναμα εκ των προτέρων, τότε το $p(x|M)$ είναι ανάλογο με την εκ των υστέρων πιθανότητα ότι τα δεδομένα προσαρμόζονται καλύτερα από το μοντέλο M . Συνεπώς, όσο μεγαλύτερη η τιμή του BIC, τόσο περισσότερα τα στοιχεία υπέρ του μοντέλου M .

Η προσαρμογή ενός μοντέλου μίξεων σε ένα συγκεκριμένο σετ δεδομένων μπορεί να βελτιωθεί (και η πιθανοφάνεια να αυξηθεί) μόνο αν περισσότεροι όροι προστεθούν σε αυτό. Στο κριτήριο BIC ένας όρος προστίθεται στην λογαριθμοποιημένη πιθανοφάνεια έτσι ώστε η τελευταία να μπορεί να μεγιστοποιηθεί για περισσότερο φειδωλές παραμετροποιήσεις και μικρότερο αριθμό ομάδων από την λογαριθμοποιημένη πιθανοφάνεια. Επίσης,

το BIC μπορεί να χρησιμοποιηθεί για να συγκρίνει μοντέλα με διαφορετικές παραμέτρους, διαφορετικό αριθμό συνιστωσών, κλπ. Ένας συμβατικός τρόπος για την αξιολόγηση των διαφορών BIC, είναι ο παρακάτω. Διαφορές μικρότερες του 2 αντιστοιχούν σε μικρή ένδειξη ότι το μοντέλο που προσαρμόσαμε είναι το σωστό, διαφορές μεταξύ 2 και 6 σε θετική ένδειξη, ενώ διαφορές μεγαλύτερες του 10 σε ισχυρή ένδειξη υπέρ του μοντέλου.

Τέλος, πρέπει να επισημάνουμε ότι υπάρχουν και άλλα τεστ (κριτήρια) για την εκτίμηση του κατάλληλου αριθμού συστάδων (π.χ. το AWE, AIC (Akaike's Information Criterion), CAIC (consistent Akaike's Information Criterion)).

4.7 Στρατηγική για Ανάλυση σε Συστάδες με Χρήση Μίξης Κανονικών Κατανομών

Στην πράξη οι μέθοδοι ανάλυσης σε συστάδες ιεραρχικής συσσωμάτωσης βασιζόμενοι σε πιθανοφάνεια με κανονικούς (γκαουσιανούς) όρους συχνά δίνουν καλούς, αλλά όχι και βέλτιστους διαμερισμούς. Ο αλγόριθμος EM μπορεί να βελτιώσει τις διαμερίσεις αν οι αρχικές τιμές που θα οριστούν είναι κοντά στην βέλτιστη τιμή.

Μια στρατηγική η οποία δίνει αρκετά καλά αποτελέσματα είναι η εξής:

- Καθόρισε ένα μέγιστο αριθμό συστάδων (M) και ένα σετ υποψηφίων παραμετροποιήσεων του Γκαουσιανού μοντέλου. Γενικά το M πρέπει όσο το δυνατόν πιο μικρό
- Πραγματοποίησε ανάλυση κατά συστάδες με τη μέθοδο της ιεραρχικής συσσωμάτωσης για το Γκαουσιανό μοντέλο χωρίς περιορισμούς και απέκτησε τις αντίστοιχες κατατάξεις για περισσότερα από M ομάδες
- Τρέξε τον EM αλγόριθμο για κάθε παραμετροποίηση και για κάθε αριθμό συστάδων $2, \dots, M$ ξεκινώντας από την ταξινόμηση της ιεραρχικής ανάλυσης κατά συστάδες
- Υπολόγισε το BIC ή κάποια άλλα κριτήρια για την εύρεση του μοντέλου με τον βέλτιστο αριθμό συστάδων. Στην περίπτωση χρήσης του BIC κάνε το γράφημα των τιμών του για κάθε μοντέλο. Το πρώτο

τοπικό μέγιστο υποδεικνύει το σωστό μοντέλο (μέθοδο παραμετροποίησης και βέλτιστο αριθμό συστάδων).

Είναι σημαντικό να αποφεύγεται η εφαρμογή της παραπάνω διαδικασίας για μεγαλύτερο αριθμό συνιστώσων από ότι είναι απαραίτητο. Ένας λόγος είναι για να ελαχιστοποιηθεί η υπολογιστική προσπάθεια, οι άλλοι αφορούν αυτούς που αναφέραμε στην παράγραφο 4.3. Μια μέθοδος, η οποία δουλεύει καλά στην πράξη, είναι η επιλογή του αριθμού των συστάδων που αντιστοιχούν στο πρώτο τοπικό μέγιστο, σε οποιαδήποτε παραμετροποίηση εμφανιστεί αυτό.

4.8 Μίξεις για Κατηγορικά Δεδομένα

Έως τώρα ασχοληθήκαμε με μίξεις πολυμεταβλητών κανονικών κατανομών, οι οποίες δεν είναι φυσικά κατάλληλες για την περίπτωση κατηγορικών δεδομένων. Στην περίπτωση αυτή κυρίως χρησιμοποιούνται πυκνότητες πολυμεταβλητών Bernoulli κατανομών, οι οποίες υποθέτουν ότι οι κατηγορικές μεταβλητές είναι ανεξάρτητες μεταξύ τους, η ονομαζόμενη δεσμευμένη υπόθεση ανεξαρτησίας (Everitt B.S. (1993)).

Αναλυτικότερα, ας υποθέσουμε ότι υπάρχουν g ομάδες στα δεδομένα και ότι στην ομάδα i , το διάνυσμα θ_i δίνει την πιθανότητα ότι

$$\Pr(x_{ij} = 1 \text{ γκρουπ } i) = \theta_{ij} \quad (4.21)$$

όπου το x_{ij} είναι η τιμή η οποία παίρνει η μεταβλητή j στην ομάδα i . Από την υπόθεση της δεσμευμένης ανεξαρτησίας συνεπάγεται ότι η πιθανότητα ενός παρατηρούμενου διανύσματος των σκορ, \mathbf{x} , στην ομάδα i δίνεται από την σχέση

$$f(\mathbf{x} / \text{γκρουπ } i) = \prod_{j=1}^p \theta_{ij}^{x_{ij}} (1 - \theta_{ij})^{1-x_{ij}} \quad (4.22)$$

Αν οι αναλογίες κάθε ομάδας στον πληθυσμό είναι p_1, p_2, \dots, p_g , τότε η μη δεσμευμένη υπόθεση ανεξαρτησίας της παρατήρησης \mathbf{x} δίνεται από την μίξη

$$\Pr(\mathbf{x}) = \sum_{i=1}^g p_i \prod_{j=1}^p \theta_{ij}^{x_{ij}} (1 - \theta_{ij})^{1-x_{ij}} \quad (4.23)$$

Δεν θα επεκτείνουμε την ανάλυση στην περίπτωση κατηγορικών δεδομένων. Περισσότερες λεπτομέρειες για την εκτίμηση των παραμέτρων μέσω μεγιστοποίησης της πιθανοφάνειας και δημιουργία των ομάδων θεωρώντας τις εκτιμώμενες εκ των υστέρων πιθανότητες παρατίθενται στο άρθρο των McLachlan και Basford (1988).



ΚΕΦΑΛΑΙΟ 5

ΕΦΑΡΜΟΓΗ ΑΝΑΛΥΣΗΣ ΚΑΤΑ ΣΥΣΤΑΔΕΣ ΣΕ ΔΥΟ ΣΕΤ ΔΕΔΟΜΕΝΩΝ - ΣΥΓΚΡΙΣΗ ΙΕΡΑΡΧΙΚΩΝ ΜΕΘΟΔΩΝ ΚΑΙ ΑΝΑΛΥΣΗΣ ΚΑΤΑ ΣΥΣΤΑΔΕΣ ΜΕ ΧΡΗΣΗ ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

5.1 Εισαγωγή

Στο κεφάλαιο αυτό θα χρησιμοποιήσουμε τις μεθόδους ανάλυσης κατά συστάδες που περιγράψαμε στα προηγούμενα κεφάλαια σε δυο σετ δεδομένων. Το ένα αφορά στοιχεία οικονομικά και δημογραφικά 25 χωρών (Econ Data). Το άλλο αποτελείται από δύο μεταβλητές που αφορούν την διάρκεια των εκρήξεων και το χρόνο αναμονής ανάμεσα σε δύο εκρήξεις του θερμοπίδακα Old Faithful. Στα παραπάνω σετ επιχειρήσαμε να δημιουργήσουμε ομάδες και με τη χρήση ιεραρχικών μεθόδων, καθώς και με τη χρήση πιθανοθεωρητικού μοντέλου. Η στατιστική ανάλυση πραγματοποιήθηκε με το στατιστικό πακέτο S-Plus.

5.2 Econ Data

Στον Πίνακα 5.1 παραθέτουμε τις τιμές 5 δημογραφικών και οικονομικών δεικτών ενός δείγματος 25 χωρών για το έτος 1990 όπως καταγράφηκαν στον Ετήσιο Βιβλίο των Ηνωμένων Εθνών. Οι δείκτες αφορούν τον εκατοστιαίο ετήσιο ρυθμό ανάπτυξης του πληθυσμού (increase), τον αναμενόμενο χρόνο ζωής (life expectancy) (Life), τον ρυθμό θνησιμότητας των νεογνών ανά 1000 (IMR), το συνολικό ρυθμό γονιμότητας (TFR) και τις ακαθάριστες οικιακές εισπράξεις (Gross Domestic Product) ανά πρωτεύουσα σε αμερικάνικα δολάρια (GDP).

Σκοπός της ανάλυσης είναι να δημιουργηθούν ομάδες κρατών με κοινά χαρακτηριστικά με βάση την ανάπτυξή τους, όπως υποδεικνύουν οι παραπάνω 5 δείκτες. Σε γενικές γραμμές οι ανεπτυγμένες χώρες χαρακτηρίζονται από

χαμηλό ρυθμό ανάπτυξης, υψηλό αναμενόμενο χρόνο ζωής, χαμηλό ρυθμό θνησιμότητας, χαμηλό ρυθμό γονιμότητας και υψηλό GDP. Ήταν λογικό κανείς, λοιπόν να περιμένει διαχωρισμό των αναπτυγμένων και λιγότερο αναπτυγμένων χωρών σε διαφορετικές ομάδες.

Πίνακας 5.1 Econ Data

	Increase	Life	IMR	TFR	GDP
Albania	1.2	69.2	30	2.9	659.91
Argentina	1.2	68.6	24	2.8	4343.04
Australia	1.1	74.7	7	1.9	17529.98
Austria	1	73	7	1.5	20561.88
Benin	3.2	45.9	86	7.1	398.21
Bolivia	2.4	57.7	75	4.8	812.19
Brazil	1.5	64	58	2.9	3219.22
Cambodia	2.8	50.1	116	5.3	97.39
China	1.1	66.7	44	2	341.31
Colombia	1.7	66.4	37	2.7	1246.87
Croatia	-1.5	67.1	9	1.7	5400.66
ElSalvador	2.2	63.9	46	4	988.58
France	0.4	73	7	1.7	21076.77
Greece	0.6	75	10	1.4	6501.23
Guatemala	2.9	62.4	48	5.4	831.81
Iran	2.3	67	36	5	9129.34
Italy	-0.2	74.2	8	1.3	19204.92
Malawi	3.3	45	143	7.2	229.01
Netherlands	0.7	74.4	7	1.6	18961.9
Pakistan	3.1	60.6	91	6.2	385.59
PapuaNG	1.9	55.2	68	5.1	839.03
Peru	1.7	64.1	64	3.4	1674.15
Romania	-0.5	66.6	23	1.5	1647.97
US	1.1	72.5	9	2.1	21965.08
Zimbabwe	4.4	52.4	67	5	686.75

Προκειμένου να διερευνήσουμε την πιο πάνω εικασία εφαρμόσαμε αρχικά τις ιεραρχικές μεθόδους single linkage, complete linkage και average linkage. Στη συνέχεια κάναμε χρήση πιθανοθεωρητικού μοντέλου με πίνακα διακύμανσης- συνδιακύμανσης S,S* και σφαιρικό (με ευμετάβλητο μέγεθος) και με τη μέθοδο του Ward (trace). Ο πίνακας S (Murtagh και Raftery (1984))

αντιστοιχεί στον πίνακα συνδιακύμανσης με κατανομή ελλειψοειδούς, προσανατολισμό ευμετάβλητο και σταθερό όγκο και σχήμα, ενώ ο S* σε εκείνον με κατανομή και πάλι ελλειψοειδούς και όγκο και προσανατολισμό ευμετάβλητο. Ο σφαιρικός πίνακας διακύμανσης-συνδιακύμανσης (Banfield και Raftery (1992)) έχει κατανομή σφαιρική και όγκο ευμετάβλητο, ενώ ο πίνακας του Ward (1963) έχει απλά σφαιρική κατανομή και σταθερό όγκο και σχήμα (σφαιρικό Γκαουσιανό μοντέλο).

5.2.1 Ιεραρχικές Μέθοδοι

Τρεις μέθοδοι ιεραρχικής ομαδοποίησης χρησιμοποιήθηκαν για την εύρεση ομάδων μεταξύ των κρατών: ο αλγόριθμος ατομικής σύνδεσης (single linkage), ο αλγόριθμος πλήρους σύνδεσης (complete linkage) και ο αλγόριθμος group-average. Χρησιμοποιήθηκε πίνακας αποστάσεων (Ευκλείδιες αποστάσεις) σε μη τυποποιημένα δεδομένα. Πιο αναλυτικά, για τον αλγόριθμο group-average η απόσταση μεταξύ δυο συστάδων είναι ο μέσος όρος των αποστάσεων των στοιχείων της μιας συστάδας με τα στοιχεία της άλλης, για τον αλγόριθμο ατομικής σύνδεσης είναι η μικρότερη απόσταση μεταξύ ενός στοιχείου της μιας συστάδας από το στοιχείο της άλλης, ενώ ο αλγόριθμος πλήρους σύνδεσης χρησιμοποιεί τη μεγαλύτερη απόσταση.

Τα δενδρογράμματα των αντίστοιχων μεθόδων απεικονίζονται στο Γράφημα 5.1. Το δενδρόγραμμα είναι ένα πολύτιμο οπτικό εργαλείο για τις ιεραρχικές μεθόδους μιας και απεικονίζει την ιστορία της ομαδοποίησης. Ενώνει τις χώρες με μια γραμμή και αυτό επαναλαμβάνεται σε κάθε βήμα, έτσι ώστε όλες οι χώρες να είναι ενωμένες. Στον ένα άξονα έχουμε τις ομάδες, ενώ στον άλλο την τιμή της απόστασης με την οποία ενώσαμε τις χώρες ώστε να έχουμε μια ένδειξη πώς προχώρησε η διαδικασία. Από το Γράφημα 5.1 καταλήγουμε ότι οι χώρες κατηγοριοποιούνται σε 3 ομάδες (με υψηλό δείκτη ανάπτυξη, λιγότερο αναπτυγμένες και φτωχές).

Στον Πίνακα 5.2 δίνονται οι ομάδες που σχηματίζονται για κάθε μέθοδο. Παρατηρούμε ότι οι 3 μέθοδοι δίνουν παρόμοια αποτελέσματα. Πιο συγκεκριμένα οι μέθοδοι group-average και πλήρους σύνδεσης έχουν κοινή την πρώτη ομάδα η οποία περιλαμβάνει τις λιγότερο αναπτυγμένες χώρες, και η διαφορά τους είναι η μετακίνηση των χωρών Ελλάδα, Κροατία, Ρουμανία

από την ομάδα 3 για την μέθοδο group-average στην ομάδα 2 για τον αλγόριθμο πλήρους σύνδεσης. Δηλαδή, η μέθοδος πλήρους σύνδεσης κατατάσσει τις 3 παραπάνω χώρες στην ομάδα με τις πιο ανεπτυγμένες χώρες της Ευρώπης, των Ηνωμένων Πολιτειών και της Αυστραλίας, ενώ η μέθοδος group-average τις κατατάσσει στις χώρες με ενδιάμεση ανάπτυξη. Τέλος παρατηρούμε ότι η μέθοδος ατομικής σύνδεσης οδηγεί και αυτή σε τρεις ομάδες, οι οποίες όμως είναι ανισομεγέθεις. Η μέθοδος αυτή παρουσιάζει αρκετές αποκλίσεις από τις άλλες δύο. Παρατηρούμε, όμως, ότι η ομάδα 2 για αυτή τη μέθοδο είναι ίδια με αυτή του group-average. Έχουμε ξανά την ομάδα των ανεπτυγμένων χωρών που αποτελείται από τις Ηνωμένες Πολιτείες Αμερικής, την Αυστραλία και τις πιο ανεπτυγμένες Ευρωπαϊκές χώρες. Οι Ελλάδα, Κροατία και Ρουμανία ανήκουν σε ομάδα λιγότερο αναπτυγμένων χωρών. Τέλος σχηματίζεται μια ομάδα που αποτελείται μόνο από 4 χώρες (με την μικρότερη ανάπτυξη), ενώ οι υπόλοιπες 15 αποτελούν την τρίτη ομάδα.

Συμπεραίνουμε λοιπόν ότι και οι 3 μέθοδοι iεραρχικής ομαδοποίησης καταλήγουν σε τρεις ομάδες. Ο group-average και ο πλήρους σύνδεσης αλγόριθμος συμφωνούν περισσότερο, ενώ η ανισομεγέθης ομαδοποίηση στην οποία καταλήγει ο αλγόριθμος ατομικής σύνδεσης μας βάζει σε κάποια σκέψη σχετικά με την καταλληλότητα της συγκεκριμένης μεθόδου.

5.2.2 Ανάλυση Κατά Συστάδες με Χρήση Πιθανοθεωρητικού Μοντέλου

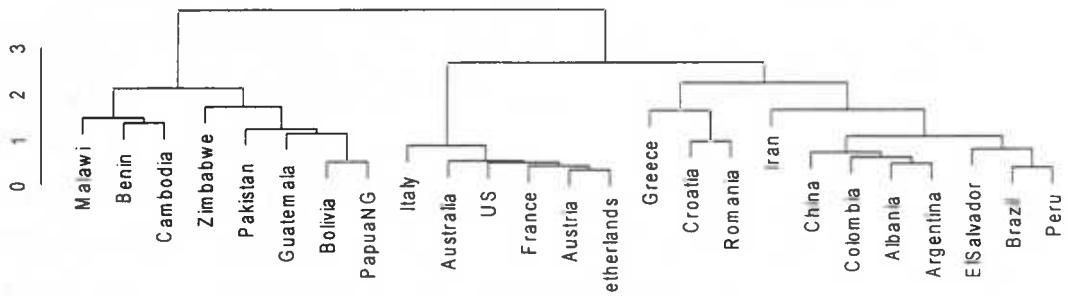
Όπως αναφέραμε και πιο πάνω για την ανάλυση των οικονομικών και δημογραφικών δεδομένων των 25 χωρών χρησιμοποιήσαμε και πιθανοθεωρητικό μοντέλο με διαφορετικό πίνακα διακύμανσης-συνδιακύμανσης κάθε φορά. Στον Πίνακα 5.3 παρατίθεται η ομαδοποίηση που προκύπτει κάθε φορά με πίνακα συνδιακύμανσης σφαιρικό, του Ward, S και S*, ενώ το Γράφημα 5.2 δίνει τα αντίστοιχα δενδρογράμματα.

Πίνακας 5.2 Ιεραρχική ομαδοποίηση σε 3 ομάδες για τα δεδομένα των 25 χωρών

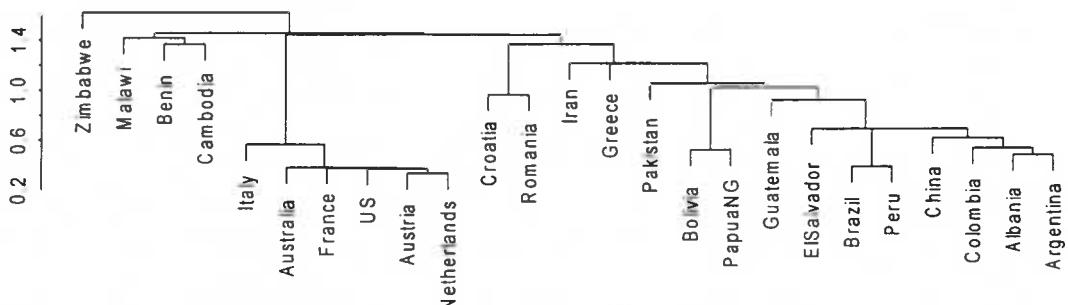
	Ομάδα 1	Ομάδα 2	Ομάδα 3	
Ατομικής Σύνδεσης	<ul style="list-style-type: none"> • Zimbabwe • Malawi • Benin • Cambodia 	<ul style="list-style-type: none"> • Italy • Australia • France • US • Austria • Netherlands 	<ul style="list-style-type: none"> • Croatia • Romania • Iran • Greece • Pakistan • Bolivia • PapuaNG • Guatemala 	<ul style="list-style-type: none"> • El Salvador • Brazil • Peru • China • Colombia • Albania • Argentina
Πλήρους Σύνδεσης	<ul style="list-style-type: none"> • Malawi • Benin • Cambodia • Zimbabwe • Pakistan • Guatemala • Bolivia • PapuaNG 	<ul style="list-style-type: none"> • Italy • Australia • France • US • Austria • Netherlands • Greece • Croatia • Romania 	<ul style="list-style-type: none"> • Iran • Albania • Argentina • China • Colombia • El Salvador • Brazil • Peru 	
Group-Average	<ul style="list-style-type: none"> • Malawi • Benin • Cambodia • Zimbabwe • Pakistan • Guatemala • Bolivia • PapuaNG 	<ul style="list-style-type: none"> • Italy • Australia • France • US • Austria • Netherlands 	<ul style="list-style-type: none"> • Greece • Croatia • Romania • Iran • China • Colombia • Albania • Argentina 	<ul style="list-style-type: none"> • El Salvador • Brazil • Peru

Γράφημα 5.1 Δενδρογράμματα των μεθόδων ιεραρχικής ομαδοποίησης

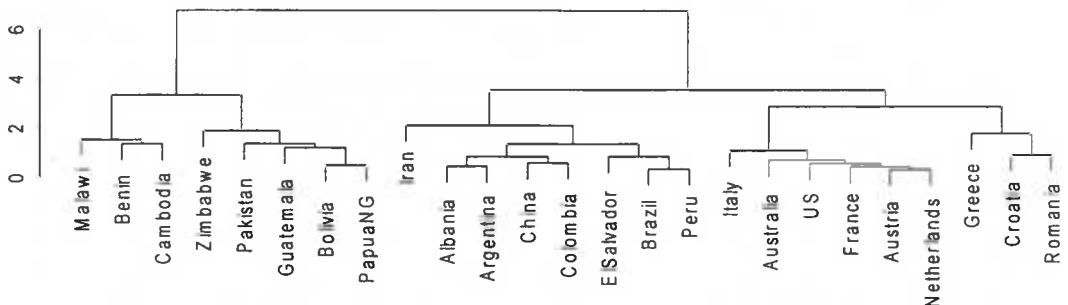
Δενδρόγραμμα της μεθόδου Group-Average



Δενδρόγραμμα της μεθόδου Ατομικής Σύνδεσης (Single Linkage)



Δενδρόγραμμα της μεθόδου Πλήρους Συνδεσης (Complete Linkage)



Από την ανάλυση προέκυψε ότι και οι τέσσερις μέθοδοι καταλήγουν σε τρεις ομαδοποιήσεις – ανεπτυγμένες, λιγότερο ανεπτυγμένες και φτωχές χώρες. Αναλυτικότερα, η μέθοδος που Ward και του σφαιρικού πίνακα συνδιακύμανσης έδωσαν τα ίδια αποτελέσματα. Η ομάδα 1 περιλαμβάνει τις ενδιάμεσα ανεπτυγμένες χώρες, η ομάδα 2 τις ευρωπαϊκές (μαζί και τις βαλκανικές που δεν έχουν τον ίδιο βαθμό ανάπτυξης με τις υπόλοιπες ευρωπαϊκές), την Αμερική και την Αυστραλία, ενώ η ομάδα 3 περιλαμβάνει τις πιο φτωχές. Ο πίνακας S και S* έδωσαν τα ίδια αποτελέσματα όσον αφορά την ομάδα 2 με τις πιο ανεπτυγμένες χώρες. Εδώ παρατηρούμε ότι

βαλκανικά κράτη δεν ανήκουν στην ομάδα αυτή, αλλά ανήκουν σε εκείνη με τις ενδιάμεσα ανεπτυγμένες χώρες. Διαφοροποίηση παρουσιάζουν οι δυο τελευταίες μέθοδοι στην κατάταξη των ενδιάμεσα ανεπτυγμένων χωρών και των πιο φτωχών κρατιδίων. Πιο συγκεκριμένα ο S^* θεωρεί ότι το Ιράν, η Βραζιλία, το Περού, το Ελ Σαλβαδόρ και η Γουατεμάλα ανήκουν στις ενδιάμεσα ανεπτυγμένες χώρες, ενώ ο S τις κατατάσσει στις φτωχές.

5.2.3 Σύγκριση των Δυο Διαφορετικών Προσεγγίσεων, Ιεραρχικών Μεθόδων και Πιθανοθεωρητικού Μοντέλου- Τελικά Συμπεράσματα

Συγκρίνοντας τις μεθόδους που εφαρμόσαμε πιο πάνω συμπεραίνουμε ότι όλες διαχωρίζουν τα κράτη σε ανεπτυγμένα, λιγότερο ανεπτυγμένα και φτωχά. Ο αλγόριθμος της ατομικής σύνδεσης φαίνεται να προσαρμόζει πιο φτωχά τα δεδομένα από τις υπόλοιπες μεθόδους. Επίσης ο αλγόριθμος πλήρους σύνδεσης δίνει τα ίδια αποτελέσματα με τον σφαιρικό πίνακα διακύμανσης και την μέθοδο του Ward. Τέλος οι αλγόριθμοι group-average, S και S^* συμφωνούν μόνο στην ομαδοποίηση των ανεπτυγμένων χωρών στην οποία δεν περιλαμβάνουν τα κράτη της βαλκανικής χερσονήσου.

5.3 Old Faithful Data

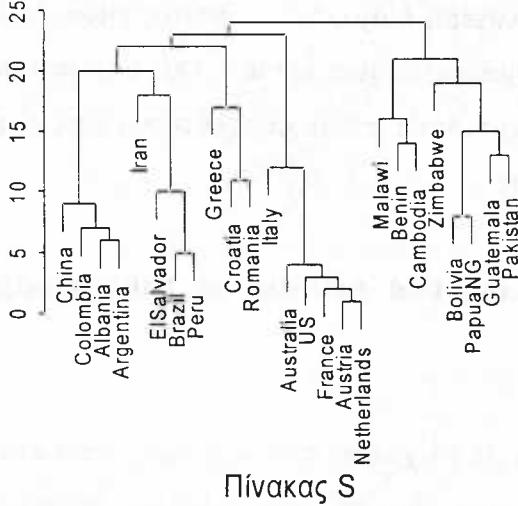
Ο Old Faithful είναι ένας θερμοπίδακας του Εθνικού Πάρκου Yellowstone της πόλης Wyoming των Ηνωμένων Πολιτειών της Αμερικής. Καθώς ο θερμοπίδακας είναι μεγάλης τουριστικής σημασίας, το πρόβλημα της πρόβλεψης του ακριβή χρόνου που θα συμβεί μια έκρηξη του θερμοπίδακα και του διαστήματος που μεσολαβεί ανάμεσα σε δυο εκρήξεις έχει απασχολήσει αρκετούς επιστήμονες. Η βάση δεδομένων του Old Faithful (Παράρτημα II) περιλαμβάνει δυο μεταβλητές, μια για τη διάρκεια των εκρήξεων (duration) και μια του χρόνου που μεσολαβεί μεταξύ δυο διαδοχικών εκρήξεων (waiting). Και σε αυτή τη βάση δεδομένων χρησιμοποιήσαμε τις ίδιες μεθόδους, ιεραρχικές και πιθανοθεωρητικό μοντέλο, με την προηγούμενη βάση. Τα πλήρη δεδομένα παρατίθενται στο Παράρτημα II.

Πίνακας 5.3 Αποτελέσματα της ανάλυσης κατά συστάδες στα οικονομικά-δημογραφικά δεδομένα με τη χρήση πιθανοθεωρητικού μοντέλου.

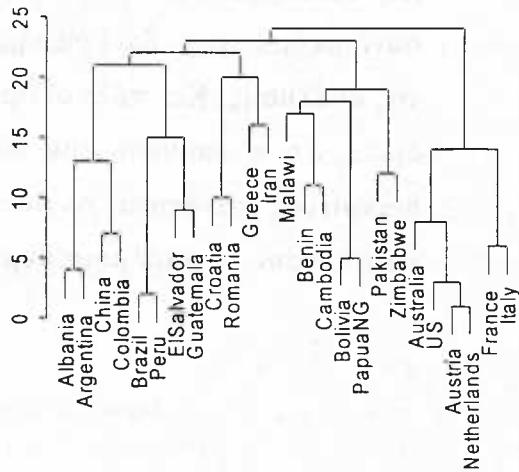
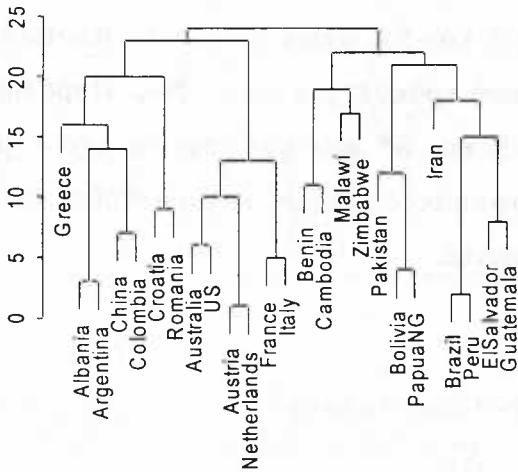
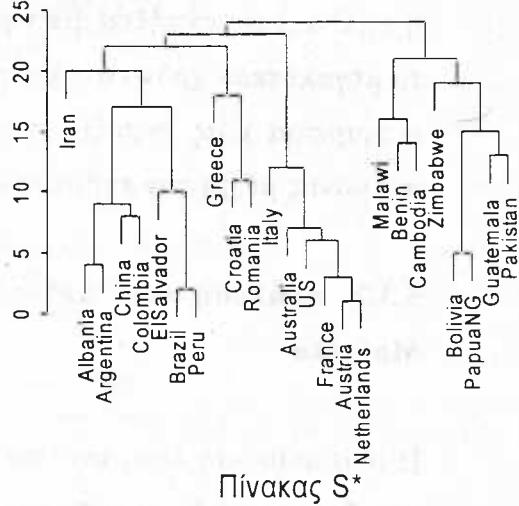
	Σφαιρικός	Ward	S	S*
Ομάδα 1	• China	• Iran	• Greece	• Albania • Greece
	• Colombia	• Albania	• Albania	• Argentina • Iran
	• Albania	• Argentina	• Argentina	• China
	• Argentina	• China	• China	• Colombia
	• Iran	• Colombia	• Colombia	• Brazil
	• El Salvador	• El Salvador	• Croatia	• Peru
	• Brazil	• Brazil	• Romania	• El Salvador
	• Peru	• Peru		• Guatemala • Croatia • Romania
Ομάδα 2	• Greece	• Greece	• Australia	• Australia
	• Croatia	• Croatia	• US	• US
	• Romania	• Romania	• Austria	• Austria
	• Italy	• Italy	• Netherland	• Netherlands
	• Australia	• Australia	s	• France
	• US	• US	• France	• Italy
	• France	• France	• Italy	
	• Austria	• Austria		
Ομάδα 3	• Malawi	• Malawi	• Benin	• Malawi
	• Benin	• Benin	• Cambodia	• Benin
	• Cambodia	• Cambodia	• Malawi	• Cambodia
	• Zimbabwe	• Zimbabwe	• Zimbabwe	• Bolivia
	• Bolivia	• Bolivia	• Pakistan	• PapuaNG
	• Papuang,	• PapuaNG	• Bolivia	• Pakistan
	• Guatemala	• Guatemala	• PapuaNG	• Zimbabwe
	• Pakistan	• Pakistan	• Iran • Brazil • Peru • El Salvador • Guatemala	

Γράφημα 5.2 Δενδρογράμματα των μεθόδων με τη χρήση πιθανοθεωρητικού μοντέλου

Σφαιρικός πίνακας διακύμανσης-συνδιακύμανσης



Μέθοδος του Ward



5.3.1 Ανάλυση των Δεδομένων του Old Faithful με Ιεραρχικές Μεθόδους

Το Γράφημα 5.3 απεικονίζει τα αποτελέσματα από την ανάλυση κατά συστάδες με τη χρήση ιεραρχικών μεθόδων. Τα αντίστοιχα δενδρογράμματα δεν δίνονται μιας και λόγω των πολλών δεδομένων δεν μπορούν να οδηγήσουν σε εξαγωγή συμπερασμάτων. Και οι τρεις μέθοδοι καταλήγουν σε τρεις ομάδες. Παρατηρούμε, όμως, ότι ο αλγόριθμος ατομικής σύνδεσης

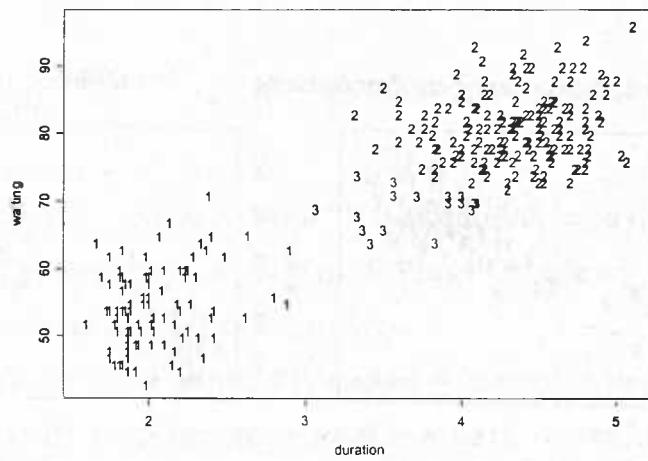
αποτυγχάνει να δημιουργήσει τις κατάλληλες ομάδες μιας και η ομάδα 3 περιλαμβάνει μόνο μια παρατήρηση. Οι δυο άλλες μέθοδοι δίνουν παρόμοια αποτελέσματα. Έτσι, η πρώτη ομάδα που δημιουργήθηκε αναφέρεται σε εκείνες τις εκρήξεις οι οποίες έχουν μικρή διάρκεια, αλλά ταυτόχρονα και το μικρότερο χρόνο αναμονής μεταξύ δυο διαδοχικών εκρήξεων. Η ομάδα 3 περιλαμβάνει τις εκρήξεις με μεγαλύτερη διάρκεια και χρόνο αναμονής, ενώ η ομάδα 2 τις εκρήξεις με τη μεγαλύτερη διάρκεια, οι οποίες όμως έχουν και το μεγαλύτερο χρόνο αναμονής. Συμπεραίνουμε, λοιπόν, πως όσο μεγαλύτερη η διάρκεια μιας έκρηξης του θερμοπίδακα τόσο μεγαλύτερος και ο χρόνος αναμονής μέχρι την επόμενη έκρηξη.

5.3.2 Ανάλυση των Δεδομένων του Old Faithful με Πιθανοθεωρητικό Μοντέλο

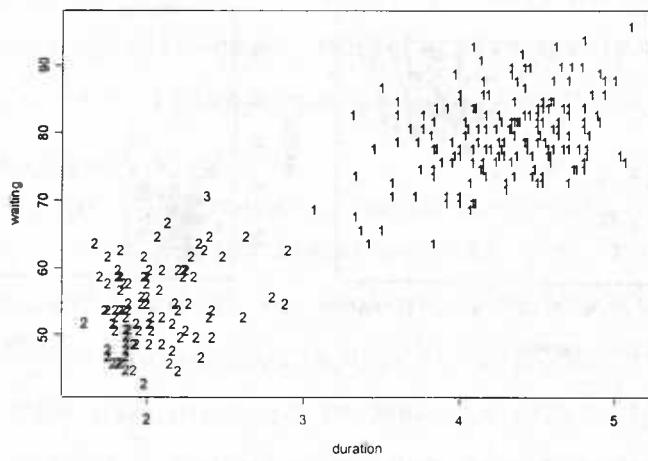
Η ίδια ανάλυση πραγματοποιήθηκε με τη χρήση πιθανοθεωρητικών μοντέλων με διαφορετικές υποθέσεις για την μορφή του πίνακα διακύμανσης-συνδιακύμανσης. Τα Γραφήματα 5.3 και 5.4 απεικονίζουν τα αποτελέσματα της ανάλυσης. Και πάλι οι ομάδες που προκύπτουν είναι τρεις. Παρατηρούμε, όμως, ότι η υπόθεση του πίνακα S και S* δεν φαίνεται να έχουν μεγάλη διακριτική ικανότητα. Αντίθετα ο σφαιρικός πίνακας και η μέθοδος του Ward καταλήγουν σε παρόμοια συμπεράσματα.

Γράφημα 5.3 Χρήση ιεραρχικών μεθόδων στα δεδομένα Old Faithful

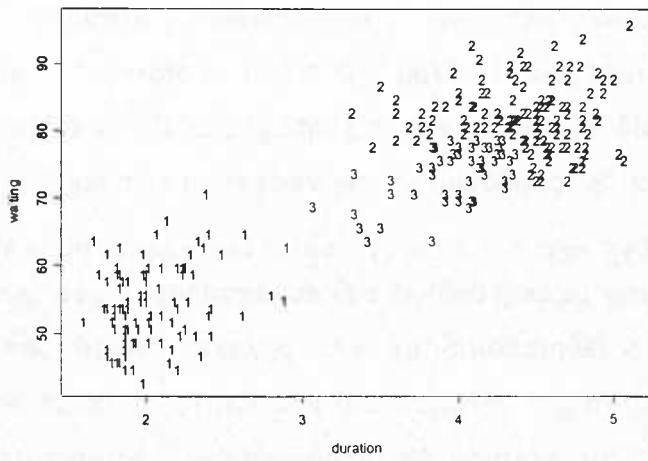
Μέθοδος Group- Average



Μέθοδος Ατομικής Σύνδεσης (Single Linkage)

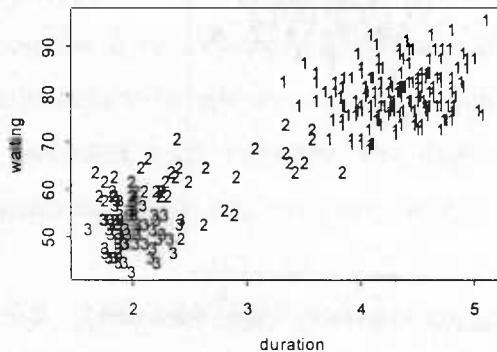


Μέθοδος Πλήρους Συνδεσης (Complete Linkage)

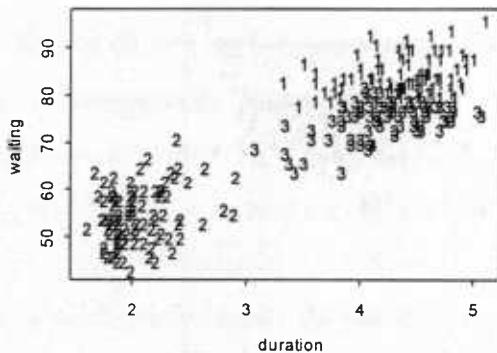


Γράφημα 5.4 Χρήση πιθανοθεωρητικών μοντέλων στα δεδομένα Old Faithful

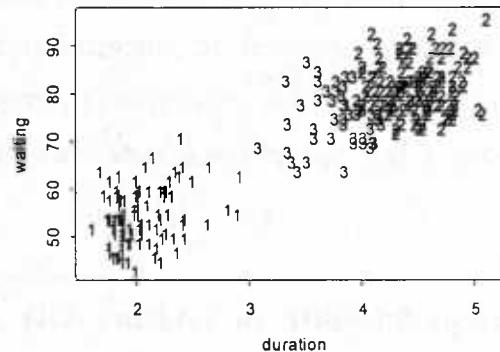
Σφαιρικός πίνακας διακύμανσης-συνδιακύμανσης



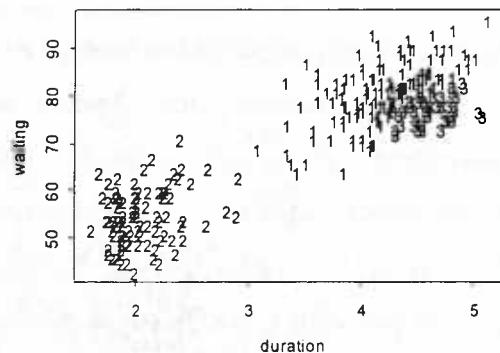
Πίνακας S



Μέθοδος του Ward



Πίνακας S*



ΠΑΡΑΡΤΗΜΑ Ι

ΛΟΓΙΣΜΙΚΟ

S-Plus Λογισμικό

Η στατιστική ανάλυση του Κεφαλαίου 5 πραγματοποιήθηκε εξ ολοκλήρου με το στατιστικό πακέτο S-Plus. Οι κύριες εντολές για την ανάλυση κατά συστάδες είναι οι παρακάτω:

- Ιεραρχικές μέθοδοι: `hclust(dist, method = "compact", sim =)`

Η εντολή `hclust` πραγματοποιεί ανάλυση κατά συστάδες σε ένα πίνακα αποστάσεων ή ομοιότητας (παράμετρος `dist` ή `sim`) με ιεραρχικές μεθόδους (παράμερος `method`). Επιλογές για ιεραρχικές μοθόδους είναι οι `compact` (αλγόριθμος πλήρους σύνδεσης),`average` (αλγόριθμος group-average) και `connected` (αλγόριθμος ατομικής σύνδεσης).

- Πιθανοθεωρητικό μοντέλο:

```
mclust(x, method = "S", shape= , workspace=100000)
```

Η παραπάνω εντολή πραγματοποιεί ανάλυση κατά συστάδες με χρήση πιθανοθεωρητικού μοντέλου και υπολογίζει ένα Μπεϋζιανό κριτήριο για την επιλογή του κατάλληλου αριθμού συστάδων. Η παράμετρος `x` αναφέρεται στον πχρ πίνακα με τα δεδομένα (δεν επιτρέπονται missing values), η παράμετρος `method` αναφέρεται στην επιλογή του κατάλληλου κριτηρίου. Πιθανές τιμές είναι οι "S", "S*", "trace" (Ward μέθοδος), "spherical", "centroid", "determinant", "weighted average link", "group average link", "complete link" ή "farthest neighbor", "single link" ή "nearest neighbor". Η παράμετρος `shape` είναι ένα διάνυσμα το οποίο καθορίζει το σχήμα των συστάδων για τις μεθόδους "S" και "S*".

- Δενδρογράμματα: `prclust(tree, labels = <<see below>>, plot=T)`

Η παράμετρος `tree` αναφέρεται σε ένα δενδρόγραμμα που αφορά ανάλυση με χρήση της `hclust` εντολής. Αν χρησιμοποιηθεί κάποια ιεραρχική μέθοδος τότε αρκεί το όνομα του διανύσματος της αντίστοιχης μεθόδου, ενώ αν χρησιμοποιηθεί πιθανοθεωρητικό μοντέλο τότε η σύνταξη είναι «όνομα διανύσματος μεθόδου με πιθανοθεωρητικό μοντέλο»\$tree. Η δε

παράμετρος `plot=T` (`True`) ορίζει ότι θα γίνει το γράφημα του δενδρογράμματος. Τέλος, η παράμετρος `label` αναφέρεται στο διάνυσμα με τα ονόματα που θέλουμε να δώσουμε στα κλαδιά του δενδρογράμματος. Αν δεν δώσουμε κάποιο συγκεκριμένο, τότε το πακέτο τοποθετεί αριθμούς.

ΠΑΡΑΡΤΗΜΑ II

OLD FAITHFUL ΔΕΔΟΜΕΝΑ



No	Duration	Waiting	No	Duration	Waiting
1	3,6	79	46	3,317	83
2	1,8	54	47	3,833	64
3	3,333	74	48	2,1	53
4	2,283	62	49	4,633	82
5	4,533	85	50	2	59
6	2,883	55	51	4,8	75
7	4,7	88	52	4,716	90
8	3,6	85	53	1,833	54
9	1,95	51	54	4,833	80
10	4,35	85	55	1,733	54
11	1,833	54	56	4,883	83
12	3,917	84	57	3,717	71
13	4,2	78	58	1,667	64
14	1,75	47	59	4,567	77
15	4,7	83	60	4,317	81
16	2,167	52	61	2,233	59
17	1,75	62	62	4,5	84
18	4,8	84	63	1,75	48
19	1,6	52	64	4,8	82
20	4,25	79	65	1,817	60
21	1,8	51	66	4,4	92
22	1,75	47	67	4,167	78
23	3,45	78	68	4,7	78
24	3,067	69	69	2,067	65
25	4,533	74	70	4,7	73
26	3,6	83	71	4,033	82
27	1,967	55	72	1,967	56
28	4,083	76	73	4,5	79
29	3,85	78	74	4	71
30	4,433	79	75	1,983	62
31	4,3	73	76	5,067	76
32	4,467	77	77	2,017	60
33	3,367	66	78	4,567	78
34	4,033	80	79	3,883	76
35	3,833	74	80	3,6	83
36	2,017	52	81	4,133	75
37	1,867	48	82	4,333	82
38	4,833	80	83	4,1	70
39	1,833	59	84	2,633	65
40	4,783	90	85	4,067	73
41	4,35	80	86	4,933	88
42	1,883	58	87	3,95	76
43	4,567	84	88	4,517	80
44	1,75	58	89	2,167	48
45	4,533	73	90	4	86



No	Duration	Waiting	No	Duration	Waiting
91	2,2	60	135	1,833	46
92	4,333	90	136	4,383	82
93	1,867	50	137	1,883	51
94	4,817	78	138	4,933	86
95	1,833	63	139	2,033	53
91	2,2	60	140	3,733	79
92	4,333	90	141	4,233	81
93	1,867	50	142	2,233	60
94	4,817	78	143	4,533	82
95	1,833	63	144	4,817	77
96	4,3	72	145	4,333	76
97	4,667	84	146	1,983	146
98	3,75	75	147	4,633	147
99	1,867	51	148	2,017	148
100	4,9	82	149	5,1	149
101	2,483	62	150	1,8	150
102	4,367	88	151	5,033	151
103	2,1	49	152	4	152
104	4,5	83	153	2,4	153
105	4,05	81	154	4,6	154
106	1,867	47	155	3,567	155
107	4,7	84	156	4	156
108	1,783	52	157	4,5	81
109	4,85	86	158	4,083	93
110	3,683	81	159	1,8	53
111	4,733	75	160	3,967	89
112	2,3	59	161	2,2	45
113	4,9	89	162	4,15	86
114	4,417	79	163	2	58
115	1,7	59	164	3,833	78
116	4,633	81	165	3,5	66
117	2,317	50	166	4,583	76
118	4,6	85	167	2,367	63
119	1,817	59	168	5	88
120	4,417	87	169	1,933	52
121	2,617	53	170	4,617	93
122	4,067	69	171	1,917	49
123	4,25	77	172	2,083	57
124	1,967	56	173	4,583	77
125	4,6	88	174	3,333	68
126	3,767	81	175	4,167	81
127	1,917	45	176	4,333	81
128	4,5	82	177	4,5	73
129	2,267	55	178	2,417	50
130	4,65	90	179	4	85
131	1,867	45	180	4,167	74
132	4,167	83	181	1,883	55
133	2,8	56	182	4,583	77
134	4,333	89	183	4,25	83

No	Duration	Waiting	No	Duration	Waiting
184	3,767	83	231	4,083	70
185	2,033	51	232	2,417	54
186	4,433	78	233	4,183	86
187	4,083	84	234	2,217	50
188	1,833	46	235	4,45	90
189	4,417	83	236	1,883	54
190	2,183	55	237	1,85	54
191	4,8	81	238	4,283	77
192	1,833	57	239	3,95	79
193	4,8	76	240	2,333	64
194	4,1	84	241	4,15	75
195	3,966	77	242	2,35	47
196	4,233	81	243	4,933	86
197	3,5	87	244	2,9	63
198	4,366	77	245	4,583	85
199	2,25	51	246	3,833	82
200	4,667	78	247	2,083	57
201	2,1	60	248	4,367	82
202	4,35	82	249	2,133	67
203	4,133	91	250	4,35	74
204	1,867	53	251	2,2	54
205	4,6	78	252	4,45	83
206	1,783	46	253	3,567	73
207	4,367	77	254	4,5	73
208	3,85	84	255	4,15	88
209	1,933	49	256	3,817	80
210	4,5	83	257	3,917	71
211	2,383	71	258	4,45	83
212	4,7	80	259	2	56
213	1,867	49	260	4,283	79
214	3,833	75	261	4,767	78
215	3,417	64	262	4,533	84
216	4,233	76	263	1,85	58
217	2,4	53	264	4,25	83
218	4,8	94	265	1,983	43
219	2	55	266	2,25	60
220	4,15	76	267	4,75	75
221	1,867	50	268	4,117	81
222	4,267	82	269	2,15	46
223	1,75	54	270	4,417	90
224	4,483	75	271	1,817	46
225	4	78	272	4,467	74
226	4,117	79			
227	4,083	78			
228	4,267	78			
229	3,917	70			
230	4,55	79			

REFERENCES

- Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles, *Journal of Royal Statistical Society A.*, 144, 419-448
- Banfield, J.D., and Raftery, A.E. (1992). Ice floe identification in satellite images using mathematical morphology and clustering about principle curves, *Journal of the American Statistical Association*, 87, 7-16
- Banfield, J.D., and Raftery, A.E. (1993). Model- based Gaussian and non-Gaussian clustering, *Biometrics*, 49, 803-821
- Celeux, G., and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions, *Computational Statistics and Data Analysis*, 14, 315-332
- Constantine, A.G., and Gower, J.C. (1978). Graphical representation of asymmetric matrices, *Applied Statistics*, 27, 297-304
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B.*, 39, 1, 1-38
- Everitt, B.S. (1993). *Cluster Analysis*, Edward Arnold, Ltd., London, UK.
- Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus, H., and Zubrzchi, S. (1951). Sur la liason et la division des points d'un ensemble fini, *Colloquium Mathematicum*, 2, 282-285
- Fraley, C., and Raftery, A.E. (1998). How many clusters? Which clustering method?- Answers via model-based cluster analysis, *The Computer Journal*, 41, 579-588
- G. Celeux and J. Diebold. (1985) *The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problems*. Comp. Stat. Quart., 2:73-82
- Gower, J.C. (1967). A comparison of some methods of cluster analysis, *Biometrics*, 23, 623-628
- Gower, J.C. (1971). A general coefficient of similarity and some of its properties, *Biometrics*, 27, 857-872
- Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). Data clustering: A review,



ACM Computing Surveys, 31, 3, 265-323

Jardine, N., and Sibson, R. (1971). *Mathematical Taxonomy*, John Wiley & Sons, Chichester

Johnson, S.C. (1967). Hierarchical clustering schemes, *Psychometrika*, 32, 241-254

Karlis, D. Lecture Notes, **Cluster analysis**

Lance, G.N., and Williams, W.T., (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems, *Comp. J.*, 9, 373-380

M. Broniatowski, G. Celeux, and J. Diebolt (1984). *Reconnaissance de melanges de densites par un algorithme d'apprentissage probabiliste*. In E. Diday, M. Jambu, L. Lebart, J.-P. Pages, and R. Tomassone, editors, *Data Analysis and Informatics*, III, pages 359-373. Elsevier Science

Mc Lachlan, G. (1982). *The classification and mixture maximum likelihood approaches to cluster analysis*. In *Handbook of Statistics*, 2, P.R. Krishnaiah and L.N. Kanal (eds), 199-208, Amsterdam: North Holland

Mc Lachlan, G.J., and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York

Mc Lachlan, G.J., Peel D., Basford K.E., Adams, P. (2002). The EMMIX Software for the fitting of mixtures of normal and t-components

Murtagh, F. and Raftery, A.E. (1984). Fitting straight lines to point patterns, *Pattern Recognition*, 17, 479-483

Sneath, P.H.A. (1957). The application of computers to taxonomy, *J. Gen. Microbiol.*, 17, 201-226

Sneath, P.H.A., and Sokal, R.R. (1973). *Numerical Taxonomy*, W.H. Freedman & Co., San Francisco

Symons, M.J. (1981). Clustering criteria and multivariate normal mixtures, *Biometrics*, 37, 35-43

Ward, J.H. (1963). Hierarchical grouping to optimise an objective function, *Journal of American Statistical Association*, 58, 236-244

Wolfe, J.H. (1971). *A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions*. Nava Personnel and Training Research Laboratory, Technical Bulletin, STB 72-2 (San Diego, California 92152)



