ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Δημήτρης Καρλής

Πολυμεταβλητή Στατιστική Ανάλυση

ΑΘΗΝΑ- ΜΑΙΟΣ 2003

ΕΚΔΟΣΕΙΣ ΟΙΚΟΝΟΜΙΚΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΑΘΗΝΩΝ

Πρόλογος

Η πολυμεταβλητή στατιστική ανάλυση αποτελεί ένα πολύτιμο εργαλείο στα χέρια του ερευνητή σε πολλές επιστήμες. Η διάδοση της χρήσης ηλεκτρονικών υπολογιστών και η αφθονία στατιστικών πακέτων που προσφέρουν τις τεχνικές αυτές τις έχει κάνει ένα αναπόσπαστο μέρος των περισσότερων αναλύσεων. Τα περισσότερα φαινόμενα είναι από τη φύση τους πολυμεταβλητά. Συγχρόνως η αφθονία δεδομένων που προέρχεται από αυτοματοποιημένα συστήματα συλλογής δεδομένων επιτρέπει στον ερευνητή να έχει εύκολα και γρήγορα χιλιάδες παρατηρήσεις και ίσως μερικές εκατοντάδες μεταβλητές για κάθε παρατήρηση.

Οι σημειώσεις που κρατάτε στα χέρια σας αποτελούν μια σύντομή εισαγωγή στην πολυμεταβλητή στατιστική ανάλυση. Θα πρέπει να τονιστεί ότι η προσέγγιση είναι από στατιστική οπτική και συνεπώς σε κάθε μέθοδο το θεωρητικό της υπόβαθρο συζητείτε σε μεγάλο βαθμό.

Χρειάστηκαν 3 χρόνια για να πάρουν οι σημειώσεις τη σημερινή τους μορφή. Συνεπώς θα πρέπει να ευχαριστήσω τους φοιτητές του τμήματος Στατιστικής (προπτυχιακούς και μεταπτυχιακούς) που με ποικίλα σχόλια βοήθησαν τη διαμόρφωση των σημειώσεων αυτών. Θερμές ευχαριστίες στους συναδέλφους μου για την πολύτιμη βοήθεια τους όλα αυτά τα χρόνια αλλά και πολλά σχόλια που πιστεύω πως βελτίωσαν την παρουσίαση. Ιδιαίτερες ευχαριστίες στους συναδέλφους Βασίλη Βασδέκη και Γιάννη Ντζούφρα καθώς οι δικές τους σημειώσεις αποτέλεσαν τη βάση για τα κεφάλαια 6 και 10 αντίστοιχα.

Παρά την προσπάθεια είναι βέβαιο πως υπάρχουν ακόμα κάποια τυπογραφικά λάθη. Θα βοηθήσει σημαντικά, όποιος βρει τέτοια λάθη να με ενημερώσει στέλνοντας ένα email στη διεύθυνση karlis@hermes.auab.gr ώστε σε επόμενη έκδοση να απαλειφθούν όσα λάθη παραμένουν.

Αθήνα, Μάρτιος 2003

Δημήτρης Καρλής

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφάλαιο 1. Εισαγωγή **1**
1.1 Εισαγωγή 1
1.2 Πολυμεταβλητές Μέθοδοι 4

Κεφάλαιο 2. Πολυμεταβλητή Περιγραφική Στατιστική **9**
2.1 Γραφήματα 9
2.1.1 Matrix Plot 9
2.1.2 Starplots 14
2.1.3 Bubble Plots 17
2.1.4 Glyph Plot 19
2.1.5 Τα Πρόσωπα του Chernoff 20
2.1.6 Οι Καμπύλες του Andrews 22
2.2 Πολυμεταβλητά Περιγραφικά Μέτρα 26
2.2.1 Πολυμεταβλητά Δεδομένα 26
2.2.2 Μέτρα Θέσης 28
2.2.3 Μέτρα Μεταβλητότητας 29
2.2.4 Πίνακας Συσχετίσεων R 32
2.2.5 Στοιχεία Πινάκων 36
2.2.6 Μέτρα Πολυμεταβλητής Ασυμμετρίας και Κύρτωσης 38
2.3 Τρόποι Πολυμεταβλητής Ανάλυσης 41

Κεφάλαιο 3. Πολυμεταβλητες Κατανομές **45**
3.1 Πολυμεταβλητές κατανομές 45
3.2 Πολυμεταβλητή Κανονική Κατανομή 52
3.3 Διδιάστατη κανονική κατανομή 54
3.4 Ιδιότητες της πολυμεταβλητής κανονικής κατανομής 60
3.5 Περιθώριες και δεσμευμένες κατανομές 68
3.6 Εκτίμηση παραμέτρων 74
3.7 Προσομοίωση δεδομένων από πολυμεταβλητή κανονική κατανομή 76

Κεφάλαιο 4. Δειγματοληπτικές Κατανομές **79**
4.1 Μη κεντρικές κατανομές 79
4.2 Η κατανομή Wishart 81
4.3 Ιδιότητες της κατανομής Wishart 83
4.4 Η κατανομή T2 του Hotelling 84
4.5 Η κατανομή Λάμδα του Wilks 85

Κεφάλαιο 5. Ελεγχοι Υποθέσεων **89**
5.1 Εισαγωγή 89
5.2 Έλεγχοι για ένα διάνυσμα μέσων τιμών 90

5.2.1 Γνωστός πίνακας διακύμανσης *90*
5.2.2 Άγνωστος πίνακας διακύμανσης *91*
5.3 Έλεγχος για διαφορά δύο μέσων 92
5.4 Έλεγχος Ισότητας Πινάκων Διακύμανσης 93
5.5 Συμπεράσματα 98
5.6 Έλεγχοι για την πολυμεταβλητή κανονική κατανομή 99

Κεφάλαιο 6. Πολυμεταβλητή Ανάλυση Διακύμανσης 103
6.1 Εισαγωγή 103
6.2 MANOVA ως προς έναν Παράγοντα 103
6.3 Έλεγχοι Υποθέσεων 106
6.4 Πίνακας Ανάλυσης Διακύμανσης 108
6.5 Πολυμεταβλητή Παλινδρόμηση 109
6.6 Η MANOVA ως Γραμμικό Μοντέλο 110
6.7 Άλλα Θέματα 110
6.8 Παράδειγμα (Fisher's Iris Data). 111
6.9 MANOVA με τη Χρήση του Πακέτου SPSS 113

Κεφάλαιο 7. Ανάλυση σε Κύριες Συνιστώσες 115
7.1 Εισαγωγή 115
7.2 Η Βασική Ιδέα 116
7.3 Εύρεση των Κυριών Συνιστωσών 117
7.4 Αλλαγή Κλίμακας 120
7.5 Βήματα της Ανάλυσης Σε Κύριες Συνιστώσες 123
7.5.1 Έλεγχος συσχετίσεων *123*
7.5.2 Επιλογή πίνακα που θα δουλέψουμε *124*
7.5.3 Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων *124*
7.5.4 Απόφαση για τον αριθμό των συνιστωσών που θα κρατήσουμε *124*
7.5.5 Εύρεση των συνιστωσών *128*
7.5.6 Ερμηνεία των συνιστωσών *128*
7.5.7 Δημιουργία νέων μεταβλητών *129*
7.6 Αποτελέσματα για Ανάλυση σε Κύριες Συνιστώσες από Δείγμα 129
7.7 Μερικά Χρήσιμα Αποτελέσματα 131
7.8 Χρήση των Κυρίων Συνιστωσών 132
7.9 Παραλλαγές της Μεθόδου 133
7.10 Case Study: Αποτελέσματα Επτάθλου (Ολυμπιακοί αγώνες, Λος Άντζελες 1984) 134
7.11 Bootstrap στην ανάλυση κυρίων συνιστωσών 149
7.11.1 Η κατανομή των ιδιοτιμών. *152*
7.11.2 Ερμηνεία των ιδιοδιανυσμάτων – συνιστωσών *155*

Κεφάλαιο 8. Παραγοντική Ανάλυση 161
8.1 Εισαγωγή 161
8.2 Το Ορθογώνιο Μοντέλο 162

8.3 Υποθέσεις του Ορθογώνιου Μοντέλου 164
8.4 Έλεγχος Συσχετίσεων 165
8.5 Αριθμός Παραγόντων και Εκτίμηση των Παραγόντων 169
8.5.1 Εκτίμηση με τη μέθοδο Κυρίων Συνιστωσών *170*
8.5.2 Εκτίμηση με τη μέθοδο μεγίστης πιθανοφάνειας *172*
8.5.3 Κριτήρια Επιλογής Μοντέλου *174*
8.5.4 Άλλες μέθοδοι Εκτίμησης *174*
8.6 Περιστροφή 175
8.7 Υπολογισμός των Σκορ των Παραγόντων 176
8.8 Confirmatory Factor Analysis 178
8.9 Μη Ορθογώνια Παραγοντική Ανάλυση 178
8.10 Συμπεράσματα και Σχόλια 179
8.11 Εφαρμογή της Μεθόδου 180
8.11.1 Καταλληλότητα των δεδομένων *181*
8.11.2 Επιλογή αριθμού παραγόντων *184*
8.11.3 Εκτίμηση των παραμέτρων *185*
8.11.4 Αξιολόγηση του μοντέλου *194*
8.11.5 Περιστροφή *195*
8.11.6 Δημιουργία των factor scores *199*
8.11.7 Χρήση των σκορ *200*
8.12 Παραγοντική Ανάλυση με τη Χρήση του Στατιστικού Πακέτου SPSS for Windows 203

Κεφάλαιο 9. Ανάλυση κατά Συστάδες 209
9.1 Εισαγωγή 209
9.2 Η Απόσταση 212
9.2.1 Η έννοια της απόστασης 212
9.2.2 Μέτρα απόστασης 215
9.3 Προβλήματα που πρέπει να αντιμετωπίσει ο ερευνητής 224
9.4 Η μέθοδος K-Means 226
9.4.1 Ο αλγόριθμος *226*
9.4.2 Χαρακτηριστικά του αλγορίθμου *228*
9.4.3 K-means στο SPSS *230*
9.4.4 Εφαρμογή *232*
9.5 Ιεραρχική ομαδοποίηση 235
9.5.1 Ο αλγόριθμος *236*
9.5.2 Επιλογή μεθόδου *236*
9.5.3 Παράδειγμα και σύγκριση των μεθόδων *239*
9.5.4 Χαρακτηριστικά του αλγόριθμου *244*
9.5.5 Εφαρμογή *244*
9.6 Ανάλυση σε ομάδες με τη χρήση πιθανοθεωρητικού μοντέλου 252
9.7 Άλλοι αλγόριθμοι 255
9.8 Κριτήρια επιλογής αριθμού ομάδων 256
9.9 Διάφορα άλλα θέματα 259

9.9.1 Μεγάλα σετ δεδομένων 259
9.9.2 Ενδιαφέροντα σημεία 260
9.9.3 Επιτυχία της μεθόδου 261

Κεφάλαιο 10. Διαχωριστική Ανάλυση 263
10.1 Εισαγωγή 263
10.2 Ο Βασικός Κανόνας Διαχωρισμού Δυο Ομάδων 266
10.3 Διαχωρισμός Δυο Ομάδων με τη Χρήση της Κανονικής Κατανομής 268
10.4 Η Λογική της Διαχωριστικής Συνάρτησης του Fisher 272
10.5 Γενίκευση Διαχωριστικής Ανάλυσης σε Κ ομάδες 273
10.6 Γενίκευση Διαχωριστικής Ανάλυσης του Fisher σε Κ ομάδες 274
10.7 Άλλες Προσεγγίσεις για το Διαχωρισμό Ομάδων 274
10.8 Άλλα θέματα 277
10.8.1 Καλή προσαρμογή του μοντέλου 277
10.8.2 Μη παραμετρική διαχωριστική ανάλυση 278
10.8.3 Σχέση με την ανάλυση κατά συστάδες 279
10.9 Διαχωριστική ανάλυση με το SPSS 280
10.9.1 Ένα Απλό Παράδειγμα 280
10.9.2 Ανάλυση του Παραδείγματος με το SPSS 281
10.10 Παράδειγμα Διαχωρισμού Τεσσάρων Ομάδων. 294

Βιβλιογραφία 305

1 ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγή

Οι μέθοδοι της πολυμεταβλητής στατιστικής ανάλυσης, όπως φανερώνει και η ονομασία τους, αναφέρονται σε διαδικασίες και μεθοδολογίες όπου προσπαθούμε να καταλήξουμε σε στατιστική συμπερασματολογία με τη χρήση πολλών μεταβλητών. Στην πράξη τα δεδομένα ενός ερευνητή είναι από τη φύση τους πολυμεταβλητά, και ο σκοπός του ερευνητή σπάνια είναι να μελετήσει μια μεταβλητή ανεξάρτητα και απομονωμένα από τις υπόλοιπες. Συνεπώς, ουσιαστικά όλες οι στατιστικές μέθοδοι είναι από τη φύση τους πολυμεταβλητές, ή τουλάχιστον τα δεδομένα που έχει ένας ερευνητής στη διάθεση του είναι σχεδόν πάντα πολυμεταβλητά και εξαρτάται πια από εκείνον το κατά πόσο θέλει να χρησιμοποιήσει όλα τα δεδομένα τους για να αποκομίσει τη μεγαλύτερη δυνατή πληροφορία από τα δεδομένα του.

Από τα παραπάνω μπορεί να παρατηρήσει κανείς πως οι πολυμεταβλητές τεχνικές δεν αναπτύχθηκαν ξεχωριστά από τις (λίγες) μονομεταβλητές τεχνικές. Ο βασικός λόγος που δεν είναι κάποιες από αυτές τόσο διαδεδομένες έχει να κάνει κυρίως με την πολυπλοκότητά τους που οδήγησε σε σοβαρούς περιορισμούς στην πρακτική τους εφαρμογή. Κάτι τέτοιο δεν ισχύει πια με τη γενικευμένη χρήση υπολογιστών στη στατιστική, καθώς υπάρχει πάντα μια ποικιλία στατιστικών πακέτων που μπορούν να χρησιμοποιηθούν ακόμα και για ιδιαίτερα πολύπλοκες μεθόδους.

Αρκετές πολυμεταβλητές μέθοδοι είναι ιδιαίτερα διαδεδομένες από μόνες τους ως ξεχωριστές τεχνικές. Για παράδειγμα η πολλαπλή γραμμική παλινδρόμηση (και επομένως και το γενικό γραμμικό μοντέλο) στοχεύει στη μελέτη πολλών μεταβλητών συγχρόνως. Η τεχνική της παλινδρόμησης είναι ιδιαίτερα διαδεδομένη σε ερευνητές από μια μεγάλη ποικιλία επιστημών και αποτελεί την πιο βασική πολυμεταβλητή τεχνική. Αν και στην ουσία της η παλινδρόμηση είναι πολυμεταβλητή τεχνική δεν θα μας απασχολήσει καθώς υπάρχει μια ποικιλία συγγραμμάτων που την περιγράφουν σε μεγάλη έκταση. Το ίδιο ισχύει και για την ανάλυση διακύμανσης, μια άλλη πολυμεταβλητή στην ουσία τεχνική, η οποία αναλύεται σε πολλά συγγράμματα και δεν θα μας απασχολήσει στις σημειώσεις αυτές παρά μόνο όταν μιλήσουμε για μια γενίκευση της, την πολυμεταβλητή ανάλυση διακύμανσης.

Οι λόγοι για τους οποίους οι πολυμεταβλητές τεχνικές είναι ιδιαίτερα χρήσιμες είναι:

- *Έχουμε περισσότερη πληροφορία* (περισσότερες μεταβλητές ερμηνεύουν καλύτερα το φαινόμενο). Συνήθως ο ερευνητής σκοπεύει με τα δεδομένα που έχει στα χέρια του να περιγράψει ή να ερμηνεύσει κάποιο φαινόμενο ή κάποιο μηχανισμό. Είναι ευνόητο ότι όσο περισσότερη πληροφορία έχει κανείς τόσο περισσότερο μπορεί να περιορίσει την αβεβαιότητα του και επομένως να εξάγει συμπεράσματα με μεγαλύτερη βαρύτητα.
- *Μελετάμε συσχετισμούς* (μεταξύ μεταβλητών και μεταξύ υποκειμένων). Ο κόσμος μέσα στον οποίο ζούμε είναι ένας κόσμος γεμάτος από συσχετίσεις μεταξύ διαφορετικών πραγμάτων και οντοτήτων και θα ήταν απλοϊκό να τον μελετά κανείς χωρίς να τις λαμβάνει υπόψη του. Από την άλλη η ανακάλυψη τέτοιων συσχετίσεων ανάμεσα σε διαφορετικές μεταβλητές μπορεί από μόνη της να οδηγήσει σε καινούριες ερμηνείες για τα υπό μελέτη φαινόμενα. Επομένως μοιάζει καλή ιδέα να μελετήσουμε συγχρόνως ένα σύνολο μεταβλητών με σκοπό να αντλήσουμε όσο γίνεται περισσότερα από τα δεδομένα μας.

Από τα παραπάνω γίνεται σαφές πως οι πολυμεταβλητές τεχνικές μας δίνουν τη δυνατότητα για καλύτερη μελέτη των φαινομένων. Οι πολυμεταβλητές τεχνικές χρησιμοποιούνται για:

- Την εύρεση και ερμηνεία συσχετίσεων μεταξύ των μεταβλητών. Όπως είπαμε και πριν τα περισσότερα φαινόμενα παρουσιάζουν μια πολυπλοκότητα και επομένως δύσκολα εξηγούνται από μια μόνο μεταβλητή. Χρειάζεται λοιπόν να εξετάσουμε και να λάβουμε υπόψη μας το πώς συσχετίζονται διάφορες μεταβλητές προκειμένου να αποκτήσουμε καλύτερη γνώση για το φαινόμενο που εξετάζουμε.
- Τη δημιουργία ομάδων είτε από παρατηρήσεις είτε από μεταβλητές σύμφωνα με κάποια χαρακτηριστικά. Και οι δύο αυτές προσεγγίσεις είναι ιδιαίτερα χρήσιμες. Ο ερευνητής αγοράς ενδιαφέρεται να ξέρει ποια είναι τα χαρακτηριστικά των αγοραστών μιας συγκεκριμένης μάρκας προϊόντος. Επομένως μελετώντας δεδομένα που αφορούν την καταναλωτική συνήθεια του αγοραστή ως προς το προϊόν, δημιουργεί ομάδες από αυτούς και μπορεί με βάση αυτή την ομαδοποίηση να κατευθύνει τις μελλοντικές του κινήσεις πχ διαφήμιση, προσφορές κλπ.

  Από την άλλη ένας γιατρός ενδιαφέρεται να ομαδοποιήσει διαφορετικές μεταβλητές. Αν για παράδειγμα οι μεταβλητές είναι τα διάφορα χαρακτηριστικά του αρρώστου (τα οποία μπορεί να είναι είτε ποσοτικά, πχ ο αριθμός των αιμοπεταλίων, είτε ποιοτικά, η παρουσία ή απουσία ενός συμπτώματος) ο γιατρός ενδιαφέρεται να δει ποια από αυτά τα χαρακτηριστικά εμφανίζονται μαζί και επομένως κάποιες από τις εξετάσεις είναι πλεονάζουσες και θα μπορούσαν να αποφευχθούν.
- Τη μείωση των διαστάσεων του προβλήματος (συμπύκνωση της πληροφορίας που περιέχουν πολλές μεταβλητές σε λιγότερες). Ιδίως στις μέρες μας που πολλά δεδομένα μαζεύονται με αυτόματο ηλεκτρονικό τρόπο (πχ οι βάσεις δεδομένων των σούπερ μάρκετ), τα δεδομένα εμφανίζουν έναν γιγαντισμό. Δημιουργούνται τεράστιες βάσεις δεδομένων με

χιλιάδες μεταβλητές από τις οποίες όμως πολλές είτε είναι άχρηστες για το σκοπό της έρευνας μας, είτε η πληροφορία που μας παρέχουν περιέχεται και σε κάποια άλλη μεταβλητή και επομένως είναι πλεονάζουσες. Στο παράδειγμα του γιατρού που προηγούμενα αναφέραμε, πολλά από τα συμπτώματα εμφανίζονται μαζί και συνεπώς η καταγραφή αλλά και η εξέταση και των δύο ενώ δεν προσφέρει καμιά πληροφορία από την άλλη απαιτεί χώρο στο σκληρό δίσκο για να αποθηκευτεί (η άχρηστη για τον ερευνητή) πληροφορία. Από στατιστική άποψη πλεονάζουσες μεταβλητές μπορεί να δημιουργούν προβλήματα στη στατιστική ανάλυση (θυμηθείτε το πρόβλημα της πολυσυγγραμικότητας στην παλινδρόμηση). Επομένως θα ήταν ενδιαφέρον (και συμφέρον) να μπορούσε κανείς να μειώσει τις υποψήφιες μεταβλητές.

Έχετε επίσης υπόψη σας ότι σε μερικές επιστήμες όπως για παράδειγμα στην αρχαιολογία και τη συντήρηση έργων τέχνης, ο αριθμός των παρατηρήσεων είναι πολύ μικρός σε σχέση με τον αριθμό των μεταβλητών που έχουμε διαθέσιμες. Επομένως για να προχωρήσουμε σε οποιαδήποτε στατιστική ανάλυση χρειάζεται να δημιουργήσουμε καινούριες μεταβλητές οι οποίες κατά κάποιον τρόπο θα περιλαμβάνουν μεγάλο μέρος της πληροφορίας που είχε ένας μεγάλος αριθμός των αρχικών μεταβλητών, ώστε να προκύψουν λίγες μεταβλητές και να είναι δυνατή η στατιστική ανάλυση.

- Την πρόβλεψη νέων τιμών. Αυτό έχει να κάνει με δύο διαφορετικές περιπτώσεις. Η πρώτη αφορά την περίπτωση των χαμένων παρατηρήσεων (missing observations). Σε πολυμεταβλητά προβλήματα ενδέχεται από το σύνολο των μεταβλητών μας για κάποια παρατήρηση να λείπει η τιμή κάποιας από αυτές. Μερικές φορές είναι πολύ επίπονο να αγνοήσουμε αυτή την παρατήρηση και επομένως πρέπει με κάποια ισχυρή στατιστική μέθοδο να εκτιμήσουμε αυτή την τιμή που λείπει. Στη δεύτερη περίπτωση η τιμή που μας λείπει είναι αυτή που θα μας έδειχνε σε ποια ομάδα (από ένα σύνολο δυνητικών ομάδων) ανήκει η παρατήρηση. Σε αυτή την περίπτωση θέλουμε από παρατηρήσεις για τις οποίες η κατάταξη σε ομάδες μας είναι ήδη γνωστή, να κατασκευάσουμε κανόνες ώστε να μπορούμε να κατατάσσουμε νέες παρατηρήσεις.

Ένα κλασσικό τέτοιο παράδειγμα αφορά το αν θα δοθεί δάνειο ή όχι σε έναν υποψήφιο πελάτη τράπεζας. Από τα δεδομένα που έχει κανείς στα χέρια του γνωρίζει την εξέλιξη για ένα μεγάλο πλήθος δανείων, αν το δάνειο αποπληρώθηκε κανονικά ή όχι καθώς και όλα τα στοιχεία του δανειολήπτη. Επομένως μπορεί να κατασκευάσει έναν κανόνα σχετικά με το ποια χαρακτηριστικά του δανειολήπτη είναι εκείνα που επιδρούν στο να αποπληρώσει κανονικά ή όχι το δάνειο. Όταν λοιπόν ένας καινούριος πελάτης ζητήσει δάνειο μπορεί ο ερευνητής χρησιμοποιώντας τον κανόνα αυτό να κατατάξει τον νέο πελάτη είτε στην κατηγορία των 'καλών' πελατών είτε στην κατηγορία των 'κακών' πελατών, δηλαδή να προβλέψει την τιμή της μεταβλητής που αφορά το αν ο πελάτης θα αποπληρώσει το δάνειο ή όχι. Στην πράξη μπορεί να υπάρχει και μια τρίτη κατηγορία, οι πελάτες που αποπλήρωσαν μεν το δάνειο αλλά αφού δημιούργησαν προβλήματα κατά τη διάρκεια του (καθυστέρηση πληρωμών κλπ) και τότε θα πρέπει να κατατάξουμε τον πελάτη σε μια εκ των τριών κατηγοριών και ούτω καθεξής. Προφανώς το πρόβλημα μπορεί να γενικευτεί επιτρέποντας να υπάρχουν πολλές ομάδες πελατών.

- Μοντελοποίηση σε πολλές διαστάσεις (για την ερμηνεία πολλών μεταβλητών σε σχέση με άλλες). Για παράδειγμα η γραμμική παλινδρόμηση έχει μια εξαρτημένη μεταβλητή και πολλές ανεξάρτητες. Υπάρχουν μοντέλα πολυμεταβλητής παλινδρόμησης (προσοχή δεν εννοούμε πολλαπλή παλινδρόμηση) όπου υπάρχουν περισσότερες από μια εξαρτημένες μεταβλητές. Τέτοια μοντέλα μας επιτρέπουν να λάβουμε υπόψη τυχόν συσχετίσεις ανάμεσα στις μεταβλητές.
- Ποσοτικοποίηση μη παρατηρήσιμων ποσοτήτων. Σε πολλές κοινωνικές επιστήμες υπάρχουν βασικές έννοιες οι οποίες δεν είναι άμεσα μετρήσιμες, όπως πχ η ευφυΐα. Κάποιες από τις πολυμεταβλητές τεχνικές επιτρέπουν να δημιουργηθούν συνδυασμοί άλλων μετρήσιμων μεταβλητών (πχ οι βαθμοί σε κάποιο τεστ) οι οποίοι στη συνέχεια να θεωρηθούν ότι ποσοτικοποιούν την αφηρημένη και μη μετρήσιμη έννοια. Βέβαια αυτή η προσέγγιση εμπεριέχει έναν βαθμό αυθαιρεσίας αλλά από την άλλη προσφέρει τη δυνατότητα για ποσοτικοποίηση και άρα για επιστημονική τεκμηρίωση υποθέσεων σε αυτές τις επιστήμες.

Σε αυτή τη λίστα με τους σκοπούς της πολυμεταβλητής ανάλυσης θα μπορούσε κανείς να προσθέσει και άλλα σημεία, όπως για παράδειγμα η γραφική αναπαράσταση των δεδομένων (data visualization). Επίσης πολλές φορές οι προσωπικοί σκοποί του ερευνητή είναι εξίσου σημαντικοί αν και δεν μπορούν να μπουν σε αυτή τη λίστα.

1.2 Πολυμεταβλητές Μέθοδοι

Ας δούμε με συντομία διάφορες πολυμεταβλητές τεχνικές. Είναι σαφές ότι η λίστα δεν είναι πλήρης καθώς δεν περιλαμβάνει εξεζητημένες μεθόδους χωρίς μεγάλη απήχηση. Επίσης, μερικές από τις τεχνικές δεν πρόκειται να περιγραφούν με λεπτομέρειες στη συνέχεια και επομένως ο αναγνώστης πρέπει να καταφύγει σε αντίστοιχα ξενόγλωσσα συγγράμματα

- Ανάλυση σε κύριες συνιστώσες (Principal Components Analysis)

Η μέθοδος αυτή είναι ιδιαίτερα διαδεδομένη κυρίως λόγω της ευκολίας της και συνάμα της απλής ερμηνείας των αποτελεσμάτων. Αποσκοπεί στην εύρεση γραμμικών συνδυασμών των αρχικών δεδομένων έτσι ώστε η πληροφορία να μην χάνετε άλλα οι νέες μεταβλητές, οι συνιστώσες, να είναι ασυσχέτιστες μεταξύ τους. Η μέθοδος ουσιαστικά ερμηνεύει τις συσχετίσεις ανάμεσα στις αρχικές μεταβλητές, μειώνει τις διαστάσεις του προβλήματος και επιτρέπει απλή ερμηνεία με τη δημιουργία απλών γραφημάτων που όμως επιτρέπουν την αποκάλυψη ενδιαφερουσών σχέσεων στα δεδομένα μας. Στην πράξη είναι απλά ένας μαθηματικός μετασχηματισμός των δεδομένων. Η μέθοδος θα αναπτυχθεί στη συνέχεια.

- Παραγοντική ανάλυση (Factor Analysis)

Η παραγοντική ανάλυση είναι μια μέθοδος ιδιαίτερα αναπτυγμένη στις κοινωνικές επιστήμες και αποσκοπεί στην εύρεση και ερμηνεία παραγόντων που δεν είναι μετρήσιμοι αλλά υπάρχουν και προκαλούν τη συσχέτιση μεταξύ των παρατηρούμενων μεταβλητών. Σε αντίθεση με την ανάλυση σε κύριες συνιστώσες έχει ένα ισχυρό θεωρητικό υπόβαθρό και επομένως επιτρέπει την στατιστική εξέταση διαφόρων υποθέσεων σχετικά με το υπό μελέτη φαινόμενο. Η μέθοδος θα αναπτυχθεί στη συνέχεια.

- Ανάλυση σε ομάδες (Cluster Analysis)

Σκοπός της ανάλυσης σε ομάδες είναι η δημιουργία ομάδων (clusters) από παρατηρήσεις για τις οποίες τα δεδομένα δείχνουν πως έχουν παρόμοια χαρακτηριστικά. Για το σκοπό αυτό έχουν αναπτυχθεί διάφορες μεθοδολογίες. Κάποιες από αυτές είναι εμπειρικές, δηλαδή χωρίς σημαντικό θεωρητικό υπόβαθρο, και βασίζονται στην έννοια της απόστασης. Έχουν ιδιαίτερη απήχηση σε εφαρμοσμένα προβλήματα καθώς δεν χρειάζονται υποθέσεις και λειτουργούν εύκολα στην πράξη. Από την άλλη υπάρχουν και μέθοδοι βασισμένες σε μοντέλα (model-based clustering) οι οποίες αφενός έχουν ένα σημαντικό θεωρητικό υπόβαθρο αφετέρου προσφέρουν μια σειρά από μεθοδολογικά εργαλεία για να μπορεί κανείς να αξιολογήσει τα αποτελέσματα. Σε μερικές επιστήμες τη συναντάμε με άλλες ονομασίες όπως Ταξινόμηση (Taxonomy), Κατηγοριοποίηση (Classification) αλλά και ως Segmentation (αυτός ο όρος είναι ιδιαίτερα διαδεδομένος στις οικονομικές επιστήμες). Η ανάλυση σε ομάδες θα μελετηθεί αναλυτικότερα σε επόμενο κεφάλαιο.

- Διακριτική ανάλυση (Discriminant Analysis)

Η διακριτική (ή διαχωριστική) ανάλυση έχει σκοπό να δημιουργήσει κανόνες από ήδη υπάρχοντα δεδομένα ώστε να είναι σε θέση κανείς να κατατάξει μελλοντικές παρατηρήσεις σε έναν από τους υπό εξέταση πληθυσμούς. Στη βιβλιογραφία υπάρχουν διαφορετικές προσεγγίσεις για να επιτευχθεί αυτό είτε μέσω συγκεκριμένων παραμετρικών υποθέσεων είτε με τη χρήση μη παραμετρικών τεχνικών. Η διακριτική ανάλυση χρησιμοποιείται πολύ από τις επιστήμες των υπολογιστών όπου και την συναντά κανείς με την ονομασία Αναγνώριση Προτύπων (Pattern Recognition) και αποσκοπεί να δημιουργήσει κανόνες που θα μπορούν να ξεχωρίζουν συγκεκριμένα πρότυπα (πχ κείμενα, εικόνες κλπ) από έναν μεγάλο όγκο πληροφορίας. Και για την διακριτική ανάλυση θα μιλήσουμε αναλυτικότερα σε επόμενη ενότητα.

- Ανάλυση αντιστοιχιών (Correspondence Analysis)

Η ανάλυση αντιστοιχιών είναι μια μέθοδος συγγενής της ανάλυσης σε κύριες συνιστώσες. Ο σκοπός είναι παρόμοιος αλλά στη μέθοδο της ανάλυσης σε αντιστοιχίες τα δεδομένα είναι κατηγορικά. Αυτό έχει κάνει τη μέθοδο σημαντικό εργαλείο στις κοινωνικές επιστήμες όπου η συλλογή των στοιχειών γίνεται με τη χρήση ερωτηματολογίου και άρα τα δεδομένα είναι από τη

φύση τους κατηγορικά. Ο σκοπός είναι να δημιουργηθούν άξονες πάνω στους οποίους προβάλλει κανείς τις παρατηρήσεις αλλά και τις μεταβλητές και επομένως μπορεί να δει και να ερμηνεύσει πως σχετίζονται μεταξύ τους οι μεταβλητές.

- Ανάλυση κανονικών συσχετίσεων (Canonical Correlation Analysis)

Και αυτή η μέθοδος είναι παρόμοια με την ανάλυση σε κύριες συνιστώσες με τη διαφορά πως οι συνιστώσες που προκύπτουν έχουν μεταξύ τους κάποια ελεγχόμενη συσχέτιση. Με αυτό τον τρόπο επιτυγχάνουμε μια πιο ρεαλιστική περιγραφή των δεδομένων.

- Πολυδιάστατη κλιμακοποίηση (Multidimensional Scaling)

Η πολυδιάστατη κλιμακοποίηση (ή κλιμάκωση) είναι μια μαθηματική μέθοδος που σκοπό έχει να προβάλλει τις διαστάσεις του προβλήματος στο χώρο των δύο (συνήθως) ή περισσοτέρων διαστάσεων. Με αυτό βελτιώνουμε την ικανότητα μας να ερμηνεύσουμε τα αποτελέσματα καθώς είναι πολύ πιο εύκολο να μελετήσουμε ένα διάγραμμα λίγων διαστάσεων σε σχέση με δεδομένα πολλών διαστάσεων χωρίς ουσιαστικά κανένα εργαλείο απεικόνισης τους. Είναι επίσης σημαντικό να αναφερθεί πως η μέθοδος με αυτό τον τρόπο καταφέρνει να δημιουργήσει δείκτες βασισμένους σε όλα τα δεδομένα οι οποίοι είναι πιο εύκολα κατανοητοί.

- Πολυμεταβλητό Γραμμικό Μοντέλο (Multivariate Linear Model)

Το απλό γραμμικό μοντέλο θεωρείται, και είναι, ένα από τα πιο σημαντικά εργαλεία για στατιστική συμπερασματολογία. Σε μια γενική θεώρηση το γραμμικό μοντέλο περιλαμβάνει την παλινδρόμηση και την ανάλυση διακύμανσης ως ειδικές περιπτώσεις. Στο μοντέλο υπάρχει μια εξαρτημένη μεταβλητή και πολλές συνήθως ανεξάρτητες, οι οποίες στην περίπτωση της ανάλυσης διακύμανσης είναι κατηγορικές. Το μοντέλο αυτό μπορεί να γενικευτεί στις πολλές διαστάσεις, επιτρέποντας να υπάρχουν τώρα πια πολλές εξαρτημένες μεταβλητές ενώ επίσης επιτρέπουμε αυτές οι εξαρτημένες μεταβλητές να έχουν και μεταξύ τους κάποια συσχέτιση. Επομένως γενικεύοντας το μονοδιάστατο γραμμικό μοντέλο προκύπτει η μέθοδος της πολυμεταβλητής παλινδρόμησης (Multivariate Regression) και η μέθοδος της πολυμεταβλητής ανάλυσης διακύμανσης (MANOVA).

- Μέθοδοι για δεδομένα διεύθυνσης (Directional data)

Σε αρκετές περιπτώσεις τα δεδομένα του ερευνητή αφορούν παρατηρήσεις όπου σχετίζονται με την έννοια της διεύθυνσης. Ένα τέτοιο παράδειγμα είναι ανεμολογικά δεδομένα όπου η διεύθυνση είναι μια βασική μεταβλητή. Από την άλλη συνήθως όταν η διεύθυνση είναι μια μεταβλητή υπάρχουν και διάφορες άλλες που περιγράφουν το φαινόμενο. Δηλαδή τα δεδομένα διεύθυνσης συνήθως είναι πολυμεταβλητά. Για το σκοπό αυτό και για τη μελέτη τέτοιων δεδομένων χρειαζόμαστε ειδικές μεθόδους. Σκεφτείτε το απλό παράδειγμα μιας μόνο μεταβλητής που αφορά διεύθυνση. Αν οι παρατηρήσεις είναι σε μοίρες τότε οι τιμές που έχουμε

είναι από 0 έως και 360 μοίρες αλλά η τιμή 1 μοίρα είναι τόσο κοντά στην τιμή 2 μοίρες όσο και στις 360 μοίρες. Συνεπώς υπάρχει μια σειρά από κατάλληλες μεθοδολογίες οι οποίες επιτρέπουν τη μελέτη και τη στατιστική συμπερασματολογία σε τέτοιας μορφής δεδομένα.

Οι μέθοδοι που μόλις τώρα συνοπτικά αναφέραμε δεν είναι οι μόνες πολυμεταβλητές μέθοδοι. Σαφώς και υπάρχουν και άλλες και ελπίζουμε στο μέλλον να προκύψουν και καινούριες. Ουσιαστικά κάθε μέθοδος που χρησιμοποιούμε στη μονομεταβλητή περίπτωση γενικεύεται εύκολα και στην πολυμεταβλητή περίπτωση. Εξαιτίας περιορισμών χώρου δεν θα συζητήσουμε για όλες τις μεθόδους που προαναφέρθηκαν αλλά για λίγες από αυτές.

Στο κεφάλαιο 2 θα συζητήσουμε περιγραφικές μεθόδους για πολυμεταβλητά δεδομένα και συγκεκριμένα περιγραφικά γραφήματα και περιγραφικά μέτρα. Στη συνέχεια στο κεφάλαιο 3 θα δούμε την πιο σημαντική πολυμεταβλητή κατανομή που είναι η πολυμεταβλητή κανονική κατανομή και θα συζητήσουμε τις ιδιότητες της και τη χρησιμότητα της για στατιστική συμπερασματολογία. Στο κεφάλαιο 4 θα δούμε δειγματοληπτικές κατανομές που αφορούν την πολυμεταβλητή περίπτωση ενώ στο κεφάλαιο 5 θα δούμε πολυμεταβλητούς ελέγχους υποθέσεων και στο κεφάλαιο 6 την πολυμεταβλητή ανάλυση διακύμανσης. Εν συνεχεία, στο κεφάλαιο 7 θα δούμε τη μέθοδο των κυρίων συνιστωσών ενώ στο κεφάλαιο 8 την μέθοδο της παραγοντικής ανάλυσης και την μέθοδο της ανάλυσης κατά συστάδες στο 9ο κεφάλαιο. Οι σημειώσεις θα τελειώσουν με το κεφάλαιο 10 που αφορά την διαχωριστική ανάλυση.

2 ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

2.1 Γραφήματα

Όχι λανθασμένα πολλοί υποστηρίζουν πως οποιαδήποτε στατιστική ανάλυση πρέπει να ξεκινά κατασκευάζοντας διάφορα γραφήματα με τα δεδομένα ώστε να αποκτήσει κανείς μια εικόνα για αυτά. Δεδομένου μάλιστα πως μια εικόνα είναι κατά πολύ προτιμότερη από το να προσπαθήσει κανείς να εξάγει συμπεράσματα κοιτάζοντας απλά αριθμούς, τα γραφήματα αποτελούν ένα πολύτιμο εργαλείο για τον ερευνητή. Στην περίπτωση μονομεταβλητών γραφημάτων, όταν δηλαδή εξετάζαμε μια μόνο μεταβλητή κάθε φορά, μπορούσε να δημιουργήσει κανείς διάφορα γραφήματα όπως Διάγραμμα Σημείων (Dotplot), Διάγραμμα Πλαισίου-Απολήξεων (Boxplot) ή Ιστόγραμμα (Histogram), μεταξύ άλλων για να αποκτήσει μια εικόνα για αυτά. Στην περίπτωση κατηγορικών μεταβλητών μπορούσαμε να κατασκευάσουμε ραβδογράμματα (barcharts) ή κυκλικά διαγράμματα (pie-charts).

Όταν έχουμε περισσότερες από μια μεταβλητή χρειαζόμαστε γραφήματα που να μπορούν να αναπαραστήσουν περισσότερες από μια μεταβλητές κάθε φορά. Γενικά κάτι τέτοιο δεν είναι καθόλου εύκολο. Διαγράμματα τριών διαστάσεων αν και μπορούν να κατασκευαστούν από τα σύγχρονα πακέτα δεν είναι εύκολο να ερμηνευτούν. Συνεπώς δεν είναι πολύ εύκολο να αναπαρασταθούν γραφικά πολυμεταβλητά δεδομένα.

Στη συνέχεια θα παρουσιάσουμε κάποιες ιδέες που μας επιτρέπουν να απεικονίσουμε γραφικά πολυμεταβλητά δεδομένα. Θα πρέπει να τονιστεί πως υπάρχει μια μεγάλη γκάμα από μεθόδους παρουσίασης των δεδομένων σε πολλές διαστάσεις και πως οι τεχνικές οπτικοποίησης δεδομένων (visualization methods) αναπτύσσονται ραγδαία τα τελευταία χρόνια, λόγω της ολοένα αυξανόμενης ισχύος των υπολογιστών.

Στην πολυδιάστατη περίπτωση προσπαθούμε να παρουσιάσουμε τα δεδομένα στον χώρο των δύο διαστάσεων, χάνοντας εν γνώσει μας πληροφορία. Ας δούμε μερικούς τρόπους για απεικόνιση πολυμεταβλητών δεδομένων.

2.1.1 Matrix Plot

Το Matrix Plot δεν είναι παρά ένας οργανωμένος πίνακας από απλά διαγράμματα σημείων για ζευγάρια μεταβλητών. Το πλεονέκτημα του είναι πως μπορούμε να δούμε όλα τα

δυνατά ζευγάρια αλλά και επειδή οι κλίμακες είναι σταθερές μπορούμε να συγκρίνουμε ζεύγη μεταβλητών μεταξύ τους.

Ας δούμε ένα παράδειγμα:

Παράδειγμα 2.1. Τα δεδομένα αφορούν τα 25 σημαντικότερα πανεπιστήμια και τα 25 κολέγια κλασσικών σπουδών (liberal arts colleges) της Αμερικής. Τόσο τα πανεπιστήμια όσο και τα κολέγια αυτά θεωρούνται ισάξια από την άποψη της σημαντικότητας των πτυχίων και το μόνο που αλλάζει είναι η κατεύθυνση των σπουδών που προσφέρουν. Για τις 50 παρατηρήσεις υπάρχουν 8 μεταβλητές και συγκεκριμένα οι μεταβλητές:

Τα δεδομένα υπάρχουν στον πίνακα 2.1

Όνομα σχολείου	School Type	SAT	Acceptance	Student	Top 10%	%PhD	Grad %
Amherst	Lib Arts	1315	22	26636	85	81	93
Swarthmore	Lib Arts	1310	24	27487	78	93	88
Williams	Lib Arts	1336	28	23772	86	90	93
Bowdoin	Lib Arts	1300	24	25703	78	95	90
Wellesley	Lib Arts	1250	49	27879	76	91	86
Pomona	Lib Arts	1320	33	26668	79	98	80
Wesleyan (CT)	Lib Arts	1290	35	19948	73	87	91
Middlebury	Lib Arts	1255	25	24718	65	89	92
Smith	Lib Arts	1195	57	25271	65	90	87
Davidson	Lib Arts	1230	36	17721	77	94	89
Vassar	Lib Arts	1287	43	20179	53	90	84
Carleton	Lib Arts	1300	40	19504	75	82	80
Claremont McKenna	Lib Arts	1260	36	20377	68	94	74
Oberlin	Lib Arts	1247	54	23591	64	98	77
Washington & Lee	Lib Arts	1234	29	17998	61	89	78
Grinnell	Lib Arts	1244	67	22301	65	79	73
Mount Holyoke	Lib Arts	1200	61	23358	47	83	83
Colby	Lib Arts	1200	46	18872	52	75	84
Hamilton	Lib Arts	1215	38	20722	51	86	85
Bates	Lib Arts	1240	36	17554	58	81	88
Haverford	Lib Arts	1285	35	19418	71	91	87
Colgate	Lib Arts	1258	38	17520	61	78	85
Bryn Mawr	Lib Arts	1255	56	18847	70	81	84
Occidental	Lib Arts	1170	49	20192	54	93	72
Barnard	Lib Arts	1220	53	17653	69	98	80
Harvard	Univ	1370	18	46918	90	99	90
Stanford	Univ	1370	18	61921	92	96	88
Yale	Univ	1350	19	52468	90	97	93
Princeton	Univ	1340	17	48123	89	99	93
Cal Tech	Univ	1400	31	102262	98	98	75
MIT	Univ	1357	30	56766	95	98	86
Duke	Univ	1310	25	39504	91	95	91

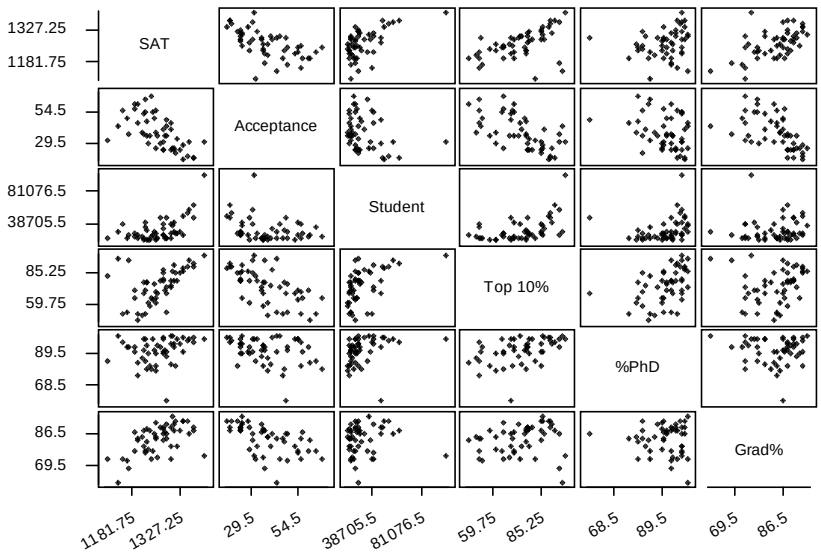
(ο πίνακας συνεχίζεται στην επόμενη σελίδα)

Dartmouth	Univ	1306	25	35804	86	100	95
Cornell	Univ	1280	30	37137	85	90	83
Columbia	Univ	1268	29	45879	78	93	90
U of Chicago	Univ	1300	45	38937	74	100	73
Brown	Univ	1281	24	24201	80	98	90
U Penn	Univ	1280	41	30882	87	99	86
Berkeley	Univ	1176	37	23665	95	93	68
Johns Hopkins	Univ	1290	48	45460	69	58	86
Rice	Univ	1327	24	26730	85	95	88
UCLA	Univ	1142	43	26859	96	100	61
U Va.	Univ	1218	37	19365	77	91	88
Georgetown	Univ	1278	24	23115	79	89	89
UNC	Univ	1109	32	19684	82	84	73
U Michican	Univ	1195	60	21853	71	93	77
Carnegie Mellon	Univ	1225	64	33607	52	84	77
Northwestern	Univ	1230	47	28851	77	79	82
Washington U (MO)	Univ	1225	54	39883	71	98	76
U of Rochester	Univ	1155	56	38597	52	96	73

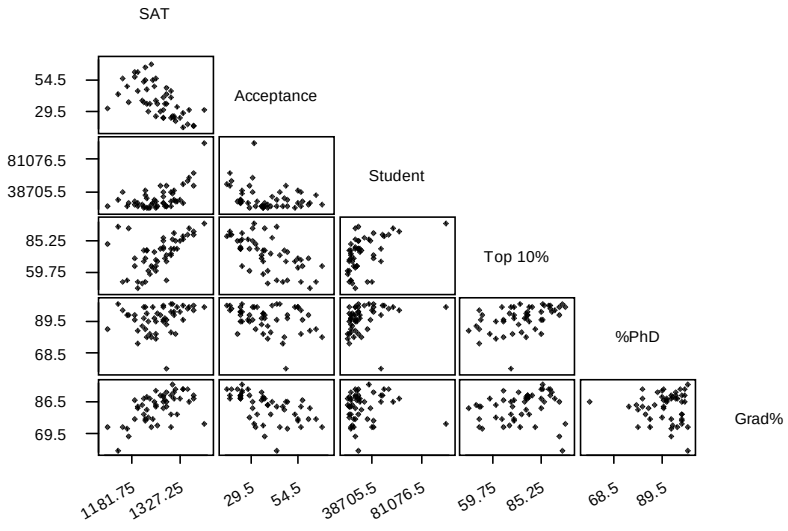
Πίνακας 2.1. Δεδομένα αμερικάνικων πανεπιστημίων και κολεγίων

Στο γράφημα 2.1 μπορεί κανείς να δει το matrix plot για τα δεδομένα αυτά. Για να κατασκευαστεί το γράφημα χρειάζεται απλά να οργανώσουμε πολλά απλά διαγράμματα σημείων μαζί. Παρατηρείστε πως κάθε γράφημα εμφανίζεται δύο φορές λόγω της συμμετρίας του πίνακα. Για αυτό το λόγο, συνήθως, το matrix plot εμφανίζεται με τη μορφή ενός διαγώνιου πίνακα όπως στο γράφημα 2.2.

Όνομα μεταβλητής	**Περιγραφή**
School:	Το όνομα του πανεπιστήμιου
School_Type	Ο τύπος του πανεπιστημίου (Univ ή liberal arts college)
SAT:	Η διάμεσος του βαθμού των φοιτητών στα διαγωνίσματα SAT
Acceptance	Το ποσοστό των φοιτητών που έγιναν δεκτοί στο πρόγραμμα του πανεπιστημίου
$/Student	Το μέσο ποσό σε δολάρια που ξοδεύει κάθε φοιτητής
Top 10%:	Το ποσοστό των τελειόφοιτων με βαθμό πάνω από κάποιο όριο
%PhD:	Ποσοστό διδασκόντων με διδακτορικό
Grad%	Το ποσοστό των φοιτητών που τελικά αποφοιτούν



Γράφημα 2.1. Matrix plot για τα δεδομένα του παραδείγματος 2.1

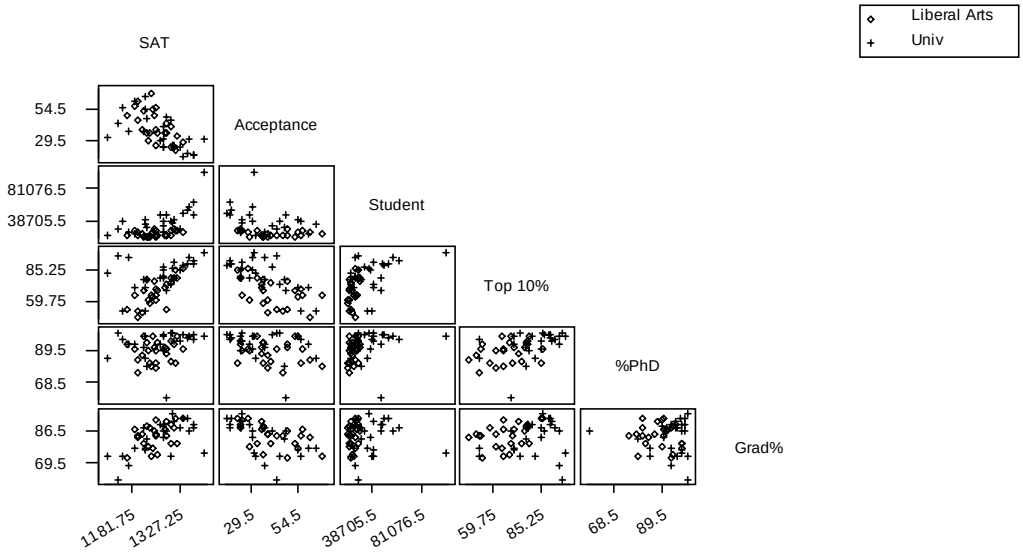


Γράφημα 2.2. Matrix plot, σε διαγώνια μορφή, για τα δεδομένα του Παραδείγματος 2.1

Από ένα matrix plot μπορεί κανείς να αποκτήσει γρήγορα μια εικόνα για το ποιες μεταβλητές συσχετίζονται με ποιες άλλες. Οι ονομασίες των μεταβλητών υπάρχουν στη διαγώνιο του πίνακα. Για να βρει κανείς ποιο είναι το ζεύγος των μεταβλητών για το οποίο έχει σχηματιστεί το διάγραμμα σημείων αρκεί να βρει ποια είναι η μεταβλητή που απεικονίζεται σε κάθε γραμμή και σε κάθε στήλη του πίνακα. Εναλλακτικά κάποια πακέτα έχουν στη διαγώνιο, αντί για τα ονόματα των μεταβλητών, κάποια μονοδιάστατα γραφήματα όπως boxplots, ή ιστογράμματα.

Έτσι από το γράφημα 2.1 μπορούμε να δούμε πως ανάμεσα στις μεταβλητές top 10% και SAT υπάρχει μια έντονη γραμμική σχέση ενώ αντίθετα η σχέση ανάμεσα στη μεταβλητή Acceptance και τη μεταβλητή Grad% φαίνεται να μην είναι ισχυρή.

Τα matrix plot είναι ιδιαίτερα χρήσιμα όταν θέλουμε να διαλέξουμε ζεύγη μεταβλητών με ισχυρή συσχέτιση, όπως πχ στην Γραμμική Παλινδρόμηση. Επίσης είναι πολύ χρήσιμα αν χρησιμοποιήσουμε διαφορετικά σύμβολα για διαφορετικές ομάδες. Για παράδειγμα, μπορεί κανείς να δει το γράφημα 2.3 όπου τα δύο διαφορετικά είδη πανεπιστήμιων συμβολίζονται με διαφορετικά σύμβολα.



Γράφημα 2.3. Matrix plot για τα δεδομένα του Παραδείγματος 2.1, διακρίνοντας μεταξύ πανεπιστημίων και κολεγίων.

Μπορεί λοιπόν να δει κανείς από το γράφημα 2.3 πως τα πανεπιστήμια έχουν αρκετά διαφορετικά χαρακτηριστικά για κάποιες μεταβλητές. Για παράδειγμα από το διάγραμμα σημείων των μεταβλητών Top 10 % και Grad % βλέπουμε πως τα πανεπιστήμια ξεχωρίζουν από τα κολέγια κλασσικών σπουδών, γιατί τα '+' είναι τα περισσότερα σε διαφορετικό σημείο από ότι τα 'ο'.

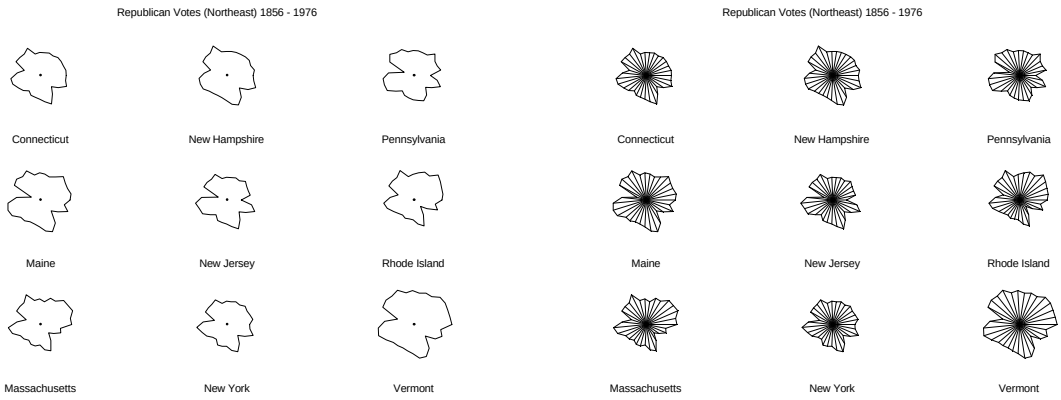
Θα πρέπει να τονιστεί πως ο ερευνητής θα μπορούσε να χρησιμοποιήσει και μια ακόμα κατηγορική μεταβλητή χρησιμοποιώντας το χρώμα για να ξεχωρίζει ανάμεσα από τις κατηγορίες αυτής της μεταβλητής. Αυτό αν και είναι ιδιαίτερα χρήσιμο σε έγχρωμες παρουσιάσεις θα μπορούσε να δημιουργήσει προβλήματα σε ασπρόμαυρες εκτυπώσεις των αποτελεσμάτων. Για αυτό η χρήση χρώματος για να υποδηλώνονται κατηγορικές μεταβλητές, αν και χρήσιμη οπτικά, μπορεί να οδηγήσει σε παρερμηνείες όταν τα χρώματα δεν διακρίνονται καλά (πχ σε μια εκτύπωση).

2.1.2 Starplots

Τα διαγράμματα-αστέρια (starplots) είναι ένας έξυπνος τρόπος να αναπαραστήσει κανείς γραφικά πολυδιάστατα γραφήματα. Για κάθε μια παρατήρηση ο ερευνητής κατασκευάζει ένα 'αστέρι' με τόσες ακτίνες όσες είναι και οι μεταβλητές, δηλαδή το μέγεθος κάθε ακτίνας αναπαριστά την τιμή της παρατήρησης για κάποια μεταβλητή. Συνεπώς κρίνοντας από το οπτικό αποτέλεσμα μπορεί κανείς να παρατηρήσει διαφορές ανάμεσα στις παρατηρήσεις καθώς αυτές θα έχουν διαφορετικά σχήματα. Τέτοια γραφήματα μπορούν να χρησιμοποιηθούν για μεγάλο αριθμό μεταβλητών. Ο περιορισμός είναι συνήθως πως επειδή κάθε παρατήρηση αντιστοιχεί σε ένα αστέρι, αν έχουμε ένα μεγάλο σετ δεδομένων με πολλές παρατηρήσεις δεν είναι εύκολο να χωρέσουμε σε μια σελίδα τόσα πολλά γραφήματα και άρα να κάνουμε οπτικές συγκρίσεις για όλες τις παρατηρήσεις.

Στις περιπτώσεις που τα δεδομένα μας ανήκουν σε διαφορετικές κατηγορίες ένα starplot για τις μέσες τιμές μπορεί να μας δώσει μια εικόνα για το πως διαφέρουν οι μέσες τιμές.

Για παράδειγμα δείτε τα starplots για ένα σετ δεδομένων, που αφορά τα ποσοστά του ρεπουμπλικανικού κόμματος στις Αμερικάνικες εκλογές για διάφορες πληθυσμιακές ομάδες. Παρουσιάζονται δύο διαφορετικές μορφές του γραφήματος, στη μια οι ακτίνες που συνδέουν το κέντρο με την περιφέρεια έχουν τυπωθεί ενώ στην άλλη περίπτωση μόνο το περίγραμμα εμφανίζεται. Από τη μορφή που παίρνει το αστέρι για κάθε παρατήρηση μπορεί κανείς να δει τις διαφορές που υπάρχουν ανάμεσα στις περιοχές. Θα πρέπει να τονιστεί πως το μήκος της ακτίνας υπολογίζεται με βάση τις τυποποιημένες τιμές.

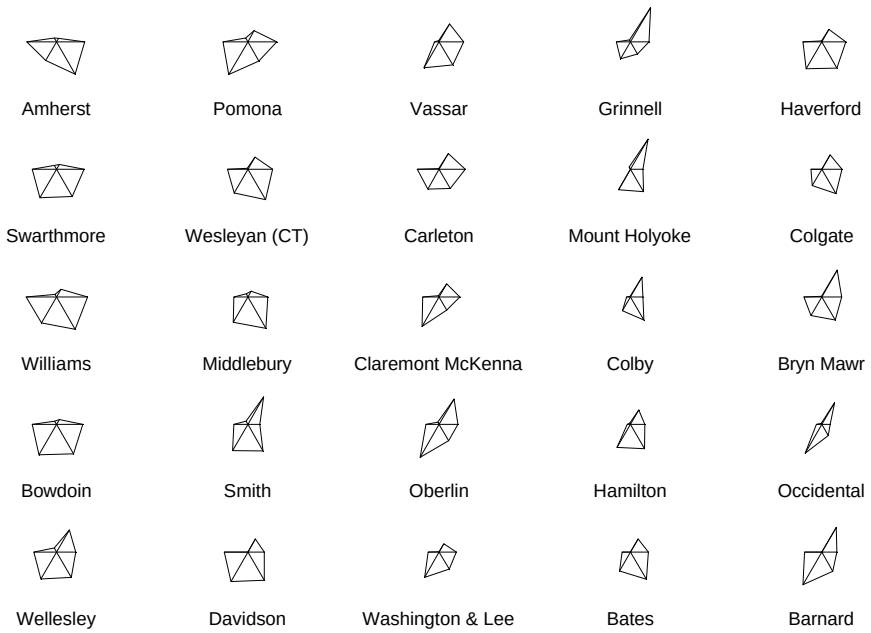


Starplot με μόνο την περίμετρο Starplot με ακτίνες

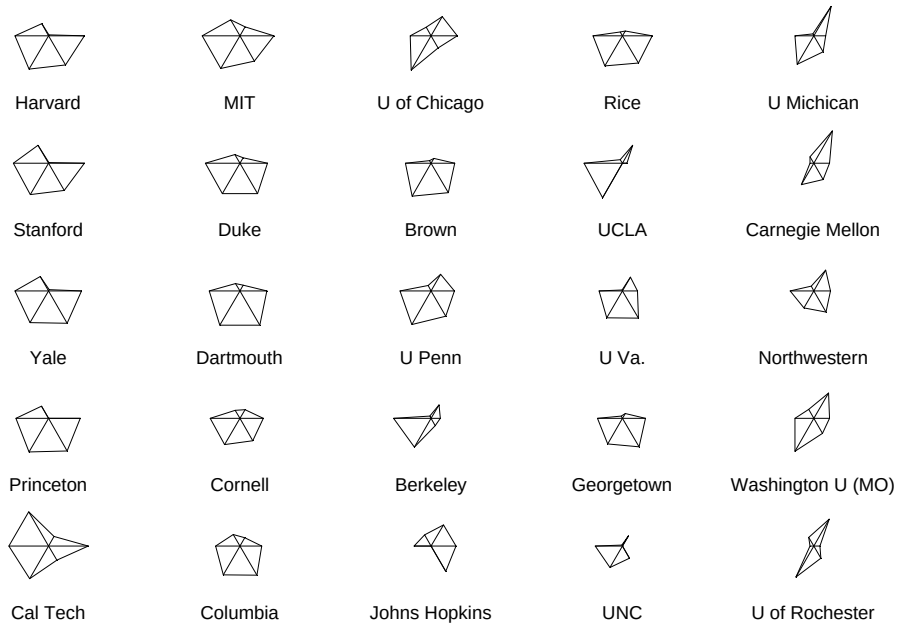
Γράφημα 2.4 Παραδείγματα γραφημάτων starplot

Παράδειγμα 2.1 (συνέχεια):

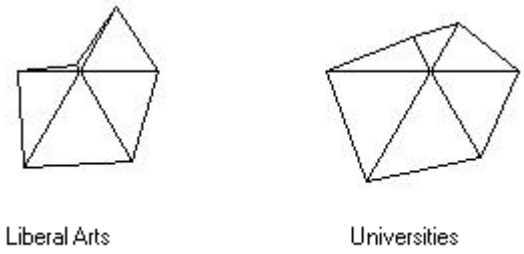
Ας επανέλθουμε στο παράδειγμα με τα αμερικάνικα πανεπιστήμια. Στα γραφήματα που ακολουθούν (2.4α και β αντίστοιχα) μπορείτε να δείτε τα starplots για τις δύο διαφορετικές κατηγορίες πανεπιστημίων. Είναι ευδιάκριτες οι διαφορές και ανάμεσα σε πανεπιστήμια της ίδιας κατηγορίας. Γενικά τα πανεπιστήμια έχουν αρκετά μεγαλύτερα αστέρια από ότι τα κολέγια κλασσικών σπουδών. Αυτό επαληθεύεται κοιτάζοντας τα αστέρια για τους μέσους κάθε ομάδας, στο γράφημα 2.4γ



Γράφημα 2.4α. Starplot για τα κολέγια κλασικών σπουδών



Γράφημα 2.4β. Starplot για τα πανεπιστήμια



Γράφημα 2.4γ. Starplot για τους μέσους των δύο ομάδων

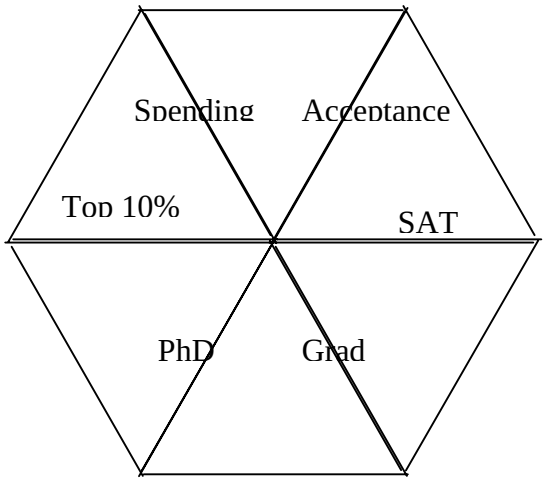
Κοιτάζοντας τις μέσες τιμές των μεταβλητών για τους δύο τύπους σχολειών (Πίνακας 2.2) βλέπουμε πως υπάρχει σημαντική διαφορά στη μεταβλητή που έχει να κάνει με τα έξοδα των φοιτητών. Εκεί παρατηρούμε και τη μεγαλύτερη διαφορά των αστεριών στο γράφημα 2.4γ

SAT	Acceptance	Student	Top 10%	%PhD	Grad%	Σχολείο
1256.64	40.56	21755.60	67.24	88.24	84.12	Liberal Arts
1271.28	35.12	38738.80	81.64	92.88	82.84	Universities

Πίνακας 2.2. Οι μέσες τιμές για τις μεταβλητές μας για τους δύο τύπους πανεπιστημίων

Θα πρέπει εδώ να σημειώσουμε πως ο τρόπος που καθορίζεται η σειρά των μεταβλητών και άρα η ακτίνα που αντιστοιχεί σε κάθε μια από αυτές, είναι συνήθως ο εξής: η πρώτη σε σειρά μεταβλητή είναι παράλληλη με τον άξονα και προς τα δεξιά δηλαδή στη θέση 3 του ρολογιού. Οι υπόλοιπες τοποθετούνται διαδοχικά σε αντίστροφή σειρά από αυτή του ρολογιού.

Στο παράδειγμα μας η σειρά των μεταβλητών απεικονίζεται στο γράφημα 2.5. Δηλαδή ξεκινάμε από την πρώτη μεταβλητή SAT που αντιστοιχεί στη δεξιά μεριά του άξονα και στη συνέχεια συνεχίζουμε με τις μεταβλητές Acceptance κλπ.



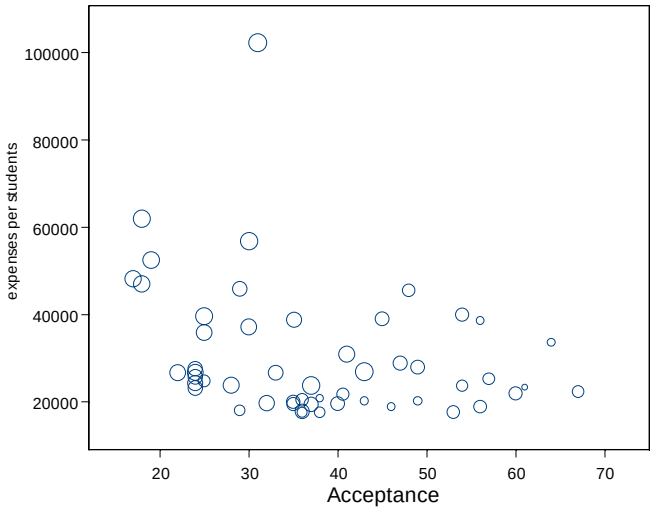
Γράφημα 2.5 Σειρά μεταβλητών σε starplot

2.1.3 Bubble Plots

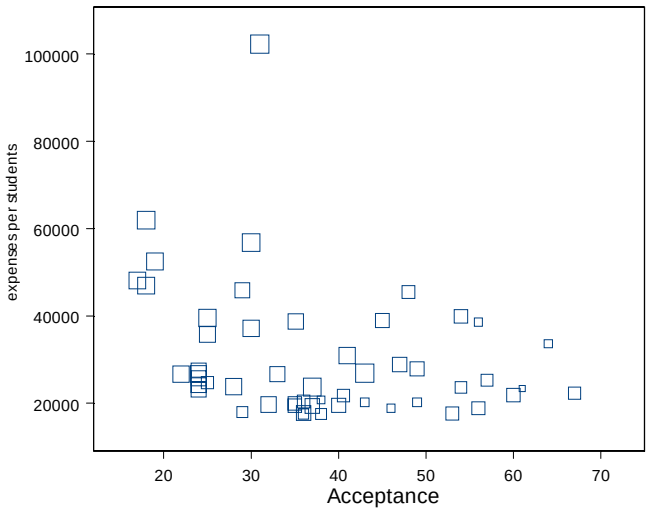
Τα bubble plots δεν είναι παρά απλά διαγράμματα σημείων για τα οποία μια τρίτη μεταβλητή απεικονίζεται ανάλογα με το μέγεθος των κύκλων που αναπαριστούν κάθε παρατήρηση. Δηλαδή ενώ στα απλά διαγράμματα σημείων κάθε παρατήρηση συμβολίζεται με ένα σύμβολο ίδιου μεγέθους για κάθε παρατήρηση, στα bubble plots, χρησιμοποιούμε διαφορετικό μέγεθος ανάλογα με μια τρίτη μεταβλητή. Συνήθως το σύμβολο είναι κύκλος, και για αυτό το γράφημα μοιάζει με σαπουνόφουσκες, από όπου πηγάζει και το όνομα. Πολλά πακέτα επιτρέπουν τη χρήση άλλων σχημάτων όπως τρίγωνα ή τετράγωνα. Το κέντρο πάντα του συμβόλου τοποθετείται στο σημείο που ορίζουν οι δύο μεταβλητές και το μέγεθος του συμβόλου καθορίζεται από μια τρίτη μεταβλητή.

Για παράδειγμα στα γραφήματα που ακολουθούν μπορεί κανείς να δει Bubble plots με τη χρήση κύκλων και τετραγώνων για τις τριάδες μεταβλητών (Acceptance, Έξοδα φοιτητή και Top 10%).

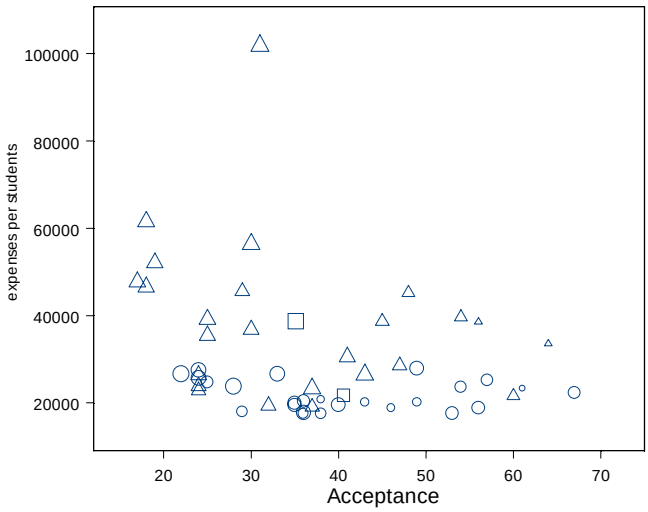
Μπορεί κανείς να εισάγει και κατηγορικές μεταβλητές στο γράφημα χρησιμοποιώντας άλλα σύμβολα ή διαφορετικά χρώματα. Για παράδειγμα στο γράφημα 2.6γ τα τριγωνάκια αντιστοιχούν σε παρατηρήσεις από πανεπιστήμια και τα κυκλάκια από κολέγια κλασσικών σπουδών.



Γράφημα 2.6α. Bubble plot για τις μεταβλητές 'ποσοστό αποδοχής' και 'έξοδα ανά φοιτητή'. Το μέγεθος κάθε κύκλου καθορίζεται από την τιμή της μεταβλητής top 10%.



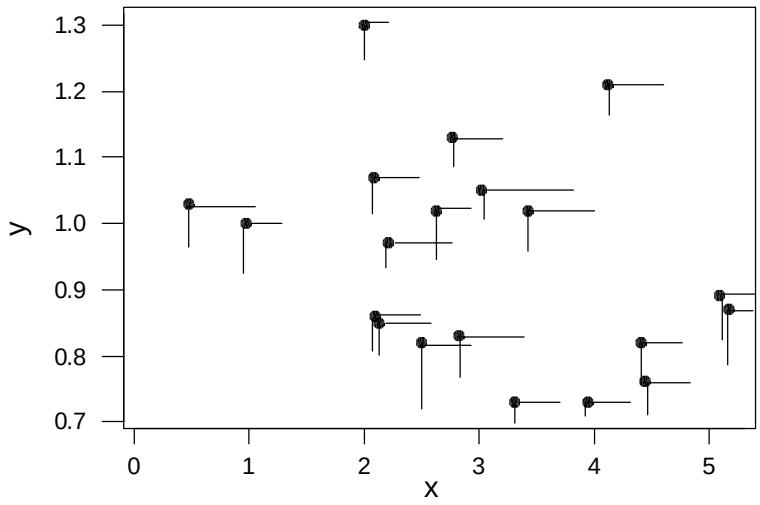
Γράφημα 2.6β. Bubble plot για τις μεταβλητές 'ποσοστό αποδοχής' και 'έξοδα ανά φοιτητή'. Το μέγεθος κάθε τετραγώνου καθορίζεται από την τιμή της μεταβλητής top 10%.



Γράφημα 2.6γ Το ίδιο Bubble plot για τις μεταβλητές 'ποσοστό αποδοχής' και 'έξοδα ανά φοιτητή' αλλά με διαφορετικά σύμβολα για τα κολέγια και τα πανεπιστήμια. Το μέγεθος κάθε τετραγώνου/τριγώνου καθορίζεται από την τιμή της μεταβλητής top 10%.

2.1.4 Glyph Plot

Το Glyph plot είναι ένας έξυπνος τρόπος να αναπαραστήσουμε τέσσερις μεταβλητές σε ένα μόνο γράφημα. Οι δύο πρώτες τοποθετούνται στο γράφημα όπως και σε ένα απλό διάγραμμα σημείων, ενώ οι άλλες δύο τοποθετούνται ως γραμμές από το σημείο που ορίζουν οι δύο πρώτες. Το μήκος αυτών των γραμμών καθορίζεται από την τιμή των παρατηρήσεων για αυτές τις δύο μεταβλητές.



Γράφημα 2.7. Παράδειγμα glyph plot. Οι δύο μεταβλητές x και y αναπαρίστανται από τη θέση του σημείου όπως σε κάθε διάγραμμα σημείων, ενώ οι άλλες δύο μεταβλητές με το μήκος των δύο κάθετων γραμμών

Στο γράφημα 2.7 μπορείτε να δείτε ένα τέτοιο glyph plot.

2.1.5 Τα Πρόσωπα του Chernoff

Η μέθοδος αυτή είναι ένας εντυπωσιακός αλλά συνάμα αρκετά αποτελεσματικός τρόπος να απεικονίσει κανείς πολυμεταβλητά δεδομένα και είναι ευρέως διαδεδομένη στις κοινωνικές επιστήμες και την ψυχολογία.

Η μέθοδος πρωτοεμφανίστηκε από τον Αμερικάνο Στατιστικό Chernoff το 1973. Στην πραγματικότητα για κάθε παρατήρηση δημιουργούμε ένα πρόσωπο του οποίου τα χαρακτηριστικά (πχ μέγεθος, θέση και μέγεθος μύτης και αυτιών κλπ) καθορίζονται από την τιμή της παρατήρησης για κάποια μεταβλητή. Έτσι ανάλογα με τις τιμές των παρατηρήσεων προκύπτουν διαφορετικά πρόσωπα. Φυσικά ανάλογα με το πόσες μεταβλητές χρησιμοποιούμε χρειαζόμαστε ανάλογο αριθμό χαρακτηριστικών του προσώπου για να το κατασκευάσουμε.

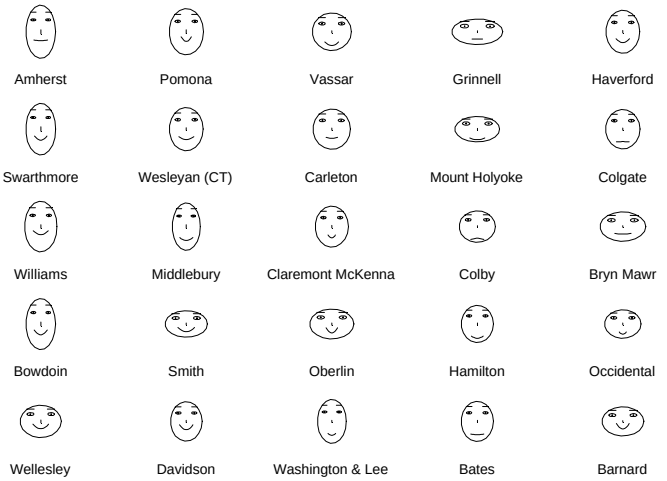
Στα γραφήματα (2.8 και 2.9) που ακολουθούν βλέπουμε τα πρόσωπα του Chernoff για τα σχολεία κλασσικών σπουδών, τα πανεπιστήμια και για τους μέσους των δύο ομάδων αντίστοιχα. Παρατηρείστε πως είναι εμφανείς οι διαφορές ως προς τα χαρακτηριστικά των προσώπων και πως υπάρχει μια έντονη ποικιλία προσώπων. Στο τρίτο γράφημα όπου πια συγκρίνουμε διαφορετικές ομάδες μεταξύ τους είναι εμφανής η διαφορά ανάμεσα στα δύο είδη σχολείων.

Βέβαια για να μπορέσει κανείς να μελετήσει τα πρόσωπα αυτά πρέπει να γνωρίζει τι αντιστοιχεί σε κάθε μεταβλητή. Έτσι στην περίπτωση μας έχουμε πως

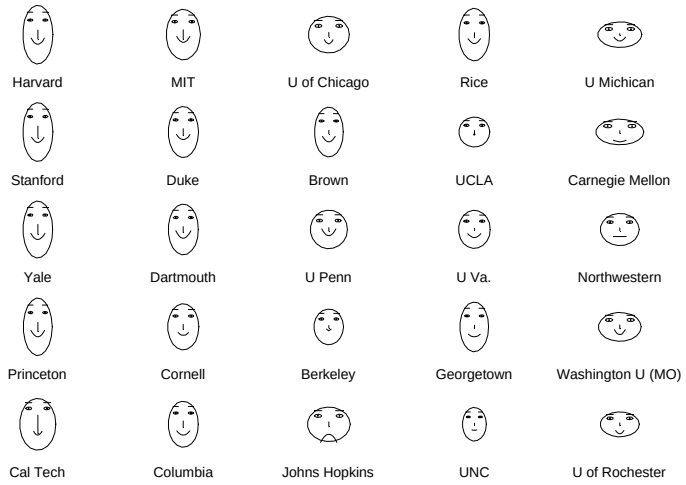
1. *Περιοχή προσώπου = βαθμός SAT*
2. *Σχήμα προσώπου = Ποσοστό φοιτητών που έγιναν δεκτοί*
3. *Μήκος μύτης = Ποσό που ξοδεύει κάθε φοιτητής*
4. *Τοποθεσία στόματος = Ποσοστό τελειόφοιτων με μεγάλο βαθμό*
5. *Καμπύλη χαμόγελου = Ποσοστό διδασκόντων με διδακτορικό*
6. *Μήκος στόματος = Ποσοστό φοιτητών που αποφοιτούν*

Παρατηρείστε πως όπως είχαμε δει και στα starplots η διαφορά στα έξοδα των φοιτητών είναι μεγάλη, το μέγεθος της μύτης είναι πολύ διαφορετικό ανάμεσα στα δύο είδη σχολείων. Θα πρέπει πάντως να παρατηρήσει κανείς πως επειδή οι διαφορές που γίνονται πιο εύκολα και άμεσα κατανοητές είναι αυτές που έχουν να κάνουν με το μέγεθος του προσώπου, συνήθως οι δύο πρώτες μεταβλητές που αντιστοιχούν σε αυτά τα χαρακτηριστικά θα πρέπει να φροντίζουμε να είναι και αυτές που έχουν μεγαλύτερη σημασία, ώστε να είναι εύκολο να αποκαλύπτουν διαφορές.

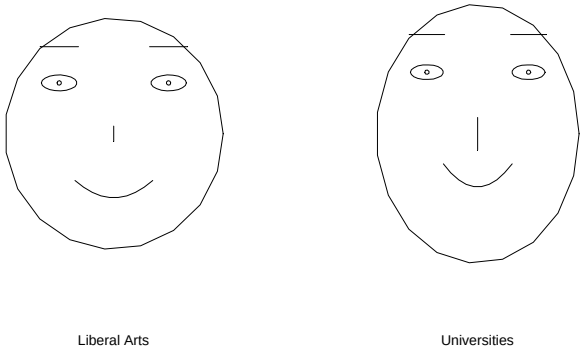
Τα διαγράμματα του Chernoff είναι πολύ χρήσιμα και για την απεικόνιση πολυμεταβλητών χρονοσειρών, στα οποία μπορούμε να δούμε πώς αλλάζει το πρόσωπο με το χρόνο.



Γράφημα 2.8α. Τα πρόσωπα του Chernoff για τα κολέγια κλασσικών σπουδών



Γράφημα 2.8β. Τα πρόσωπα του Chernoff για τα πανεπιστήμια



Γράφημα 2.9. Τα πρόσωπα του Chernoff για τους δύο τύπους σχολείων

Αξίζει να αναφερθεί πως τα πρόσωπα του Chernoff στηρίζονται στη λογική πως ο ανθρώπινος εγκέφαλος είναι καλά εκπαιδευμένος να αναγνωρίζει διαφορές και ομοιότητες σε πρόσωπα, μέρος ίσως μιας υποσυνείδητης ικανότητας του. Αυτό προέτρεψε τον Chernoff να προτείνει αυτόν τον τύπο γραφημάτων. Θα πρέπει όμως να τονιστεί πως για να μπορέσουμε να ξεχωρίσουμε δύο παρατηρήσεις θα πρέπει να υπάρχουν έντονες διαφορές σε ευδιάκριτα χαρακτηριστικά όπως πχ το σχήμα του προσώπου. Δηλαδή έχει σημασία ο τρόπος με τον οποίο αναπαριστούμε κάθε μεταβλητή, ποία μεταβλητή αντιστοιχεί σε ποιο χαρακτηριστικό του προσώπου.

2.1.6 Οι Καμπύλες του Andrews

Οι καμπύλες του Andrews (Andrews' curves) εισήχθηκαν από τον Andrews με σκοπό να απεικονιστούν πολυμεταβλητά δεδομένα..

Για κάθε παρατήρηση σχηματίζουμε την καμπύλη της συνάρτησης

$$f_X(t) = X_1/\sqrt{2} + X_2 \sin t + X_3 \cos t + X_4 \sin(2t) + X_5 \cos(2t) + \ldots, \qquad t \in (-\pi, \pi)$$

για διαφορετικές τιμές του *t* και στη συνέχεια φτιάχνουμε το γράφημα *(t, f((t))* για το διάστημα *(-π, π)*.

Όπως μπορείτε να δείτε η καμπύλη αυτή αποτελείται από ημίτονα και συνημίτονα έχει δηλαδή μια περιοδικότητα. Αυτή η περιοδικότητα όμως εξαρτάται από τις τιμές των μεταβλητών και επομένως για διαφορετικές παρατηρήσεις περιμένουμε και διαφορετικές καμπύλες. Επομένως με τη χρήση αυτών των καμπυλών ελπίζουμε πως απεικονίζοντας πολυμεταβλητές παρατηρήσεις θα μπορέσουμε να δούμε πόσο και ποιες από αυτές διαφέρουν, καθώς παρατηρήσεις με διαφορετικές τιμές σε κάθε μεταβλητή θα έχουν διαφορετικά χαρακτηριστικά (πχ διαφορετική περιοδικότητα, μεγαλύτερη συχνότητα, κλπ).

Παρά την απλή σχετικά ιδέα υπάρχουν κάποια προβλήματα που έχει να αντιμετωπίσει ο ερευνητής. Μερικά ενδιαφέροντα σημεία για τις καμπύλες αυτές είναι τα εξής

- Η επιλογή της σειράς με την οποία οι μεταβλητές θα χρησιμοποιηθούν. Δεδομένου πως η σειρά καθορίζει τη σημαντικότητα κάθε μεταβλητής στη δημιουργία της καμπύλης, οι μεταβλητές τοποθετούνται με φθίνουσα σειρά διακύμανσης, δηλαδή η X_1 είναι η μεταβλητή με τη μεγαλύτερη διακύμανση, η X_2 η μεταβλητή με τη δεύτερη μεγαλύτερη διακύμανση και ούτω καθεξής. Εναλλακτικά, θα μπορούσε να χρησιμοποιήσει κανείς τις κύριες συνιστώσες που θα δούμε όμως αργότερα
- Παρατηρήσεις που είναι σχετικά ίδιες μεταξύ τους θα έχουν σχετικά ίδιες καμπύλες και επομένως μπορούμε να διακρίνουμε ομάδες παρατηρήσεων. Αντίθετα παρατηρήσεις που διαφέρουν θα έχουν πολύ διαφορετικές καμπύλες
- Επομένως η μέθοδος είναι ικανή να βρει ακραίες παρατηρήσεις (outliers) από ένα σύνολο παρατηρήσεων. Αν δηλαδή υπάρχει κάποια(ες) παρατήρηση (εις) για την οποία η καμπύλη είναι ολότελα διαφορετική αυτό σημαίνει πως η παρατήρηση αυτή είναι πολύ διαφορετική από τις άλλες και άρα είναι πιθανό outlier.
- Δυστυχώς αν έχουμε ένα μεγάλο αριθμό παρατηρήσεων, και δεδομένου πως για κάθε παρατήρηση έχουμε μια καμπύλη το γράφημα που θα πάρουμε δεν θα είναι καθόλου χρήσιμο, εκτός και αν κάποιες παρατηρήσεις είναι πολύ μακριά από τις άλλες και ξεχωρίζουν εμφανώς
- Είναι ευνόητο πως εναλλακτικά κανείς μπορεί να σχηματίσει τις καμπύλες για τις μέσες τιμές υποομάδων και να συγκρίνει πια υποομάδες. Ουσιαστικά οι καμπύλες αυτές είναι ένα ισχυρό εργαλείο για την περιγραφή πολυμεταβλητών δεδομένων.
- Είναι πολύ σημαντικό πως οι καμπύλες Andrews έχουν την εξής ιδιότητα: η απόσταση ανάμεσα σε δύο καμπύλες είναι ανάλογη της ευκλείδειας απόστασης ανάμεσα στις παρατηρήσεις. Επομένως οι καμπύλες Andrews αναπαριστούν τις διαφορές που υπάρχουν ανάμεσα στις παρατηρήσεις.

Εν κατακλείδι, οι καμπύλες αυτές μας προσφέρουν έναν εύκολο τρόπο να πάρουμε μια πρώτη εικόνα από τα δεδομένα και κυρίως να δούμε κατά πόσο οι παρατηρήσεις μας μοιάζουν μεταξύ τους και αν υπάρχουν outliers στα δεδομένα μας

Παράδειγμα 2.2.

Στο παράδειγμα που ακολουθεί για 20 αμερικάνικες πόλεις έχουν καταγράφει μια σειρά από κοινωνικοοικονομικά χαρακτηριστικά και συγκεκριμένα, ο αριθμός καταστημάτων (σε χιλιάδες), ο αριθμός πρατηρίων χονδρικής (σε εκατοντάδες), το μέσο οικογενειακό εισόδημα (σε χιλιάδες δολάρια), ο δείκτης εξόδων και ο πληθυσμός (σε εκατομμύρια). Τα δεδομένα υπάρχουν στον πίνακα 2.3.

Αριθμός καταστημάτων (σε χιλιάδες)	**Αριθμός πρατηρίων χονδρικής (σε χιλιάδες)**	**Μέσο οικογενειακό εισόδημα (σε χιλιάδες δολάρια)**	**Δείκτης εξόδων**	**Πληθυσμός**
0.73	3.94	6.30	1.24	0.91
1.05	3.03	4.69	0.76	1.06
0.82	4.41	3.67	1.01	0.90
1.13	2.77	9.36	1.37	1.02
0.89	5.10	9.38	1.14	0.95
0.86	2.09	5.88	1.38	1.04
0.97	2.22	5.77	0.76	0.93
0.85	2.13	8.29	0.97	0.91
0.83	2.83	4.43	1.70	1.02
1.00	0.97	8.56	1.11	0.95
1.21	4.12	7.23	0.99	1.09
1.02	2.63	4.69	1.42	1.05
0.76	4.45	8.49	1.06	0.90
1.30	2.00	11.71	1.33	1.17
1.03	0.48	6.07	1.20	0.99
1.07	2.08	9.12	2.08	1.11
0.82	2.50	6.66	0.86	0.72
0.73	3.31	8.62	1.36	0.72
0.87	5.18	10.38	0.57	0.75
1.02	3.43	6.44	1.89	1.04

Πίνακας 2.3. Δεδομένα παραδείγματος 2.2

Για τα δεδομένα του πίνακα 2.3. βρήκαμε πως οι διακυμάνσεις των μεταβλητών είναι

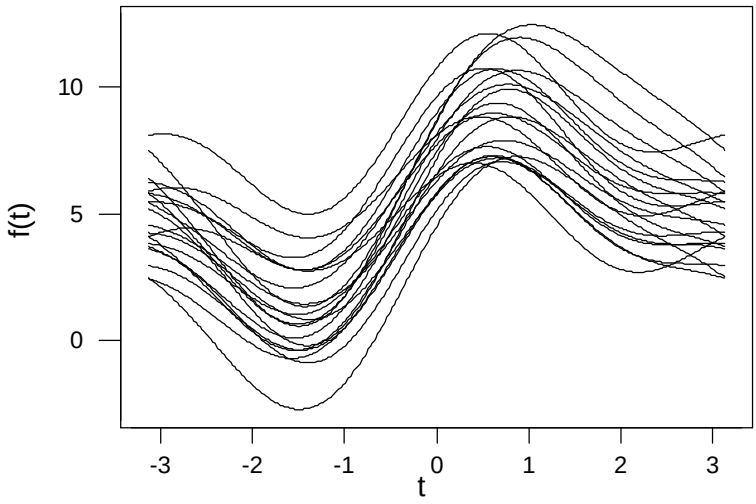
Μεταβλητή	**Αριθμός καταστημάτων (σε χιλιάδες)**	**Αριθμός πρατηρίων χονδρικής (σε χιλιάδες)**	**Μέσο οικογενειακό εισόδημα (σε χιλιάδες δολάρια)**	**Δείκτης εξόδων**	**Πληθυσμός**
	X_1	X_2	X_3	X_4	X_5
Διακύμανση	0.0250063	1.61200	4.71420	0.142968	0.0155292

Πίνακας 2.4. Διακυμάνσεις δεδομένων πίνακα 2.3

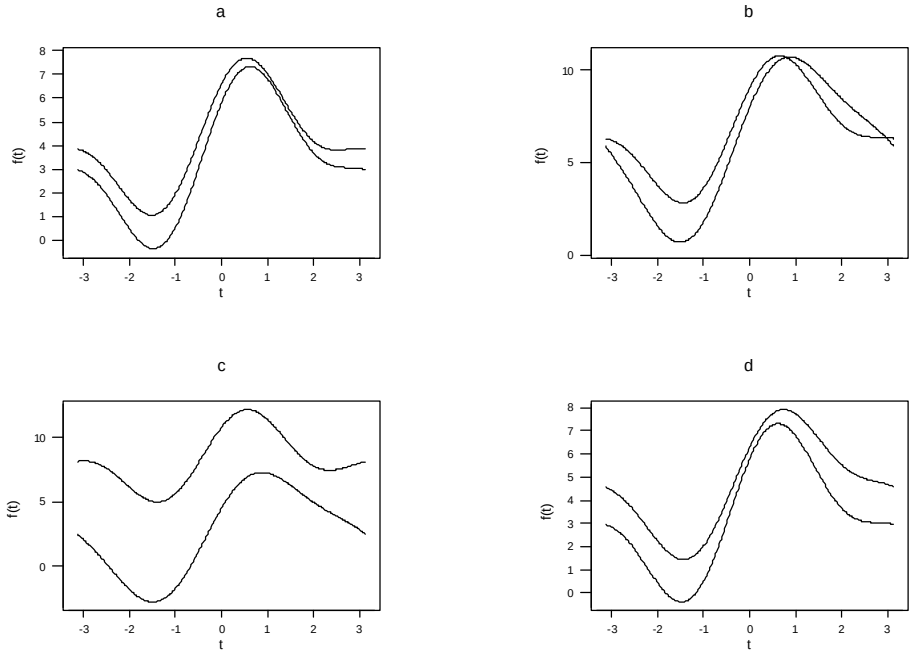
Επομένως η σειρά τους με φθίνουσα διακύμανση είναι X_3, X_2, X_4, X_1 και X_5 και επομένως η καμπύλη για οποιαδήποτε παρατήρηση υπολογίζεται ως

$$f_X(t) = X_3/\sqrt{2} + X_2 \sin t + X_4 \cos t + X_1 \sin(2t) + X_5 \cos(2t),\ t \in (-\pi, \pi)$$

Στο γράφημα 2.10 μπορεί κανείς να δει τις 20 καμπύλες για όλες τις παρατηρήσεις. Παρατηρείστε το ενδιαφέρον σημείο πως δεν φαίνεται να υπάρχουν αφενός πόλεις που μοιάζουν πολύ μεταξύ τους και αφετέρου δεν φαίνεται να υπάρχει κάποια πόλη με μεγάλη διαφορά από τις υπόλοιπες. Παρατηρείστε πως η πόλη 14 που αποκλίνει περισσότερο από τις άλλες στην περιοχή (-2, -1) σε άλλα σημεία της καμπύλης ταυτίζεται με τις υπόλοιπες.



Γράφημα 2.10. Οι καμπύλες του Andrews για τα δεδομένα των αμερικάνικων πόλεων



Γράφημα 2.11. Σύγκριση των καμπυλών του Andrews για επιλεγμένα ζεύγη παρατηρήσεων

Ενδιαφέρον όμως παρουσιάζει και το γράφημα 2.11 όπου υπάρχουν οι καμπύλες για επιλεγμένα ζεύγη παρατηρήσεων

Στο γράφημα μπορεί κανείς να δει 4 ζεύγη παρατηρήσεων και συγκεκριμένα τα ζεύγη (7,13, γράφημα a), (5,14, γράφημα b), (4,15, γράφημα c), (13,18, γράφημα d). Από το γράφημα c μπορούμε να δούμε τις 2 παρατηρήσεις που διαφέρουν περισσότερο. Παρατηρείστε ότι σχεδόν πουθενά οι καμπύλες δεν πλησιάζουν η μια την άλλη αρκετά κοντά. Από την άλλη στο γράφημα a βλέπουμε τις 2 παρατηρήσεις που είναι πιο κοντά η μια στην άλλη, οι καμπύλες τους κινούνται πολύ κοντά. Το γράφημα b έχει το ενδιαφέρον σημείο πως οι καμπύλες τέμνονται σε κάποιο σημείο κοντά στο *t=1,* ενώ στο γράφημα d παρατηρούμε πως οι καμπύλες έρχονται κοντά και μετά απομακρύνονται. Η ποικιλία των καμπυλών που μπορούμε να πάρουμε είναι τεράστια και το γράφημα μπορεί να μας δώσει μια καλή εικόνα του πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους.

2.2 Πολυμεταβλητά Περιγραφικά Μέτρα

2.2.1 Πολυμεταβλητά Δεδομένα

Όπως συζητήσαμε πριν τις περισσότερες φορές τα δεδομένα που έχουμε είναι από τη φύση τους πολυμεταβλητά, δηλαδή για κάθε παρατήρηση έχουμε περισσότερες από μια μεταβλητές. Συνήθως τα δεδομένα οργανώνονται λοιπόν υπό τη μορφή πινάκων δεδομένων. Για τα περισσότερα στατιστικά πακέτα τα δεδομένα έχουν την εξής μορφή

	Χαρακτηριστικό ή Μεταβλητή 1	Χαρακτηριστικό ή Μεταβλητή 2	…	Χαρακτηριστικό ή Μεταβλητή p
Παρατήρηση 1	x_{11}	x_{12}	…	x_{1p}
Παρατήρηση 2	x_{21}	x_{22}	…	x_{2p}
•	.	.		.
•	.	.		.
•	.	.		.
Παρατήρηση n	x_{n1}	x_{n2}	…	x_{np}

n : Αριθμός παρατηρήσεων / αντικειμένων προς μελέτη (Objects or Items)

p : Αριθμός χαρακτηριστικών / μεταβλητών προς μελέτη (Variables)

x_{ij} : Τιμή του i αντικειμένου στο j χαρακτηριστικό / μεταβλητή

Ο πίνακας $\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$ είναι διάστασης $n \times p$ και λέγεται «Μήτρα / Πίνακας Δεδομένων» (Data Matrix).

Παράδειγμα 2.3.

Έστω πως έχουμε παρατηρήσεις για 4 φοιτητές σχετικά με το ύψος, το βάρος και την ηλικία τους

	Ύψος σε cm	Βάρος σε κιλά	Ηλικία σε χρόνια
$1^{ος}$ φοιτητής	167	67	21
$2^{ος}$ φοιτητής	178	75	22
$3^{ος}$ φοιτητής	162	52	20
$4^{ος}$ φοιτητής	190	85	21

Χρησιμοποιώντας την προηγούμενη ορολογία έχουμε:

$n = 4$ *(αριθμός παρατηρήσεων)*

p=3 (μεταβλητές-χαρακτηριστικά)

Και ο πίνακας δεδομένων δίδεται ως $\mathbf{X} = \begin{bmatrix} 167 & 67 & 21 \\ 178 & 75 & 22 \\ 162 & 52 & 20 \\ 190 & 85 & 21 \end{bmatrix}$.

Σε αυτό το σημείο θα πρέπει να κάνουμε μια σημαντική παρατήρηση. Τώρα πια για κάθε παρατήρηση/άτομο, δεν έχουμε μόνο μια τιμή, έχουμε ένα διάνυσμα τιμών, δηλαδή πολλές μεταβλητές των οποίων τις τιμές χρησιμοποιούμε για να φτιάξουμε το διάνυσμα-παρατήρηση. Συνήθως από τη γραμμική άλγεβρα κάθε παρατήρηση είναι ένα διάνυσμα $\mathrm{p} \times 1$, δηλαδή το διάνυσμα στήλη. Επομένως αν με x_1 συμβολίσουμε την πρώτη παρατήρηση, δηλαδή το διάνυσμα με τις τιμές των μεταβλητών για τον πρώτο φοιτητή, τότε προκύπτει πως $x_1 = \begin{bmatrix} 167 \\ 67 \\ 21 \end{bmatrix}$.

Αυτό τον τρόπο παρουσίασης θα χρησιμοποιήσουμε από εδώ και μπρος. Αυτό που πρέπει να τονιστεί είναι πως τελικά ο πίνακας δεδομένων απαρτίζεται από τα διανύσματα γραμμή και επομένως η κατασκευή του είναι με τη μορφή

$$\mathbf{X} = \begin{bmatrix} x_1^{'} \\ x_2^{'} \\ \cdots \\ x_n^{'} \end{bmatrix}$$

Συνεπώς $\mathbf{x}_i^{'}=(x_{i1}, x_{i2}, \ldots, x_{ip})$ είναι η *i* στήλη του πίνακα και αντιστοιχεί στις τιμές των μεταβλητών για την *i* παρατήρηση. Επομένως είναι βασικό να κατανοήσει κανείς πως ο πίνακας δεδομένων, όπως αυτός εμφανίζεται σχεδόν σε όλα τα στατιστικά πακέτα περιέχει τα

διανύσματα γραμμή ως παρατηρήσεις. Από εδώ και στο εξής όταν μιλάμε για μεμονωμένες παρατηρήσεις θα εννοούμε τα διανύσματα στήλη , αλλά όταν μιλάμε για πίνακα δεδομένων έχουμε χρησιμοποιήσει τα διανύσματα γραμμή.

Έχοντας λοιπόν κατά νου αυτόν τον τρόπο παρουσίασης ας δούμε διάφορα περιγραφικά μέτρα θέσης και μεταβλητότητας.

2.2.2 Μέτρα Θέσης

Το πιο γνωστό και πιο διαδεδομένο μέτρο θέσης είναι η μέση τιμή. Σε πολυδιάστατα δεδομένα η αντίστοιχη γενίκευση είναι το διάνυσμα των μέσων τιμών, δηλαδή τίποτα διαφορετικό από ένα διάνυσμα που περιέχει τις μέσες τιμές για κάθε μια μεταβλητή. Το διάνυσμα των μέσων είναι συνήθως ένα διάνυσμα στήλη, δηλαδή

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \dots \\ \bar{x}_p \end{bmatrix},$$

όπου $\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$, είναι ο μέσος της j μεταβλητής, j=1,2,...,p

Και για το διάνυσμα των μέσων ισχύουν όλα όσα γνωρίζουμε για την απλή μέση τιμή. Ότι δηλαδή η ύπαρξη ακραίων τιμών έχει μεγάλη επίδραση στο διάνυσμα των μέσων ενώ από την άλλη υπάρχουν και πάλι αντίστοιχα του απλού κεντρικού οριακού θεωρήματος που αποδεικνύουν κάποιες χρήσιμες ιδιότητες και του διανύσματος των μέσων τιμών.

Άλλα μέτρα θέσης, διαδεδομένα για την μονομεταβλητή περίπτωση, όπως π.χ. η διάμεσος, ή ο περικομμένος μέσος, είναι πολύπλοκα στο να ορισθούν για πολυμεταβλητά δεδομένα και δεν χρησιμοποιούνται στην πολυμεταβλητή περίπτωση.

Για παράδειγμα η διάμεσος ήταν η τιμή που έκοβε στα δύο τα δεδομένα. Στην πολυμεταβλητή περίπτωση κάτι τέτοιο δεν είναι εύκολο να γίνει. Φανταστείτε πως μιλάμε για διμεταβλητά δεδομένα, έχουμε μόνο δύο μεταβλητές. Ως προς ποιόν άξονα θα ενδιαφερθούμε να κόβει τις παρατηρήσεις σε δύο ίσια κομμάτια; Μια απλή προσέγγιση θα ήταν να κατασκευάσουμε κατά αναλογία του διανύσματος των μέσων, ένα διάνυσμα με τις διάμεσους κάθε μεταβλητής. Αυτό το διάνυσμα δεν έχει όμως τις ιδιότητες της διαμέσου.

Ο περικομμένος μέσος (trimmed mean) αποτελεί ένα εναλλακτικό μέτρο θέσης ιδιαίτερα όταν υπάρχουν ακραίες τιμές που επηρεάζουν σημαντικά τη μέση τιμή. Ο περικομμένος μέσος προκύπτει όταν αφαιρέσουμε από το δείγμα μας έναν αριθμό ακραίων παρατηρήσεων και από τις δύο ουρές της κατανομής. Με αυτό τον τρόπο μειώνουμε, αν δεν

εξαφανίζουμε, την επίδραση ακραίων τιμών. Για να εφαρμόσουμε την ιδέα στην πολυμεταβλητή περίπτωση, πρέπει να αποφασίσουμε ποιες παρατηρήσεις θα αγνοήσουμε, δηλαδή ποιες είναι οι ακραίες παρατηρήσεις. Ο τρόπος όμως με τον οποίο θα ορίσουμε τις ακραίες τιμές σε πολυμεταβλητά δεδομένα δεν είναι τόσο απλός.

Η λύση και πάλι είναι να βρούμε περικομμένους μέσους για κάθε μεταβλητή και απλά να συνθέσουμε μετά το διάνυσμα των περικομμένων μέσων. Στη ουσία δηλαδή ορίζουμε τις ακραίες τιμές για κάθε μεταβλητή ξεχωριστά και επομένως αγνοούμε τις όποιες συσχετίσεις.

2.2.3 Μέτρα Μεταβλητότητας

Είναι γνωστό πως η διακύμανση αποτελεί το βασικό μέτρο μεταβλητότητας στη μονοδιάστατη περίπτωση. Επίσης είναι γνωστό πως για ζεύγη μεταβλητών μπορώ να ορίσω τη συνδιακύμανση ένα μέτρο μεταβλητότητας ως προς δύο μεταβλητές που μου λεει πως συμμεταβάλλονται αυτές.

Πιο συγκεκριμένα η δειγματική συνδιακύμανση των μεταβλητών X_j και X_k ορίζεται ως:

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)\left(x_{ik} - \overline{x}_k\right)$$

Είναι εύκολο να δει κανείς πως η διακύμανση δεν είναι παρά η συνδιακύμανση μιας μεταβλητής με τον εαυτό της.

Σε πολυμεταβλητά δεδομένα μπορώ να ορίσω τον πίνακα διακυμάνσεων συνδιακυμάνσεων (Covariance Matrix) ο οποίος είναι ένας συμμετρικός θετικά ορισμένος πίνακας που έχει στη διαγώνιο τις διακυμάνσεις των μεταβλητών και στα υπόλοιπα στοιχεία τις συνδιακυμάνσεις των μεταβλητών που αντιστοιχούν σε κάθε γραμμή και στήλη. Δηλαδή ο πίνακας **S** έχει τη μορφή

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix},$$

όπου $s_j^2 = s_{jj}$ είναι η δειγματική διακύμανση της μεταβλητής X_j

Μερικές *ιδιότητες* του πίνακα **S** είναι οι εξής:

- Ο πίνακας συνδιακύμανσης είναι συμμετρικός. Επομένως ισχύει πως $s_{ij} = s_{ji}$. Για αυτό το λόγο και για καλύτερη παρουσίασή του μπορούμε να μην εμφανίζουμε τα στοιχεία πάνω (ή κάτω) από τη διαγώνιο

- Ο πίνακας συνδιακύμανσης είναι θετικά ημιορισμένος (δηλαδή $|\mathbf{S}| \geq 0$)
- Τα στοιχεία της διαγωνίου του πίνακα είναι υποχρεωτικά θετικοί αριθμοί αφού είναι διακυμάνσεις, ενώ τα μη διαγώνια στοιχεία, αφού είναι συνδιακυμάνσεις μπορεί να είναι αρνητικοί αριθμοί

Για παράδειγμα οι επόμενοι πίνακες

$$\begin{bmatrix} -3 & 1 \\ 1 & -2 \end{bmatrix}, \quad \begin{bmatrix} 2 & -3 \\ -3 & 2 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0.5 \\ 0.7 & 1 \end{bmatrix}$$

δεν είναι πίνακες διακύμανσης καθώς έχουν μη θετικά διαγώνια στοιχεία, αρνητική ορίζουσα και δεν είναι συμμετρικός αντίστοιχα.

Ένα βασικό μ*ειονέκτημα* του πίνακα **S** ως μέτρο μεταβλητότητας είναι πως έχοντας πολλές μεταβλητές χρειαζόμαστε πολλές τιμές (διακυμάνσεις, συνδιακυμάνσεις) για να μελετήσουμε την συνολική μεταβλητότητα. Δηλαδή δεν μπορούμε να ποσοτικοποιήσουμε τη μεταβλητότητα με έναν μόνο αριθμό.

Για να το επιτύχουμε αυτό χρειαζόμαστε κάποια συνάρτηση του πίνακα που να μας δείχνει με έναν αριθμό τη μεταβλητότητα που υπάρχει. Υπάρχουν δύο τέτοια μέτρα

- Συνολική διακύμανση (Total variation): $tr(\mathbf{S}) = \sum_{i=1}^{p} S_{ii}$ και
- Γενικευμένη διακύμανση (Generalized variance): $|\mathbf{S}|$: Ορίζουσα του πίνακα **S**.

Είναι προφανές πως όσο μεγαλύτερη τιμή βρούμε και για τα δύο μέτρα τόσο μεγαλύτερη είναι η μεταβλητότητα στα δεδομένα μας. Και τα δύο μέτρα μετράνε το πόσο απλωμένες από το διάνυσμα των μέσων είναι οι παρατηρήσεις μας.

Το μειονέκτημα της συνολικής διακύμανσης είναι ότι χρησιμοποιούμε μόνο τα διαγώνια στοιχεία, δηλαδή τις διακυμάνσεις, χάνοντας έτσι πληροφορία από τις συνδιακυμάνσεις.

Για τη γενικευμένη διακύμανση, μηδενική τιμή ορίζουσας υποδηλώνει ότι κάποιες μεταβλητές είναι γραμμικά εξαρτημένες μεταξύ τους. Η γενικευμένη διακύμανση είναι ένα χρήσιμο μέτρο πολυμεταβλητής μεταβλητότητας καθώς έχει την ερμηνεία ότι μετράει τον 'όγκο' που αντιστοιχεί στον πίνακα **S** και ερμηνεύει την συνολική μεταβλητότητα των δεδομένων. Συγκεκριμένα ισχύει πως

$$|S| = \frac{(\text{ογκος})^2}{n^p}$$

δηλαδή η γενικευμένη διακύμανση μετράει κατά κάποιον τρόπο τον όγκο των δεδομένων. Στη μονομεταβλητή περίπτωση ο όγκος είναι το άπλωμα των τιμών, στη διμεταβλητή το εμβαδόν,

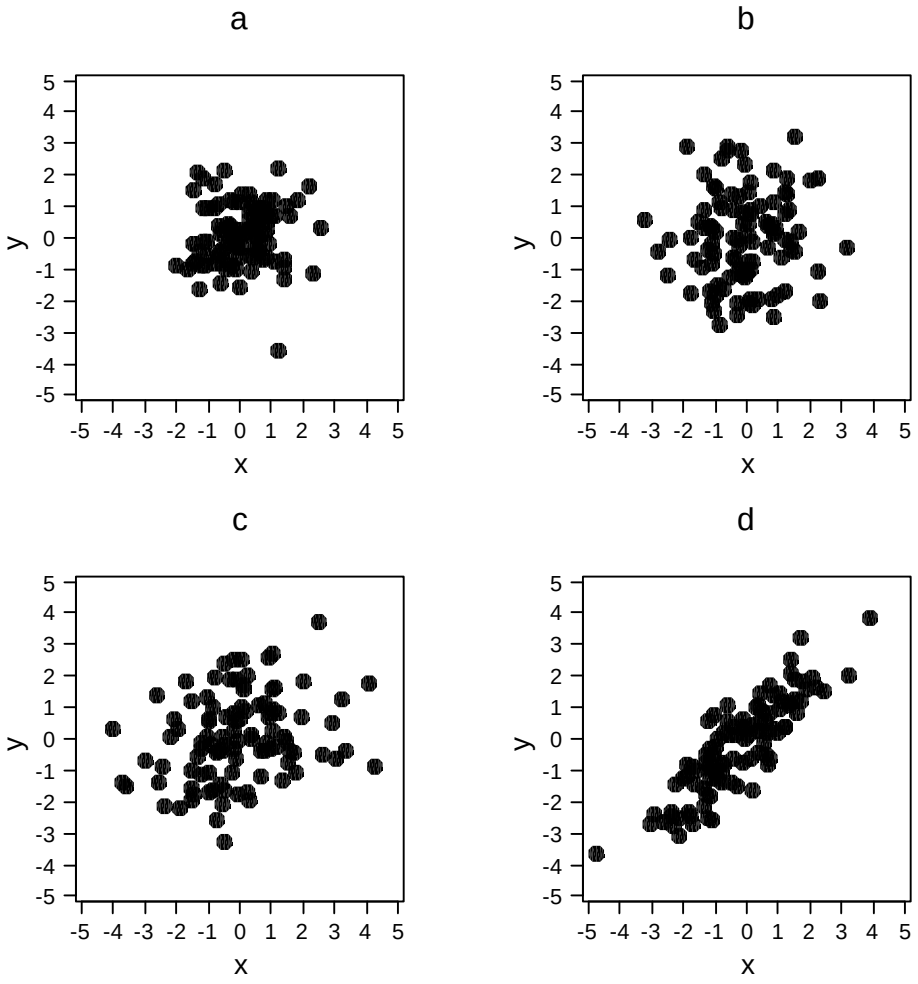
στις τρεις μεταβλητές ο όγκος στο χώρο, για περισσότερες από τρεις μεταβλητές δεν είναι εύκολο να το δούμε.

Για να πάρετε μια ιδέα τα σημεία που εμφανίζονται στα γραφήματα 2.12 αποτελούν δείγματα μεγέθους 100. Οι πίνακες συνδιακυμάνσεων που χρησιμοποιήθηκαν ήταν οι

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1.75 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix} \text{ και } \begin{bmatrix} 2 & 1.65831 \\ 1.65831 & 2 \end{bmatrix}$$

Παρατηρείστε πως από την κατασκευή τους οι 3 τελευταίοι πίνακες έχουν την ίδια ορίζουσα και άρα την ίδια γενικευμένη διακύμανση, παρόλα αυτά έχουν ολότελα διαφορετική δομή, ο δεύτερος πίνακας υποθέτει την ανυπαρξία συσχέτισης ανάμεσα στις μεταβλητές.

Στο διάγραμμα σημείων βλέπουμε πως αν και το εμβαδόν για αυτούς τους τρεις πίνακες είναι παρόμοιο το σχήμα είναι πολύ διαφορετικό, αφού για τον πίνακα 4 υπάρχει έντονη σχέση ανάμεσα στις δύο μεταβλητές και το σχήμα μοιάζει με έλλειψη.



Γράφημα 2.12. Διαγράμματα σημείων για δείγματα μεγέθους 100 . Ο πίνακας διακύμανσης του πληθυσμού είχε την ίδια γενικευμένη διακύμανση για τα γραφήματα b,c,d

Δηλαδή το μειονέκτημα της γενικευμένης διακύμανσης είναι πως αγνοεί τη δομή του πίνακα. Αξίζει να παρατηρήσει κανείς πως γενικά είναι πολύ δύσκολο να αναπαρασταθεί όλη η πολυμεταβλητή μεταβλητότητα με έναν μόνο αριθμό. Επίσης δεδομένου πως βασίζεται στην ορίζουσα, η στατιστική συμπερασματολογία σχετικά με τη γενικευμένη διακύμανση είναι δύσκολη.

Εναλλακτικά μέτρα πολυμεταβλητής μεταβλητότητας μπορούν να οριστούν με τη βοήθεια των ιδιοτιμών του πίνακα διακύμανσης.

2.2.4 Πίνακας Συσχετίσεων R

Ο πίνακας συσχετίσεων είναι ο πίνακας που περιέχει σαν στοιχεία του τους συντελεστές συσχέτισης του Pearson για κάθε ζευγάρι μεταβλητών. Σε αυτό το σημείο θα πρέπει να τονίσουμε τα εξής:

- Ο συντελεστής συσχέτισης του Pearson μετράει μόνο γραμμική συσχέτιση ανάμεσα στις μεταβλητές και επομένως δεν μπορεί να μας δώσει πληροφορία για άλλης μορφής συσχέτιση.
- Υπάρχουν πολλοί άλλοι συντελεστές συσχέτισης στη βιβλιογραφία που είτε μετρούν άλλης μορφής συσχέτιση, πχ ο συντελεστής του Spearman είναι κατάλληλος για κάθε μορφή μονότονης συσχέτισης, είτε για άλλης μορφής δεδομένα. Ο συντελεστής συσχέτισης του Pearson είναι κατάλληλος μόνο για ζεύγη ποσοτικών μεταβλητών.

Επομένως στη γενική του μορφή ο πίνακας συσχετίσεων ορίζεται ως

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix},$$

όπου $r_{jk} = \dfrac{S_{jk}}{S_j \cdot S_k} = \dfrac{S_{jk}}{\sqrt{S_{jj}^2} \cdot \sqrt{S_{kk}^2}}$, j,k=1,2,...p.

δηλαδή, το στοιχείο r_{jk} είναι, ο συντελεστής συσχέτισης του Pearson μεταξύ των μεταβλητών X_j και X_k.

Ο πίνακας έχει απαραίτητα τιμές ίσες με τη μονάδα στη διαγώνιο, είναι συμμετρικός και κανένα στοιχείο του δεν μπορεί να πάρει τιμή μεγαλύτερη σε απόλυτη τιμή από το 1. Τιμές –1 και 1 σημαίνουν απόλυτα γραμμική σχέση των δύο μεταβλητών, το πρόσημο υποδηλώνει

την ύπαρξη θετικής ή αρνητικής σχέσης. Η θετική σχέση ερμηνεύεται πως όσο αυξάνει η τιμή της μιας μεταβλητής τόσο αυξάνει και η τιμή της άλλης ενώ η αρνητική σχέση ερμηνεύεται πως όσο αυξάνει η τιμή της μιας μεταβλητής μειώνεται η τιμής της άλλης.

Ο πίνακας διακυμάνσεων-συνδιακυμάνσεων **S** τυποποιημένων μεταβλητών ταυτίζεται με τον πίνακα συσχετίσεων των αρχικών μεταβλητών πριν την τυποποίηση τους. Εύκολα βλέπει κανείς πως:

$$r_{jk} = \frac{S_{jk}}{S_j \cdot S_k} = \frac{\sum_{i=1}^{n}\left(X_{ij} - \bar{x}_j\right)\left(X_{ik} - \bar{x}_k\right)}{n \cdot S_j \cdot S_k} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_{ij} - \bar{x}_j}{S_j}\right)\left(\frac{X_{ik} - \bar{x}_k}{S_k}\right) = Cov\left(X_j^*, X_k^*\right),$$

όπου X_j^*, X_k^* είναι οι τυποποιημένες μεταβλητές.

Σχετικά με ελέγχους υποθέσεων για συντελεστές συσχέτισης θα μιλήσουμε σε επόμενο κεφάλαιο. Αυτό που πρέπει να τονιστεί είναι πως γενικά μη μηδενικές συσχετίσεις δεν σημαίνουν πως υπάρχει και κάποια αξιοσημείωτη σχέση ανάμεσα στις μεταβλητές. Για παράδειγμα η ελεγχοσυνάρτηση που ελέγχει σε ένα δείγμα μεγέθους *n* αν η συσχέτιση του πληθυσμού είναι στατιστικά σημαντικά διαφορετική από το 0 είναι η $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ η οποία ακολουθεί κάτω από τη μηδενική υπόθεση (και την υπόθεση πως τα δεδομένα προέρχονται από διμεταβλητή κανονική κατανομή) κατανομή *t* με *n-2* βαθμούς ελευθερίας. Συνεπώς σε επίπεδο στατιστικής σημαντικότητας 5% απορρίπτω τη μηδενική υπόθεση αν $\left|\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right| \geq t_{n-2,1-a/2}$ ή, λύνοντας ως προς *r,* απορρίπτω όταν

$r^2 \geq \frac{t_{n-2,1-a/2}^2}{n-2+t_{n-2,1-a/2}^2}$, όπου $t_{n,1-a/2}$ είναι το 1-α/2 ποσοστιαίο σημείο της κατανομής *t* με *n* βαθμούς ελευθερίας.

Από την παραπάνω σχέση μπορεί κανείς να υπολογίσει για κάθε μέγεθος δείγματος ποια είναι η τιμή πάνω από την οποία απορρίπτουμε τη μηδενική υπόθεση περί μηδενικής συσχέτισης. Στον πίνακα 2.5 μπορεί κανείς να δει την απόλυτη τιμή του συντελεστή συσχέτισης που μας οδηγεί σε απόρριψη της μηδενικής υπόθεσης πως τα δεδομένα μας είναι ασυσχέτιστα σε επίπεδο στατιστικής σημαντικότητας 5%.

Μέγεθος δείγματος	5	10	15	20	30	50	100	200	500	1000
Τιμή συντελεστή	0.754	0.576	0.482	0.422	0.349	0.273	0.195	0.138	0.088	0.06

Πίνακας 2.5 Η τιμή του συντελεστή συσχέτισης πάνω από την οποία απορρίπτουμε τη μηδενική υπόθεση περί μηδενικής συσχέτισης σε επίπεδο σημαντικότητας 5%.

Από τον πίνακα γίνεται σαφές πως ακόμα και με σχετικά μικρά δείγματα μικρές συσχετίσεις είναι στατιστικά σημαντικές αν και ουσιαστικά αδιάφορες από στατιστικής απόψεως. Θυμηθείτε πως ο συντελεστής συσχέτισης σχετίζεται με τον συντελεστή προσδιορισμού της γραμμικής παλινδρόμηση. Έτσι μια συσχέτιση της τάξης του 0.20, (που για δείγμα μεγέθους 100 είναι στατιστικά σημαντική) σημαίνει πως ο συντελεστής προσδιορισμού σε μια παλινδρόμηση ανάμεσα στις δύο μεταβλητές θα είναι 4%, δηλαδή πάρα πολύ μικρός για οποιαδήποτε στατιστική χρήση. Συνεπώς μας ενδιαφέρουν μεγάλες σε απόλυτη τιμή συσχετίσεις και όχι απαραίτητα στατιστικά σημαντικές συσχετίσεις.

Επίσης είναι χρήσιμο να τονιστεί πως όταν παρουσιάζουμε τον πίνακα συσχετίσεων καλό είναι να παρουσιάζουμε μόνο την κάτω διαγώνιο, ώστε ο πίνακας να είναι πιο ευανάγνωστος. Λόγω της συμμετρίας δεν χάνουμε κάποια πληροφορία. Επίσης όταν είναι δυνατόν θα πρέπει να βάζουμε κάποια μορφή διάταξης στις συσχετίσεις ώστε να μπορούμε σχετικά εύκολα να βρούμε αυτές που πραγματικά είναι μεγάλες και άρα χρήσιμες.

Για καλύτερη ερμηνεία, ειδικά όταν έχουμε πολλές μεταβλητές μπορούμε αντί του πίνακα συσχετίσεων να απεικονίσουμε τα δεδομένα με ένα πίνακα όπου θα φαίνονται μόνο οι σημαντικές συσχετίσεις ή έστω οι συσχετίσεις οι οποίες είναι μεγαλύτερες από μία τιμή πρακτικής σημαντικότητας (όπως η τιμή 0.70 που προτείνουν οι Chatfield και Collins, 1992). Έπειτα μπορούμε να απεικονίσουμε τις σχέσεις γραφικά για να έχουμε μια πιο άμεση εικόνα. Άλλη τεχνική είναι να μειώνουμε τα δεκαδικά στοιχεία σε ένα ή δύο για να μπορούμε να εντοπίσουμε καλύτερα σχέσεις μεταξύ των μεταβλητών.

Ο συντελεστής συσχέτισης έχει επίσης ενδιαφέρουσα γεωμετρική ερμηνεία, καθώς μπορεί να αναπαρασταθεί γραφικά. Έτσι ο συντελεστής συσχέτισης είναι το συνημίτονο της γωνίας που προκύπτει από τα δύο διανύσματα με τις παρατηρήσεις για κάθε μεταβλητή. Για αυτό σε πολλές γραφικές απεικονίσεις αντί να παρουσιάζεται ο πίνακας συσχετίσεων, εμφανίζεται ένα γράφημα όπου κάθε συσχέτιση αναπαρίσταται από μια γωνία. Αν η γωνία είναι 90 μοίρες τότε οι δύο μεταβλητές είναι ασυσχέτιστες, ενώ αν η γωνία έχει 0 μοίρες τότε οι μεταβλητές είναι πολύ έντονα συσχετισμένες.

Τέλος θα πρέπει να σημειωθεί πως η έννοια της συσχέτισης και της ανεξαρτησίας δεν ταυτίζονται. Οι ανεξάρτητες μεταβλητές είναι και ασυσχέτιστες αλλά οι ασυσχέτιστες δεν είναι κατ' ανάγκη και ανεξάρτητες. Όπως θα δούμε αργότερα υπάρχουν κάποιες υποθέσεις που μας εξασφαλίζουν πως οι ασυσχέτιστες μεταβλητές είναι και ανεξάρτητες, αλλά γενικά κάτι τέτοιο δεν ισχύει.

Παράδειγμα 2.2. (συνέχεια)

Με βάση τα δεδομένα του παραδείγματος 2.2 που αφορούσαν τις 20 αμερικάνικες πόλεις υπολογίζουμε τα ακόλουθα περιγραφικά μέτρα:

Το διάνυσμα των μέσων είναι

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \\ \bar{x}_5 \end{bmatrix} = \begin{bmatrix} 0.948 \\ 2.984 \\ 7.286 \\ 1.210 \\ 0.962 \end{bmatrix}$$

Για τον πίνακα διακύμανσης βρίσκουμε πως είναι ο

$$\mathbf{S} = \begin{bmatrix} 0.024 & -0.065 & 0.090 & 0.010 & 0.015 \\ -0.065 & 1.531 & 0.198 & -0.112 & -0.047 \\ 0.090 & 0.198 & 4.479 & 0.000 & -0.011 \\ 0.010 & -0.112 & 0.000 & 0.136 & 0.022 \\ 0.015 & -0.047 & -0.011 & 0.022 & 0.015 \end{bmatrix}$$

Εδώ θα πρέπει να σημειώσουμε το εξής: τα περισσότερα στατιστικά πακέτα υπολογίζουν έναν λίγο διαφορετικό πίνακα διακύμανσης. Συγκεκριμένα ο παρονομαστής δεν είναι *n* αλλά *n-1,* για τον ίδιο λόγο που στην μονομεταβλητή περίπτωση χρησιμοποιούμε την αμερόληπτη διακύμανση. Δηλαδή έχει αποδειχτεί πως ο δειγματικός πίνακας διακύμανσης δεν είναι αμερόληπτος και για αυτό διαιρούμε με *n-1* για να πάρουμε τον αμερόληπτο πίνακα διακύμανσης S^*. Είναι προφανές πως η σχέση που συνδέει τους δύο πίνακες είναι η

$$S^* = \frac{nS}{n-1}$$

Συνεπώς για τα δεδομένα μας βρίσκουμε πως

$$\mathbf{S}^* = \begin{bmatrix} 0.025 & -0.068 & 0.094 & 0.010 & 0.015 \\ -0.068 & 1.612 & 0.206 & -0.118 & -0.049 \\ 0.094 & 0.206 & 4.714 & 0.000 & -0.012 \\ 0.010 & -0.118 & 0.000 & 0.143 & 0.022 \\ 0.015 & -0.049 & -0.012 & 0.022 & 0.015 \end{bmatrix}$$

Παρατηρείστε πως μόνο εκεί που είχαμε μεγάλες τιμές υπάρχουν σαφείς διαφορές ενώ σε μικρότερες τιμές λόγω στρογγυλοποίησης οι τιμές είναι ίδιες.

Ο πίνακας συσχετίσεων είναι ο ίδιος άσχετα αν πάρουμε τον μεροληπτικό ή τον αμερόληπτο πίνακα διακύμανσης. Για τα δεδομένα μας είναι

$$\mathbf{R} = \begin{bmatrix} 1 & -0.342 & 0.277 & 0.164 & 0.752 \\ -0.342 & 1 & 0.076 & -0.246 & -0.311 \\ 0.277 & 0.076 & 1 & 0.000 & -0.043 \\ 0.164 & -0.246 & 0.000 & 1 & 0.488 \\ 0.752 & -0.311 & -0.043 & 0.488 & 1 \end{bmatrix}.$$

Παρατηρείστε λοιπόν πως υπάρχουν κάποιες αρκετά μεγάλες συσχετίσεις όπως αυτή του πληθυσμού με τον αριθμό καταστημάτων που είναι 0.756. Επίσης υπάρχουν και μερικές αρνητικές συσχετίσεις, όπως η συσχέτιση ανάμεσα στον αριθμό καταστημάτων και τον αριθμό πρατηρίων χονδρικής.

Χρησιμοποιώντας τον πίνακα S^* βρίσκουμε πως η συνολική διακύμανση είναι το άθροισμα των διαγώνιων στοιχείων του πίνακα δηλαδή 6.51 και η γενικευμένη διακύμανση είναι 0.00007245. Θα πρέπει να τονιστεί πως τόσο η συνολική όσο και η γενικευμένη διακύμανση ελάχιστα μπορούν να μας φανερώσουν για τα δεδομένα παρά μόνο αν τις χρησιμοποιήσουμε συγκριτικά με άλλα σετ δεδομένων για να δούμε πιο σετ έχει μεγαλύτερη μεταβλητότητα.

2.2.5 Στοιχεία Πινάκων

Ας δούμε τώρα πως μπορούμε να αναπαραστήσουμε με άλγεβρα πινάκων το διάνυσμα των μέσων και τον πίνακα διακύμανσης και συσχετίσεων. Όπως είπαμε και πριν κάθε παρατήρηση είναι ένα διάνυσμα στήλη. Επομένως η *i* παρατήρηση έχει τη μορφή:

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \dots \\ x_{ip} \end{bmatrix} \text{ και σαν διάνυσμα γραμμή } \mathbf{x}_i' = \begin{bmatrix} x_{i1} & \dots & x_{ip} \end{bmatrix}.$$

Μπορεί κανείς να δει πως

$\mathbf{x}_i{}'\mathbf{x}_i = \sum_{j=1}^{p} x_{ij}^2$ δηλαδή το άθροισμα τετραγώνων όλων των στοιχείων μεταβλητών και είναι φυσικά ένας μόνο αριθμός

Αντίθετα κάνοντας την πράξη

$$\mathbf{x}_i\mathbf{x}_i{}' = \begin{bmatrix} x_{i1}{}^2 & x_{i1}x_{i2} & \cdots & x_{i1}x_{ip} \\ x_{i1}x_{i2} & x_{i2}{}^2 & & \vdots \\ \vdots & & \ddots & \\ x_{i1}x_{ip} & \cdots & \cdots & x_{ip}{}^2 \end{bmatrix}$$

παίρνουμε έναν πίνακα με όλα τα γινόμενα των μεταβλητών μεταξύ τους.

Ο πίνακας δεδομένων $\mathbf{X}_{n\times p}$ είναι ο πίνακας που έχει γραμμές του τα διανύσματα γραμμές κάθε παρατήρησης.

Με βάση αυτά μπορούμε να υπολογίσουμε το διάνυσμα μέσων τιμών

$$\bar{\mathbf{x}} = \frac{1}{n}\underset{p\times n}{\mathbf{X}'}\cdot\underset{n\times 1}{\mathbf{1}}$$

όπου **1** είναι το διάνυσμα στήλη με όλα τα στοιχεία του ίσα με 1.

Ομοίως βρίσκουμε και για τον πίνακα διακύμανσης πως

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\cdot\mathbf{X} - \bar{\mathbf{x}}\cdot\bar{\mathbf{x}}'$$

ενώ για τον αμερόληπτο πίνακα έχουμε πως

$$\mathbf{S}^* = \frac{1}{n-1}\mathbf{X}'\cdot\mathbf{X} - \frac{n}{n-1}\bar{\mathbf{x}}\cdot\bar{\mathbf{x}}'$$

Επίσης ο πίνακας **S** μπορεί να γραφτεί ως

$$\mathbf{S} = \frac{1}{n}\left(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'\right)'\left(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'\right)$$

και ανάλογα προκύπτει και η αμερόληπτη εκτιμήτρια απλά διαιρώντας δια *n-1*.

Τέλος, για να κατασκευάσουμε τον πίνακα συσχετίσεων από τον πίνακα διακυμάνσεων αρκεί η σχέση

$$\mathbf{R} = \mathbf{D}^{-1}\cdot\mathbf{S}\cdot\mathbf{D}^{-1},$$

και άρα ισοδύναμα

$$\mathbf{S} = \mathbf{D}\cdot\mathbf{R}\cdot\mathbf{D}$$

όπου ο πίνακας **D** είναι διαγώνιος και ορίζεται ως

$$D = \begin{bmatrix} \sqrt{S_1^2} & 0 & \dots & 0 \\ 0 & \sqrt{S_2^2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sqrt{S_p^2} \end{bmatrix}.$$

2.2.6 Μέτρα Πολυμεταβλητής Ασυμμετρίας και Κύρτωσης

Στη μονοδιάστατη περίπτωση τα μέτρα ασυμμετρίας και κύρτωσης βασίστηκαν στις κεντρικές ροπές και συγκεκριμένα την τρίτη και την τέταρτη κεντρική ροπή αντίστοιχα. Τα μέτρα αυτά ορίζονται συνήθως ως

$$\beta_1 = \frac{m_3}{s^3} \text{ και } \beta_2 = \frac{m_4}{s^4}$$

αντίστοιχα για την ασυμμετρία και την κύρτωση. Για την ασυμμετρία το πρόσημο του συντελεστή μαρτυρά το είδος της ασυμμετρίας ενώ για την κύρτωση συνήθως συμπεραίνουμε για αυτή συγκρίνοντας με την κύρτωση της κανονικής κατανομής που είναι 3 με τον παραπάνω ορισμό. Σε πολλά βιβλία για να είναι η κύρτωση της κανονικής κατανομής ορόσημο αφαιρούμε από το συντελεστή β_2 το 3 ώστε η κύρτωση της κανονικής να είναι 0. Από τον ορισμό μπορεί κανείς να δει πως η ασυμμετρία και η κύρτωση δεν είναι παρά οι $3^{ες}$ και $4^{ες}$ ροπές των τυποποιημένων δεδομένων.

Η έννοια της ασυμμετρίας και της κύρτωσης σε μεγαλύτερες διαστάσεις δεν είναι και τόσο προφανής. Ουσιαστικά όταν μιλάμε για συμμετρία αναφερόμαστε σε 'σφαιρική συμμετρία' δηλαδή συμμετρία ως προς όλες τις διαστάσεις. Η κύρτωση στις πολλές διαστάσεις δεν έχει κάποια ερμηνεία αλλά και πάλι μας επιτρέπει σύγκριση με την πολυμεταβλητή κανονική κατανομή. Τα μέτρα λοιπόν πολυμεταβλητής ασυμμετρίας και κύρτωσης που ορίστηκαν από τον Mardia το 1973 είναι τα εξής:

Έστω πως x_i είναι ένα μια παρατήρηση (διάνυσμα p×1 διαστάσεων) και έστω πως έχουμε *n* παρατηρήσεις. Ας ορίσουμε τον πίνακα $\mathbf{Z}$ του οποίου τα στοιχεία Z_{rs}, *r,s=1,. . . , n* υπολογίζονται ως

$$Z_{rs} = (x_r - \bar{x})' S^{-1} (x_s - \bar{x})$$

τότε οι συντελεστές ασυμμετρίας και κύρτωσης ορίζονται ως

$$\beta_{1,p} = \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} Z_{rs}^3 \qquad \text{ασυμμετρία}$$

και

$$\beta_{2,p} = \frac{1}{n} \sum_{s=1}^{n} Z_{ss}^2 \qquad \text{κύρτωση}$$

Ας δούμε λίγο τι τελικά μετράνε αυτές οι ποσότητες:

- Κατά αρχάς ο πίνακας **Z** είναι συμμετρικός. Αν προς το παρόν ξεχάσουμε την ύπαρξη του πίνακα **S** (ουσιαστικά ο σκοπός του είναι να διορθώνει ως προς την κλίμακα ώστε οι συντελεστές να είναι συγκρίσιμοι αλλά και ως προς τις συσχετίσεις για να μειώνεται η επίδραση κάποιων συσχετίσεων) τότε βλέπουμε πως ο πίνακας **Z** περιέχει τα cross-products μεταξύ των παρατηρήσεων. Αν τα δεδομένα είναι συμμετρικά περιμένουμε πως αυτά τα cross-products θα ακυρώνονται, άλλα θα είναι θετικά άλλα αρνητικά (μην ξεχνάμε πως έχουμε αφαιρέσει τη μέση τιμή) και επομένως το άθροισμα τους θα μας δίνει μια εικόνα σχετικά με την ασυμμετρία. Για να γίνει αυτό πιο κατανοητό φανταστείτε 3 παρατηρήσεις. Αν κρατήσουμε τη μια σταθερή και πάρουμε τα crossproducts των άλλων δύο ως προς αυτήν επειδή τα σημεία είναι συμμετρικά ως προς το διάνυσμα των μέσων, κάποια θα δώσουν αρνητική τιμή ενώ κάποια άλλα θετική αλλά λόγω της συμμετρίας οι αρνητικές τιμές θα ακυρώνονται από τις θετικές και το τελικό αποτέλεσμα θα είναι κοντά στο 0.

- Παρατηρείστε πως και οι δυο ποσότητες αν p=1 ταυτίζονται με τους συντελεστές στη μια διάσταση. Ο συντελεστής ασυμμετρίας που υπολογίζουμε στην πολυμεταβλητή περίπτωση είναι ισοδύναμος με το τετράγωνου του απλού συντελεστή ασυμμετρίας, δηλαδή παίρνει μόνο θετικές τιμές. Για το συντελεστή ασυμμετρίας ουσιαστικά παίρνουμε την τιμή β_1^2 και άρα στην πολυμεταβλητή περίπτωση δεν μπορούμε να μιλάμε για θετική ή αρνητική ασυμμετρία

- Οι συντελεστές δεν επηρεάζονται από αλλαγές στην κλίμακα και οποιονδήποτε γραμμικό μετασχηματισμό στα δεδομένα.

- Επειδή και πάλι πρέπει να ορίσουμε ένα ορόσημο ώστε να μπορούμε να ερμηνεύουμε την κύρτωση, μπορούμε να χρησιμοποιήσουμε ως μέτρο σύγκρισης την τιμή της πολυμεταβλητής κανονικής κατανομής. Επομένως για την πολυμεταβλητή κανονική κατανομή *p* διαστάσεων ο συντελεστής κύρτωσης είναι *p(p+2)* και επομένως με βάση αυτή την τιμή μπορούμε να δούμε αν η κατανομή μας είναι πιο κυρτή από την πολυμεταβλητή κανονική κατανομή ή όχι. Σε κάθε περίπτωση είναι δύσκολη η ερμηνεία της κύρτωσης στις πολλές διαστάσεις.

- Το ίδιο, όσο αφορά την ερμηνεία, συμβαίνει και στην ασυμμετρία. Στις πολλές διαστάσεις το πρόσημο της ασυμμετρίας δεν είναι ιδιαίτερα ενδεικτικό για τη μορφή των δεδομένων παρά μόνο στο ότι δεν σχηματίζουν μια υπερσφαίρα (περίπτωση συμμετρίας)

- Είναι ενδιαφέρον πως η ποσότητα Z_{rs} που ορίσαμε πιο πάνω σχετίζεται με την απόσταση Mahalanobis. Η απόσταση Mahalanobis είναι μια πολύ σημαντική ποσότητα στην πολυμεταβλητή στατιστική και μετράει το πόσο διαφέρουν δύο παρατηρήσεις. Ο ορισμός της για δυο παρατηρήσεις X_r και X_s είναι ο ακόλουθος $D_{rs}^2 = (x_r - x_s)' S^{-1} (x_r - x_s)$. Επομένως μπορεί να δει κανείς πως $Z_{rs} = 0.5\left(Z_{rr}^2 + Z_{ss}^2 - D_{rs}^2\right)$.

- Όπως θα δούμε αργότερα οι συντελεστές που ορίσαμε μπορούν να χρησιμοποιηθούν για να ελέγξουν αν τα δεδομένα προέρχονται από πολυμεταβλητή κανονική κατανομή

Παράδειγμα 2.2 (συνέχεια).

Ας χρησιμοποιήσουμε και πάλι τα δεδομένα για τις 20 πόλεις της Αμερικής. Χρησιμοποιώντας και τις 5 μεταβλητές προκύπτει πως ο συντελεστής ασυμμετρίας είναι 4.61778 και ο συντελεστής κύρτωσης 25.0283. επειδή έχουμε 5 μεταβλητές ο συντελεστής για την 5-διάστατη πολυμεταβλητή κανονική κατανομή είναι 35 και άρα τα δεδομένα έχουν μικρότερη κύρτωση από αυτήν την κατανομή.

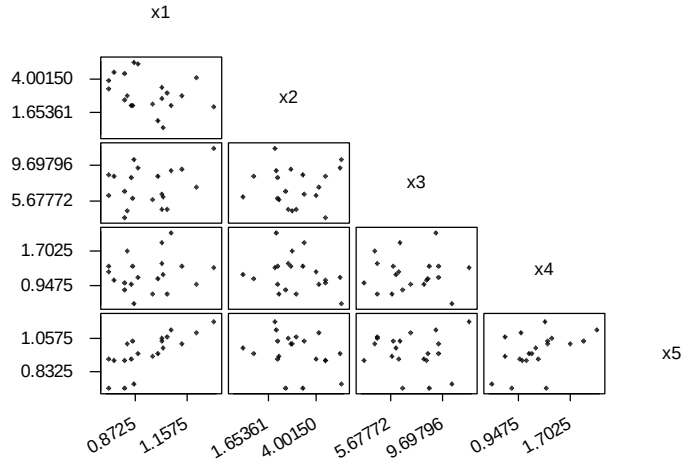
Για να αποκτήσουμε μια ιδέα των συντελεστών, στον πίνακα 2.6 που ακολουθεί έχουμε υπολογίσει τους συντελεστές ασυμμετρίας και κύρτωσης για τα ζευγάρια των μεταβλητών. Μπορείτε επίσης να δείτε και τον αντίστοιχο πίνακα με τα διαγράμματα σημείων (γράφημα 2.13) για να έχετε μια εικόνα του τι τελικά μετράνε αυτοί οι συντελεστές.

	Ασυμμετρία					Κύρτωση			
	X_1	X_2	X_3	X_4		X_1	X_2	X_3	X_4
X_2	0.415					5.619			
X_3	0.725	0.280				5.860	5.968		
X_4	0.529	0.742	0.356			5.814	5.933	6.499	
X_5	1.355	0.438	1.303	0.671		6.123	5.837	6.863	7.189

Πίνακας 2.6. Συντελεστές ασυμμετρίας και κύρτωσης για ζεύγη μεταβλητών

Ως προς την κύρτωση παρατηρούμε πως όλες οι τιμές είναι κοντά στο 6, ενώ η αντίστοιχη τιμή για την διμεταβλητή κανονική κατανομή είναι 8 (παρατηρούμε δηλαδή κάποια απόκλιση από τη διμεταβλητή κανονική κατανομή). Θυμηθείτε πως έχουμε μάλλον λίγες παρατηρήσεις, μόλις 20. Η μεγαλύτερη ασυμμετρία υπάρχει ανάμεσα στις μεταβλητές X_1 και X_5 κάτι που μπορεί κανείς να παρατηρήσει και στο γράφημα 2.13.

Ο πίνακας 2.7 περιέχει όλα τα στοιχεία κάτω από τη διαγώνιο του πίνακα **Z.** Για τον υπολογισμό της κύρτωσης χρειαζόμαστε μόνο τα στοιχεία της διαγωνίου, τα στοιχεία που δεν παρουσιάζονται μπορούν εύκολα να βρεθούν λόγω της συμμετρίας του πίνακα.



Γράφημα 2.13. Matrix plot για τα δεδομένα του Παραδείγματος 2.2.

2.3 Τρόποι Πολυμεταβλητής Ανάλυσης

Πριν τελειώσουμε αυτό το κεφάλαιο θα πρέπει να αναφέρουμε την εξής ιδέα. Ο πίνακας δεδομένων μας αποτελείται από γραμμές και στήλες όπου οι γραμμές αντιστοιχούν σε διαφορετικές παρατηρήσεις και οι στήλες σε διαφορετικές μεταβλητές. Σε πολλές εφαρμογές το ενδιαφέρον εστιάζεται στη μελέτη των μεταβλητών και των σχέσεων που υπάρχουν μεταξύ των μεταβλητών Επειδή ο πίνακας R παίζει μεγάλο ρόλο σε αυτή την προσέγγιση πολλές μέθοδοι ονομάζονται R-τεχνικές καθώς προσπαθούν να εξηγήσουν τις σχέσεις που υπάρχουν ανάμεσα στις μεταβλητές. Τέτοιες τεχνικές είναι η μέθοδος των κυρίων συνιστωσών, η παραγοντική ανάλυση και η ανάλυση κανονικών συσχετίσεων.

Πολλές όμως φορές το ενδιαφέρον εστιάζεται στο να βρεθούν σχέσεις ανάμεσα στις γραμμές του πίνακα και όχι ανάμεσα στις στήλες, σχέσεις δηλαδή ανάμεσα στις παρατηρήσεις και όχι ανάμεσα στις μεταβλητές. Τέτοιες τεχνικές είναι γνωστές ως Q –τεχνικές και τέτοιες τεχνικές είναι η διακριτική ανάλυση, η ανάλυση κατά συστάδες και η πολυδιάστατη κλιμακοποίησης.

Τέλος υπάρχουν μέθοδοι που ενδιαφέρονται και για τις γραμμές και τις στήλες συγχρόνως και προσπαθούν να αναλύσουν τα δεδομένα και ως προς τις γραμμές και ως προς τις στήλες την ίδια στιγμή. Τέτοια τεχνική είναι η ανάλυση αντιστοιχιών.

Αυτό που θα πρέπει κανείς να έχει υπόψη του είναι πως κάθε παρατήρηση είναι ένα διάνυσμα στο χώρο p-διαστάσεων που ορίζουν οι p μεταβλητές που έχουμε, αλλά συγχρόνως κάθε

μεταβλητή μπορεί να ειδωθεί σαν ένα διάνυσμα στο χώρο *n* διαστάσεων που ορίζουν οι παρατηρήσεις μας.

Ο λόγος που αυτό πρέπει να γίνει σαφές είναι πως πολλές φορές όταν έχουμε έναν πίνακα δεδομένων στα χέρια μας δεν είναι ξεκάθαρο ποιες είναι οι μεταβλητές και ποιες οι παρατηρήσεις και θα πρέπει ο ερευνητής να ξεκαθαρίσει τι ακριβώς θέλει να μελετήσει για να μπορέσει να προχωρήσει.

	Παρατήρηση																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	3.094																			
2	-0.276	5.158																		
3	1.169	2.278	4.566																	
4	-2.148	-1.101	-1.595	2.332																
5	1.789	-0.596	0.025	-0.025	4.233															
6	2.216	0.786	-0.514	-2.073	-0.201	4.031														
7	-0.922	2.631	0.859	-0.732	-1.969	0.021	2.748													
8	1.144	-0.130	-2.004	-1.314	0.056	2.295	0.811	3.560												
9	2.032	-0.324	1.613	-1.554	-0.579	2.416	-1.098	-0.647	4.037											
10	-1.254	-0.580	-2.942	0.299	-2.131	0.844	1.321	2.492	-1.402	3.664										
11	-2.169	3.228	1.682	1.784	0.938	-2.748	0.665	-3.034	-1.866	-2.207	6.222									
12	-0.623	1.462	1.872	-0.010	-1.675	-0.020	0.462	-2.002	1.704	-1.089	1.497	2.438								
13	3.081	-0.534	-0.157	-1.629	3.510	1.818	-1.270	2.053	0.299	-0.919	-1.708	-2.186	4.517							
14	-2.497	-0.096	-4.683	2.929	1.066	-0.431	-0.730	0.778	-2.981	2.053	2.243	-1.349	-0.355	7.909						
15	-1.800	0.745	-1.241	-0.002	-4.052	0.870	2.217	1.169	-0.095	3.176	-1.419	0.890	-2.916	0.331	4.466					
16	-0.507	-3.593	-2.858	1.904	0.020	0.518	-3.108	-0.993	1.742	0.224	-1.301	0.358	-0.794	2.779	-0.090	6.064				
17	-1.477	-1.300	0.924	0.068	-2.741	-2.583	1.844	-0.152	-1.490	1.433	-1.386	-0.132	-2.329	-3.360	2.154	-2.304	6.368			
18	0.200	-4.884	-0.438	0.436	-0.320	-2.011	-1.419	-0.221	-0.082	0.350	-3.408	-1.255	-0.111	-2.886	-0.356	1.213	4.417	6.648		
19	-0.322	-0.900	0.254	0.862	2.978	-3.595	-0.020	-0.004	-3.967	-0.836	1.645	-2.603	1.831	0.239	-2.930	-2.974	2.368	2.485	7.224	
20	-0.731	-1.972	1.191	1.568	-0.326	-1.638	-2.312	-3.858	2.243	-2.495	1.342	2.261	-2.200	-0.959	-1.116	3.698	-0.321	1.640	-1.736	5.721

Πίνακας 2.7. Ο πίνακας **Z** για τα δεδομένα των αμερικάνικων πόλεων (χρήση και των 5 μεταβλητών)

3 ΠΟΛΥΜΕΤΑΒΛΗΤΕΣ ΚΑΤΑΝΟΜΕΣ

3.1 Πολυμεταβλητές κατανομές

Στην πολυμεταβλητή στατιστική, το ενδιαφέρον δεν εστιάζεται στην κατανομή κάθε τυχαίας μεταβλητής X_i , =1,...,i p , ξεχωριστά αλλά στη μελέτη πολλών τυχαίων μεταβλητών X_1, X_2, ..., X_p, συγχρόνως. Δηλαδή το ενδιαφέρον πια εστιάζεται στη μελέτη της από κοινού κατανομής των τυχαίων μεταβλητών, για τις οποίες υποθέτουμε πως υπάρχει εξάρτηση μεταξύ τους και αυτή την εξάρτηση ουσιαστικά θέλουμε να μελετήσουμε.

Η θεμελιώδης έννοια της πολυμεταβλητής στατιστικής είναι η από κοινού συνάρτηση πυκνότητας πιθανότητας (joint probability density function) για τις συνεχείς τυχαίες μεταβλητές και της από κοινού συνάρτησης πιθανότητας (joint probability function) για την περίπτωση διακριτών τυχαίων μεταβλητών. Στη συνέχεια για λόγους ευκολίας θα αναφερόμαστε μόνο σε συνεχείς τυχαίες μεταβλητές. Τα αποτελέσματα ισχύουν και στην περίπτωση διακριτών τυχαίων μεταβλητών αν αντί για ολοκλήρωμα χρησιμοποιήσουμε άθροισμα.

Ας υπενθυμίσουμε λοιπόν μερικά στοιχεία από τη θεωρία κατανομών που αφορούν τη μελέτη πολλών μεταβλητών συγχρόνως.

Θα συμβολίζουμε την από κοινού συνάρτηση πυκνότητας πιθανότητας με

$f(x_1 x_2 .., x_p)$ κατά αναλογία της μονομεταβλητής περίπτωσης.

Ας δούμε τώρα μερικές άλλες χρήσιμες κατανομές που προκύπτουν από την από κοινού συνάρτηση πυκνότητας πιθανότητας.

- Η περιθώρια κατανομή μιας τυχαίας μεταβλητής (θα υποθέσουμε πως ενδιαφερόμαστε για τη μεταβλητή X_1 για λόγους ευκολίας, αλλά προφανώς θα μπορούσε να είναι κάθε μια από τις p μεταβλητές, το ίδιο θα κάνουμε και στη συνέχεια όπου οι δείκτες των τυχαίων μεταβλητών μπορεί να είναι οποιεσδήποτε μεταβλητές και όχι αναγκαστικά ο δείκτης που θα χρησιμοποιούμε)

$$f(x_1) = \int_{x_2}\int_{x_3}\ldots\int_{x_p} f(x_1, x_2, \ldots, x_p)dx_p \ldots dx_2$$

δηλαδή ολοκληρώνουμε ως προς όλες τις υπόλοιπες τυχαίες μεταβλητές.

- Η από κοινού περιθώρια κατανομή δύο ή περισσοτέρων τυχαίων μεταβλητών που έχει τη μορφή

$$f(x_1, x_2) = \int_{x_3}\int_{x_4}\ldots\int_{x_p} f(x_1, x_2, \ldots, x_p)dx_p \ldots dx_3$$

Η παραπάνω αντιστοιχεί στην από κοινού περιθώρια των X_1, X_2 και προκύπτει ολοκληρώνοντας ως προς τις υπόλοιπες μεταβλητές.

- Γενικά για να βρούμε μια από κοινού περιθώρια κατανομή ολοκληρώνουμε ως προς όλες τις μεταβλητές που θέλουμε να μην εμφανίζονται μέσα στην από κοινού συνάρτηση πυκνότητας πιθανότητας. Έτσι στη γενική περίπτωση η από κοινού περιθώρια των τμ $X_1, X_2, \ldots, X_m$ ($m<p$) θα προκύπτει ως

$$f(x_1, \ldots, x_m) = \int_{x_{m+1}}\int_{x_{m+2}}\ldots\int_{x_p} f(x_1, x_2, \ldots, x_p)dx_p \ldots dx_{m+1}$$

- Η δεσμευμένη κατανομή διανύσματος τυχαίων μεταβλητών δοθέντων των τιμών άλλων μεταβλητών. Για παράδειγμα η δεσμευμένη από κοινού κατανομή των τυχαίων μεταβλητών X_1, X_2 δοθέντων των τιμών των τυχαίων μεταβλητών X_3, …, X_p ορίζεται ως

$$f(x_1, x_2 | x_3, \ldots, x_p) = \frac{f(x_1, x_2, \ldots x_p)}{f(x_3, \ldots, x_p)}$$

δηλαδή στον αριθμητή έχουμε την από κοινού κατανομή όλων των τυχαίων μεταβλητών και στον παρονομαστή την από κοινού κατανομή των τυχαίων μεταβλητών για τις οποίες έχουμε δεσμεύσει. Εμπειρικά μπορεί να πει κανείς πως στον παρονομαστή έχουμε την πληροφορία που υπάρχει.

Όπως και στη μονομεταβλητή περίπτωση μπορούμε να ορίσουμε αναμενόμενες τιμές για οποιαδήποτε συνάρτηση των τυχαίων μεταβλητών. Συγκεκριμένα

$$E\left[g\left(X_1, X_2, \ldots, X_p\right)\right] = \int_{x_1}\ldots\int_{x_p} g\left(x_1, x_2, \ldots, x_p\right) f\left(x_1, x_2, \ldots, x_p\right) dx_p \ldots dx_1$$

Μερικές ενδιαφέρουσες ειδικές περιπτώσεις είναι οι εξής

- Έστω πως $g\left(X_1, X_2, ..., X_p\right) = X_1$. Έχουμε πως

$$E\left(X_1\right) = \int_{x_1} ... \int_{x_p} x_1 f\left(x_1, x_2, ..., x_p\right) dx_p ... dx_1 =$$

$$\int_{x_1} x_1 \left[\int_{x_2} ... \int_{x_p} f\left(x_1, x_2, ..., x_p\right) dx_p ... dx_2 \right] dx_1 = \int_{x_1} x_1 f\left(x_1\right) dx_1$$

δηλαδή η αναμενόμενη τιμή της περιθώριας κατανομής του X_1.

- Έστω $g\left(X_1, X_2\right) = \left(X_1 - E(X_1)\right)\left(X_2 - E(X_2)\right)$. Η συνάρτηση αυτή οδηγεί στον ορισμό της συνδιακύμανσης δύο τυχαίων μεταβλητών

$$Cov(X_1, X_2) = E\left[\left(X_1 - E(X_1)\right)\left(X_2 - E(X_2)\right)\right] = E(X_1 X_2) - E(X_1)E(X_2)$$

Μερικές βασικές ιδιότητες της συνδιακύμανσης είναι οι εξής

- Αν οι τ.μ. X_1, X_2 είναι ανεξάρτητες, $Cov(X_1, X_2) = 0$
- Αν $Cov(X_1, X_2) = 0$, δεν ισχύει απαραίτητα ότι οι τ.μ. X_1, X_2 είναι ανεξάρτητες
- $Cov(X_1 + X_3, X_2) = Cov(X_1, X_2) + Cov(X_3, X_2)$
- $Cov(\alpha X_1 + \beta, \gamma X_2 + \delta) = \alpha\gamma Cov(X_1, X_2)$

Μπορεί κανείς εύκολα να δει πως η συνδιακύμανση μιας τυχαίας μεταβλητής με τον εαυτό της είναι η γνωστή μας διακύμανση. Η συνδιακύμανση μπορεί να πάρει τιμές και στον αρνητικό άξονα.

Ορισμός: Τυχαίο διάνυσμα $\mathbf{x}' = \left(X_1, X_2, ..., X_p\right)$ είναι το διάνυσμα όλα τα στοιχεία του οποίου είναι τ.μ.

Έστω $\mathbf{x}_{p\times 1}$ τυχαίο διάνυσμα

$$\mathbf{x} = \begin{bmatrix} X_1 \\ ... \\ X_p \end{bmatrix}$$

τότε η αναμενόμενη τιμή τυχαίου διανύσματος ορίζεται ως

$$E(\mathbf{x}) = \mu = \begin{bmatrix} E(X_1) \\ ... \\ E(X_p) \end{bmatrix}$$

Επίσης θα συμβολίζουμε με $Cov(\mathbf{x})$ τον πίνακα διακυμάνσεων του τυχαίου διανύσματος **x**, δηλαδή

$$Cov(\mathbf{x}) = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & ... & Cov(X_1, X_p) \\ Cov(X_1, X_2) & Var(X_2) & ... & ... \\ ... & ... & ... & ... \\ Cov(X_1, X_p) & Cov(X_2, X_p) & ... & Var(X_p) \end{bmatrix}$$

Ο πίνακας $Cov(\mathbf{x})$ είναι συμμετρικός.

Έστω, τώρα ένας πίνακας $\mathbf{C}_{m\times p}$ ο οποίος περιλαμβάνει γνωστές σταθερές και όχι τυχαίες μεταβλητές. Τότε ισχύει πως :

Το διάνυσμα $Y = \mathbf{C}\mathbf{x}$ διαστάσεων $m\times 1$ είναι ένα τυχαίο διάνυσμα και

$$E(\mathbf{C}\mathbf{x}) = \mathbf{C}E(\mathbf{x})$$

$$Cov(Y) = Cov(\mathbf{C}\mathbf{x}) = \mathbf{C}Cov(\mathbf{x})\mathbf{C}' = \mathbf{C}\Sigma_x\mathbf{C}'.$$

Παρατηρείστε πως επειδή η αναμενόμενη τιμή είναι γραμμικός τελεστής θα ισχύει πως αν $Y = \mathbf{C}\mathbf{x} + b$, όπου *b* κατάλληλο διάνυσμα, τότε ισχύει

$$E(Y) = \mathbf{C}E(\mathbf{x}) + b \qquad \text{και}$$

$$Cov(Y) = \mathbf{C}\Sigma_x\mathbf{C}'.$$

Ορισμός: Έστω τ.δ. $\underset{q\times 1}{\mathbf{x}_1}$, $\underset{m\times 1}{\mathbf{x}_2}$. Θα λέμε ότι τα διανύσματα αυτά είναι ανεξάρτητα όταν όλες οι μεταβλητές του πρώτου διανύσματος είναι *ανεξάρτητες* από όλες τις μεταβλητές του δεύτερου διανύσματος.

Αντίστοιχα με τον ορισμό του τυχαίου διανύσματος έχουμε

Ορισμός: Ο πίνακας **X** διαστάσεων $m\times p$ του οποίου όλα τα στοιχεία είναι τ.μ. λέγεται τυχαίος πίνακας.

Δηλαδή ο πίνακας

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ X_{m1} & X_{m2} & \dots & X_{mp} \end{bmatrix},$$

έχει διαστάσεις $m \times p$ και όλα τα στοιχεία του X_{ij} είναι τυχαίες μεταβλητές. Και πάλι η αναμενόμενη τιμή τυχαίου πίνακα είναι ο πίνακας με τις αναμενόμενες τιμές όλων των τυχαίων μεταβλητών που αποτελούν τον τυχαίο πίνακα, δηλαδή

$$E(\mathbf{X}) = \begin{bmatrix} E(X_{11}) & E(X_{12}) & \dots & E(X_{1p}) \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ E(X_{m1}) & E(X_{m2}) & \dots & E(X_{mp}) \end{bmatrix}$$

Ισχύουν οι γενικές ιδιότητες

$$E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y}),$$

όπου $\mathbf{X}$ και $\mathbf{Y}$ πίνακες ίδιων διαστάσεων.

Επίσης αν $\mathbf{A}$, $\mathbf{B}$ κατάλληλοι πίνακες (όχι τυχαίοι πίνακες) τότε ισχύει

$$E(\mathbf{AXB}) = \mathbf{A}E(\mathbf{X})\mathbf{B}.$$

Ο πίνακας διακύμανσης ενός τυχαίου πίνακα στην ουσία αποτελείται από τις συνδιακυμάνσεις όλων των ζευγών τυχαίων μεταβλητών, αλλά είναι αρκετά δύσχρηστος από στατιστική άποψη και για αυτό δεν θα αναφερθούμε περισσότερο σε αυτόν.

Παρατηρείστε πως

- Ένα τυχαίο διάνυσμα είναι στην ουσία ειδική περίπτωση ενός τυχαίου πίνακα, όταν p=1.
- Μια τυχαία μεταβλητή είναι στην ουσία ειδική περίπτωση ενός τυχαίου πίνακα, όταν p=m=1.
- Όταν μιλάμε για την συνάρτηση πυκνότητας ενός διανύσματος μιλάμε για την από κοινού συνάρτηση πυκνότητας των τυχαίων μεταβλητών που το απαρτίζουν.

Παράδειγμα 3.1.

Έστω τ.μ. X_1, X_2. Ορίζουμε τις τ.μ. $Z_1=X_1-X_2$, $Z_2=X_1+X_2$

και άρα το τυχαίο διάνυσμα $\mathbf{Z}=\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$. Να βρεθούν τα Ε(**Z**), *Cov* (**Z**).

Είναι $\mathbf{Z}=\mathbf{CX}$, όπου $\mathbf{C}=\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ και $\mathbf{X}=\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. Οι διαστάσεις του πίνακα **C** είναι τέτοιες ώστε από το αρχικό διάνυσμα μεγέθους 2×1 να μας οδηγήσουν και πάλι σε ένα διάνυσμα 2×1 (αφού έχουμε δύο Z_i). Επομένως οι διαστάσεις του **C** είναι 2×2.

Έτσι

$$E(\mathbf{Z})=\mathbf{C}E(\mathbf{X})=\begin{bmatrix} E(X_1)-E(X_2) \\ E(X_1)+E(X_2) \end{bmatrix} \quad \text{και}$$

$$Cov(Z)=\mathbf{C}\Sigma_X\mathbf{C}'=\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}=$$

$$\begin{bmatrix} \sigma_1^2+\sigma_2^2-2\sigma_{12} & \sigma_1^2-\sigma_2^2 \\ \sigma_1^2-\sigma_2^2 & \sigma_1^2+\sigma_2^2+2\sigma_{12} \end{bmatrix}$$

Τα διαγώνια στοιχεία του πίνακα αποτελούν τα γνωστά μας αποτελέσματα για το άθροισμα και τη διαφορά συσχετισμένων τυχαίων μεταβλητών. Προφανώς ο πίνακας που προκύπτει είναι συμμετρικός.

Παράδειγμα 3.2.

Έστω $\mathbf{Z}=\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix}$, όπου $Z_3=3X_1+4X_2$, ενώ τα Z_1, Z_2 έχουν οριστεί προηγουμένως. Τώρα επειδή το διάνυσμα **Z** έχει διαστάσεις 3×1 ο πίνακας **C** θα πρέπει να έχει διαστάσεις 3×2. Προκύπτει εύκολα ότι $\mathbf{C}=\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 3 & 4 \end{bmatrix}$ και άρα

$$E(\mathbf{Z})=\mathbf{C}E(\mathbf{X})=\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix}=\begin{bmatrix} E(X_1)-E(X_2) \\ E(X_1)+E(X_2) \\ 3E(X_1)+4E(X_2) \end{bmatrix}$$

Για τον πίνακα διακύμανσης έχουμε πως

$$Cov(Z) = \mathbf{C}\Sigma_X\mathbf{C}' = \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 3 \\ -1 & 1 & 4 \end{bmatrix} =$$

$$\begin{bmatrix} \sigma_1^2 - \sigma_{12} & \sigma_{12} - \sigma_2^2 \\ \sigma_1^2 + \sigma_{12} & \sigma_{12} + \sigma_2^2 \\ 3\sigma_1^2 + 4\sigma_{12} & 3\sigma_{12} + 4\sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 3 \\ -1 & 1 & 4 \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} & \sigma_1^2 - \sigma_2^2 & 3\sigma_1^2 - 4\sigma_2^2 + \sigma_{12} \\ \sigma_1^2 - \sigma_2^2 & \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} & 3\sigma_1^2 + 4\sigma_2^2 + 7\sigma_{12} \\ 3\sigma_1^2 - 4\sigma_2^2 + \sigma_{12} & 3\sigma_1^2 + 4\sigma_2^2 + 7\sigma_{12} & 9\sigma_1^2 + 16\sigma_2^2 + 24\sigma_{12} \end{bmatrix}$$

Παράδειγμα 3.3.

Έστω τ.μ. X_1, X_2, X_3, $\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$, $Z_1=X_1-X_3$, $Z_2=X_2$

Τότε $\mathbf{C} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$, και άρα

$$E(\mathbf{Z}) = \mathbf{C}E(\mathbf{X}) = \begin{bmatrix} E(X_1) - E(X_3) \\ E(X_2) \end{bmatrix}$$

$$Cov(Z) = \mathbf{C}\Sigma_X\mathbf{C}' = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} =$$

$$\begin{bmatrix} \sigma_1^2 - \sigma_{13} & \sigma_{12} - \sigma_{23} & \sigma_{13} - \sigma_3^2 \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma_1^2 + \sigma_3^2 - 2\sigma_{13} & \sigma_{12} - \sigma_{23} \\ \sigma_{12} - \sigma_{23} & \sigma_2^2 \end{bmatrix}$$

Παρατήρηση: Οι τύποι που δόθηκαν παραπάνω αφορούν οποιοδήποτε τυχαίο διάνυσμα χωρίς καμιά υπόθεση για την από κοινού κατανομή που ακολουθεί το διάνυσμα. Συνεπώς ισχύουν γενικά και σε κάθε περίπτωση.

3.2 Πολυμεταβλητή Κανονική Κατανομή

Αν η απλή κανονική κατανομή αποτελεί τη βάση για τις περισσότερες απλές στατιστικές εφαρμογές (πχ ελέγχους υποθέσεων, γραμμική παλινδρόμηση κλπ), ανάλογη είναι και η χρήση της πολυμεταβλητής κανονικής κατανομής στην πολυμεταβλητή στατιστική. Στην πράξη, οι περισσότερες μέθοδοι αναπτύχθηκαν με βάση την κατανομή αυτή και άρα αποτελεί ακρογωνιαίο λίθο για τις περισσότερες μεθόδους.

Όπως είναι γνωστό η συνάρτηση πυκνότητας πιθανότητας για την μονοδιάστατη κανονική κατανομή είναι η

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right),\ -\infty < x, \mu < +\infty,\ \sigma > 0.$$

Η πολυμεταβλητή κανονική κατανομή που ορίζεται για το τυχαίο διάνυσμα x διαστάσεων p × 1 έχει από κοινού συνάρτηση πυκνότητας πιθανότητας

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

όπου $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$, $\boldsymbol{\mu}$ είναι ένα διάνυσμα p×1 και $\mathbf{\Sigma}$ ένας πίνακας p × p .

Είναι ιδιαίτερα ενδιαφέρον ότι η ερμηνεία των παραμέτρων είναι απλή. Το διάνυσμα **μ** περιέχει τις αναμενόμενες τιμές κάθε μιας μεταβλητής, είναι δηλαδή το διάνυσμα των μέσων, ενώ ο πίνακας **Σ** είναι ο πίνακας με τις συνδιακυμάνσεις των μεταβλητών του τυχαίου διανύσματος, δηλαδή

$\boldsymbol{\mu}$=E(**x**) και

$\mathbf{\Sigma}$=Cov(**x**)

Γενικά θα συμβολίζουμε **x**~N_p (**μ**, **Σ**) και θα εννοούμε πως το τυχαίο διάνυσμα **x** ακολουθεί την *p*-διάστατη κανονική κατανομή με διάνυσμα μέσων **μ** και πίνακα διακυμάνσεων **Σ.**

Δεδομένου πως ο πίνακας **Σ** είναι πίνακας διακύμανσης, προκύπτει πως είναι θετικά ημιορισμένος, συμμετρικός και τα διαγώνια στοιχεία του δεν μπορεί να είναι αρνητικά. Επίσης η πολυμεταβλητή κανονική κατανομή έχει *p(p+3)/2* άγνωστες παραμέτρους, εκ των οποίων *p* είναι οι αναμενόμενες τιμές των τυχαίων μεταβλητών και *p(p+1)/2* είναι τα διαφορετικά στοιχεία του πίνακα διακύμανσης.

Θα πρέπει να σημειωθεί πως γενικά στη στατιστική θεωρούμε ως πολυμεταβλητό αντίστοιχο μιας κατανομής, μια πολυμεταβλητή κατανομή που έχει περιθώριες κατανομές στην οικογένεια κατανομών που θέλουμε. Για παράδειγμα όταν μιλάμε για πολυμεταβλητή κατανομή Poisson εννοούμε μια κατανομή με περιθώριες κατανομές Poisson. Αυτό όμως πολλές φορές οδηγεί σε παρεξηγήσεις καθώς υπάρχουν περισσότερες από μια πολυμεταβλητές κατανομές για τις οποίες καταχρηστικά έχουμε χρησιμοποιήσει το ίδιο όνομα απλά και μόνο επειδή οι περιθώριες κατανομές τις ανήκουν σε κάποια συγκεκριμένη οικογένεια κατανομών. Για αυτό χρειάζεται προσοχή στην ονοματολογία και χρήση πολυμεταβλητών κατανομών. Θα πρέπει βέβαια να σημειωθεί πως για την πολυμεταβλητή κανονική κατανομή, κυρίως λόγω της γενικευμένης χρήσης της, δεν υπάρχει τέτοιο πρόβλημα.

Ας δούμε την ειδική περίπτωση ανεξάρτητων μεταβλητών. Αν ο πίνακας $\boldsymbol{\Sigma}$ είναι διαγώνιος, δηλαδή

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & \sigma_p^2 \end{bmatrix},$$

τότε $|\Sigma| = \prod_{i=1}^{p} \sigma_i^2$ και $\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1/\sigma_p^2 \end{bmatrix}$. Επομένως η πολυμεταβλητή κανονική κατανομή παίρνει τη μορφή

$$f(\mathbf{x}) = \prod_{i=1}^{p} \frac{1}{(2\pi)^{1/2}\sigma_i} \exp\left(-\frac{1}{2} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right).$$

Εύκολα αναγνωρίζει κανείς πως είναι το γινόμενο p μονοδιάστατων συναρτήσεων πυκνότητας από την κανονική κατανομή, κι επομένως οι μεταβλητές X_i είναι ανεξάρτητες.

Η από κοινού αθροιστική συνάρτηση κατανομής μιας πολυμεταβλητής κατανομής ορίζεται κατά αντιστοιχία της μονομεταβλητής περίπτωσης ως:

$$F(x_1, x_2, \ldots, x_p) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_p \le x_p) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_p} f(y_1, y_2, \ldots, y_p) dy_p \ldots dy_2 dy_1$$

Για την περίπτωση της πολυμεταβλητής κανονικής κατανομής δεν είναι καθόλου απλό να υπολογιστούν τα πολλαπλά ολοκληρώματα και χρειάζονται ειδικές αριθμητικές μέθοδοι. Αυτό κάνει την από κοινού αθροιστική κατανομή εξαιρετικά δύσχρηστη στην πολυμεταβλητή ανάλυση και η χρησιμότητά της είναι πολύ μικρή.

Στη συνέχεια θα δούμε μια ειδική περίπτωση, τη διδιάστατη κανονική κατανομή.

3.3 Διδιάστατη κανονική κατανομή

Έστω πως έχουμε μόνο 2 τυχαίες μεταβλητές Χ και Υ. Τότε η από κοινού τους κατανομή έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

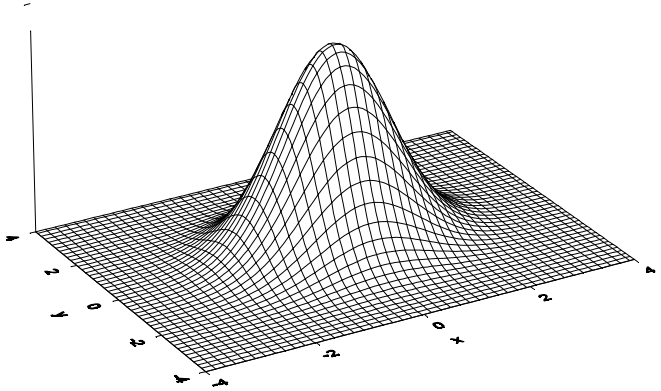
όπου $\rho=\frac{\sigma_{xy}}{\sigma_x\sigma_y}$ είναι η συσχέτιση των τ.μ. Χ και Υ.

Το διάνυσμα των μέσων είναι το $\mu=\begin{bmatrix}\mu_X\\ \mu_Y\end{bmatrix}$ και ο πίνακας διακύμανσης $\mathbf{\Sigma}=\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{bmatrix}$.

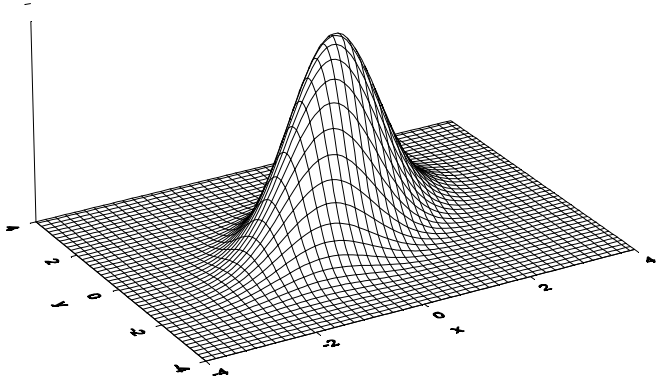
Παρατηρείστε πως αν ρ=0 τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας της διμεταβλητής κατανομής είναι απλά το γινόμενο δύο συναρτήσεων πιθανοτήτων απλών κανονικών κατανομών, δηλαδή οι δύο μεταβλητές είναι ανεξάρτητες.

Για τη διδιάστατη κατανομή είναι εύκολο να φτιάξουμε γραφήματα της από κοινού συνάρτησης πυκνότητας. Αυτά μπορούν να μας δώσουν ενδιαφέρουσες ερμηνείες για τις παραμέτρους. Στη μονοδιάστατη περίπτωση έχουμε συνδέσει τη μορφή της κανονικής κατανομής με μια καμπάνα, δηλαδή μια συμμετρική κατανομή. Το ίδιο ισχύει και στη διδιάστατη περίπτωση, όπου τώρα έχουμε μια τρισδιάστατη εικόνα. Σε όλα τα γραφήματα 3.1-3.4 βλέπουμε πως η μορφή της κατανομής είναι συμμετρική και μοιάζει με μια καμπάνα της οποίας τα χαρακτηριστικά καθορίζονται από τις παραμέτρους. Η κορυφή της καμπάνας βρίσκεται πάντα

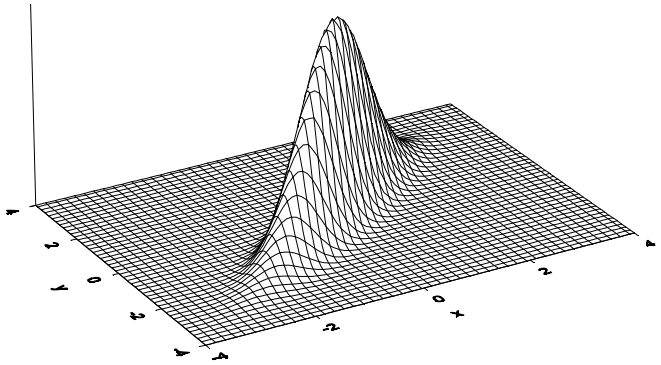
στο σημείο **μ,** δηλαδή ακριβώς στο σημείο που υποδεικνύει το διάνυσμα των μέσων.



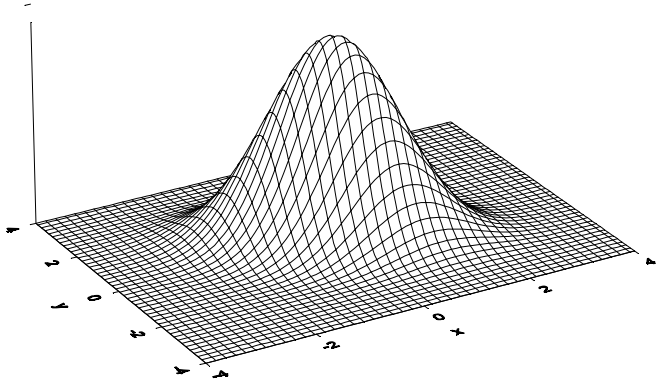
ρ=0



ρ=0.5

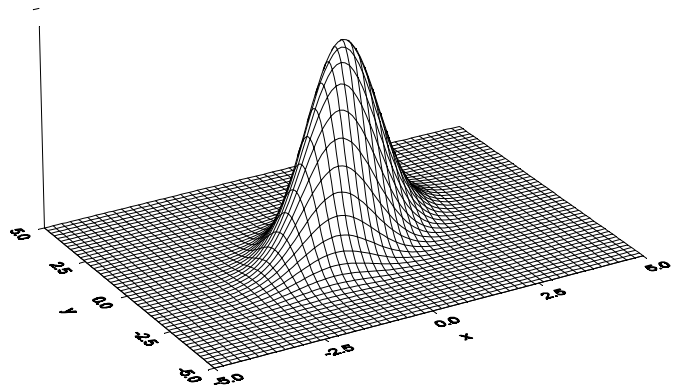


ρ=0.9

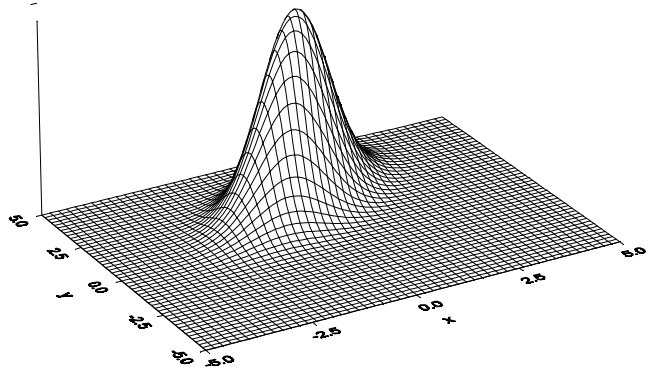


ρ=- 0.5

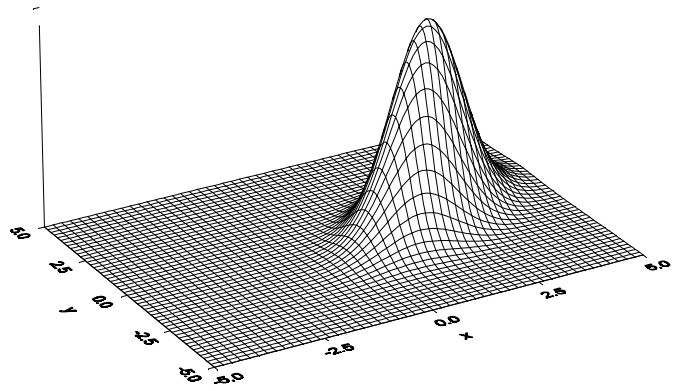
Γράφημα 3.1 Διδιάστατες κανονικές κατανομές με ίδιο διάνυσμα μέσων αλλά διαφορετική συσχέτιση ρ



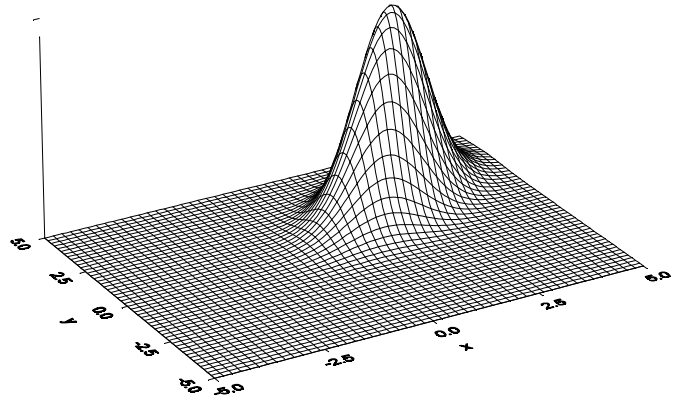
m=[0 0]



m=[0 2]

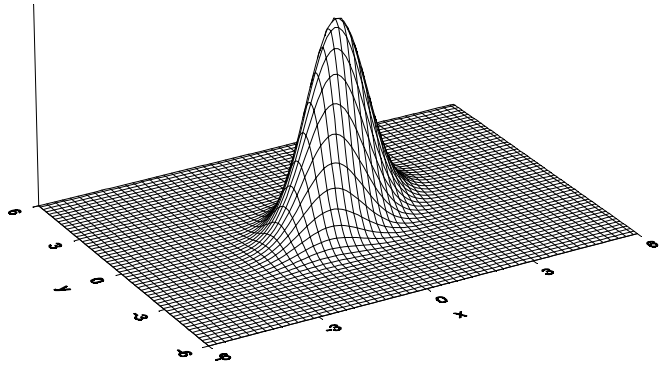


m=[2 0]

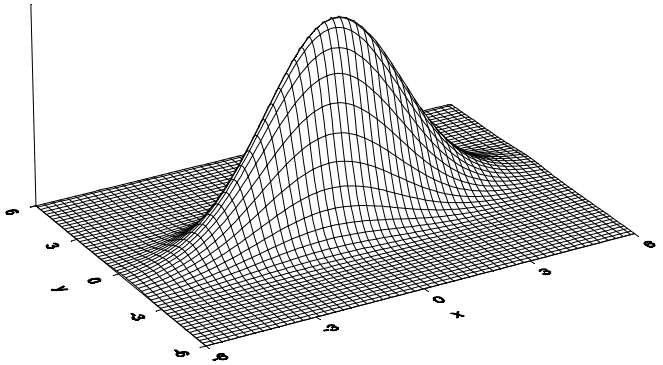


m=[2 2]

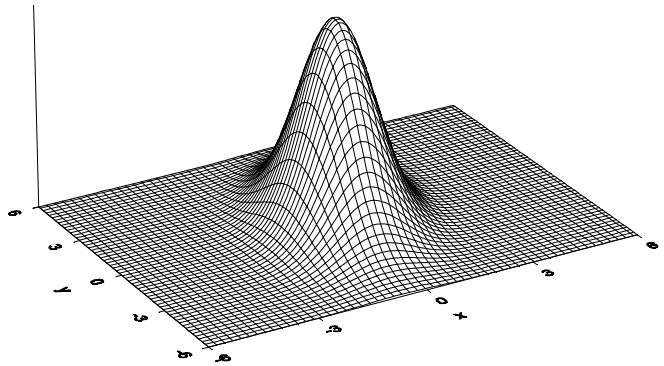
Γράφημα 3.2 Διδιάστατες κανονικές κατανομές με ίδιο πίνακα διακύμανσης αλλά διαφορετικά διανύσματα μέσων



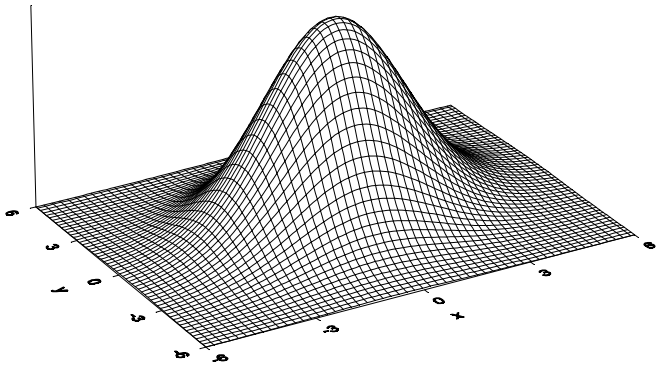
var=[1 0.5 / 0.5 1]



var=[4 0.5 / 0.5 1]

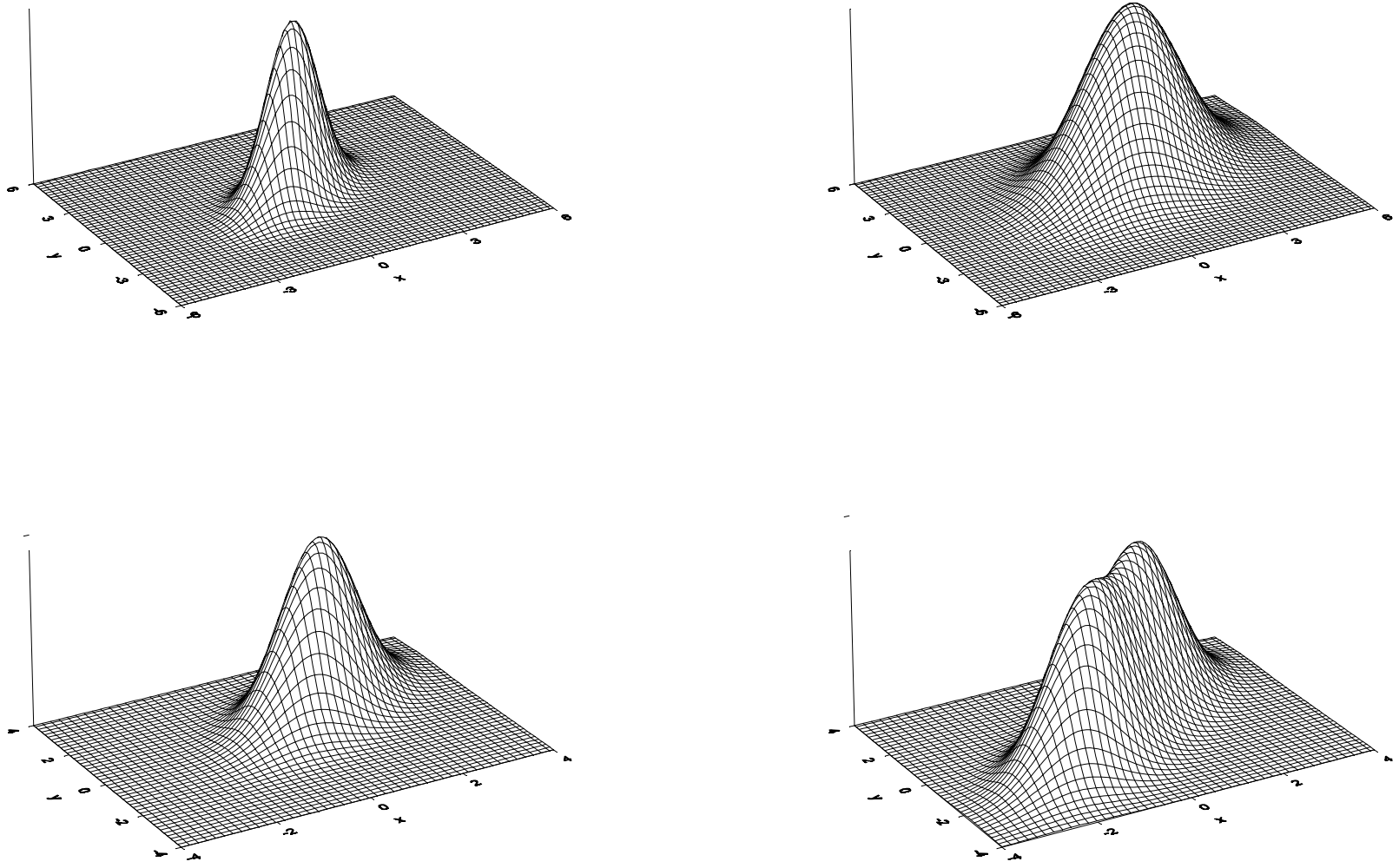


var=[1 0.5 / 0.5 4]



var=[4 0.5 / 0.5 4]

Γράφημα 3.3 Διδιάστατες κανονικές κατανομές με ίδιο μηδενικό διάνυσμα μέσων αλλά διαφορετικούς πίνακες διακύμανσης



Γράφημα 3.4. Διάφορες συνεχείς διμεταβλητές κατανομές

Έτσι

- Από το γράφημα 3.1 βλέπουμε την ερμηνεία του συντελεστή συσχέτισης *ρ*. Το διάνυσμα των μέσων είναι το ίδιο για όλα τα γραφήματα και οι διακυμάνσεις είναι επίσης ίσες με 1 και για τις δύο διαστάσεις. Αυτό που βλέπουμε είναι πως το *ρ* καθορίζει αφενός τον προσανατολισμό της από κοινού συνάρτησης πυκνότητας πιθανότητας, δηλαδή για θετικές τιμές η κατανομή τοποθετείται ως προς τη διαγώνιο από κάτω αριστερά προς το πάνω δεξιά σημείο, ενώ για αρνητικές τιμές από πάνω αριστερά προς κάτω δεξιά. Αφετέρου η τιμή του συντελεστή καθορίζει πόσο μεγάλη βάση θα έχει η κατανομή ως προς τη διαγώνιο. Έτσι για τιμές του *ρ=0.9* παρατηρείστε πως η κατανομή είναι πολύ συγκεντρωμένη πλησίον της διαγωνίου, ενώ για *ρ*=0 αυτό δεν συμβαίνει.

- Όπως βλέπουμε στο γράφημα 3.2 το διάνυσμα των μέσων καθορίζει το σημείο κορυφής δηλαδή που θα τοποθετήσουμε την καμπάνα. Στο γράφημα έχουμε χρησιμοποιήσει τον ίδιο πίνακα διακύμανσης και συγκεκριμένα $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, ενώ το διάνυσμα των μέσων αλλάζει. Ανάλογα με τις τιμές μετακινούμαστε πάνω στους άξονες οι οποίοι είναι οι ίδιοι για όλα τα γραφήματα.

- Από το γράφημα 3.3 μπορούμε να δούμε την ερμηνεία των διακυμάνσεων. Αυτές λοιπόν καθορίζουν πόση έκταση ως προς τον αντίστοιχο άξονα θα έχουν οι «καμπάνες». Αν δηλαδή είναι ίσες οι διακυμάνσεις (τα διαγώνια στοιχεία του $\mathbf{\Sigma}$) τότε η κατανομή θα απλώνεται το ίδιο και προς τους δύο άξονες.

- Τέλος, στο γράφημα 3.4 μπορεί κανείς να δει άλλες διμεταβλητές κατανομές, μη κανονικές. Παρατηρείστε την ποικιλία σχημάτων που μπορούν να σχηματιστούν, δηλαδή κατανομές με πολύ πιο παχιά βάση από ότι η κανονική, ή ακόμα και με δύο κορυφές. Όλες οι κατανομές που εμφανίζονται στο γράφημα είναι μείγματα διμεταβλητών κανονικών κατανομών. Η ποικιλία σχημάτων που μπορεί κανείς να σχηματίσει είναι ουσιαστικά άπειρη.

Ένα ακόμα ενδιαφέρον αποτέλεσμα είναι το εξής:

Για τη διμεταβλητή κανονική κατανομή ισχύει πως αν οι τυχαίες μεταβλητές X και Y είναι ασυσχέτιστες τότε είναι και ανεξάρτητες, ισχύει και το αντίστροφο.

Θυμηθείτε πως γενικά το αντίστροφο δεν ισχύει.

3.4 Ιδιότητες της πολυμεταβλητής κανονικής κατανομής

Ας δούμε τώρα μερικές χρήσιμες ιδιότητες της πολυμεταβλητής κανονικής κατανομής.

Γραμμικοί Μετασχηματισμοί

Θεώρημα Έστω $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και έστω $\boldsymbol{\alpha}$ ένα διάνυσμα διαστάσεων $(p \times 1)$ από σταθερούς αριθμούς (όχι τυχαίες μεταβλητές). Τότε για την τυχαία μεταβλητή $y = \boldsymbol{\alpha}'\mathbf{x}$ ισχύει πως : $y \sim N(\boldsymbol{\alpha}'\boldsymbol{\mu}, \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha})$

Η απόδειξη του θεωρήματος δεν θα δοθεί αλλά απλά πρόκειται για τη χρήση της γνωστής μεθόδου μετασχηματισμού τυχαίων μεταβλητών. Αυτό που έχει περισσότερη σημασία είναι η ερμηνεία του παραπάνω αποτελέσματος. Το θεώρημα λοιπόν μας επιτρέπει να δούμε πως η κατανομή γραμμικού συνδυασμού μεταβλητών με από κοινού κατανομή την πολυδιάστατη κανονική ακολουθεί την μονοδιάστατη κανονική κατανομή. Είναι γνωστό πως το αποτέλεσμα ισχύει για την περίπτωση ανεξάρτητων κανονικών τυχαίων μεταβλητών, τώρα παρατηρούμε πως το ίδιο ισχύει και για την περίπτωση που οι τμ έχουν συσχέτιση μεταξύ τους.

Εφαρμογή: Κατανομή Δειγματικού Μέσου για Συσχετισμένες Παρατηρήσεις

Μια ενδιαφέρουσα εφαρμογή των ιδιοτήτων της πολυμεταβλητής κανονικής κατανομής είναι να εξετάσουμε την κατανομή της μέσης τιμής από ένα δείγμα που οι τιμές του δεν είναι ανεξάρτητες. Συγκεκριμένα, έστω $X_1, \ldots, X_n$ συσχετισμένες παρατηρήσεις από κανονικές κατανομές και έστω πως η από κοινού τους κατανομή είναι πολυμεταβλητή κανονική. Δηλαδή υποθέτουμε πως το διάνυσμα $\mathbf{x}$ που περιέχει όλες τις τυχαίες μεταβλητές ακολουθεί την πολυμεταβλητή κανονική κατανομή. Έχουμε, δηλαδή πως

$$\mathbf{x} = \begin{bmatrix} X_1 \\ \ldots \\ X_n \end{bmatrix} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{, όπου } \underset{n \times n}{\boldsymbol{\Sigma}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \ldots & \sigma_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{1n} & \sigma_{2n} & \ldots & \sigma_{nn} \end{bmatrix} \text{ και } \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \ldots \\ \mu_n \end{bmatrix}.$$

Μπορεί κάποιος εύκολα να δει πως η μέση τιμή προκύπτει ως ένας γραμμικός συνδυασμός των τυχαίων μας μεταβλητών. Συγκεκριμένα, ορίζουμε το διάνυσμα $\boldsymbol{\alpha}_{n\times 1}: \boldsymbol{\alpha}' = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix}$

Τότε είναι $y = \boldsymbol{\alpha}'\mathbf{x} \sim N(\boldsymbol{\alpha}'\boldsymbol{\mu}, \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha})$ και συνεπώς $\bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N(\boldsymbol{\alpha}'\boldsymbol{\mu}, \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha})$. Κάνοντας τις πράξεις βρίσκουμε πως

$$\boldsymbol{\alpha}'\boldsymbol{\mu} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{bmatrix} = \frac{\sum_{i=1}^{n} \mu_i}{n} \quad \text{και}$$

$$\boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & .. & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2n} \\ \dots & \dots & \dots & \dots \\ \sigma_{1n} & \sigma_{2n} & \dots & \sigma_{nn} \end{bmatrix} \frac{1}{n} \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} =$$

$$= \frac{1}{n^2}\begin{bmatrix} \sum_{i=1}^{n}\sigma_{i1} & \sum_{i=1}^{n}\sigma_{i2} & \dots & \sum_{i=1}^{n}\sigma_{in} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}}{n^2}$$

Επομένως $\bar{\mathbf{x}} \sim N\left(\frac{\sum_{i=1}^{n}\mu_i}{n}, \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}}{n^2}\right)$. Δηλαδή η κατανομή του μέσου είναι κανονική αλλά τώρα έχει αλλάξει η διακύμανση.

Ας δούμε κάποιες πιο ειδικές περιπτώσεις.

Στην περίπτωση ανεξάρτητων και ισόνομων τ.μ. δηλαδή στην περίπτωση που κάθε $X_i \sim N(\mu, \sigma^2)$, i=1,...,n έχουμε πως $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$ και άρα $\bar{\mathbf{x}} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, το γνωστό δηλαδή αποτέλεσμα για ανεξάρτητα δείγματα.

Για λόγους ευκολίας και χωρίς να χάνουμε τίποτα στη γενικότητα των αποτελεσμάτων ας υποθέσουμε πως $\sigma^2=1$. Έστω τώρα πως ξέρουμε πως κάθε μεταβλητή έχει συσχέτιση με κάθε άλλη ίση με ρ. (Παρατήρηση: αν $\rho>0$, τότε ο πίνακας $\mathbf{\Sigma}$ που προκύπτει είναι πίνακας διακύμανσης, αυτό όμως δεν ισχύει πάντα αν $\rho<0$).

Δηλαδή υποθέτουμε πως $\mathbf{\Sigma}=\begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{bmatrix}$. Χρησιμοποιώντας το γενικό αποτέλεσμα βρίσκουμε πως $\bar{\mathbf{x}} \sim N\left(\mu, \frac{n+n(n-1)\rho}{n^2}\right)$.

Αν αναλογιστούμε πως η περίπτωση ανεξάρτητου δείγματος αντιστοιχεί στην τιμή $\rho=0$, προκύπτει πολύ εύκολα πως αν $\rho>0$ η διακύμανση της μέσης τιμής από συσχετισμένα δεδομένα θα είναι μεγαλύτερη από αυτήν όταν τα δεδομένα ήταν ασυσχέτιστα. Αυτό είναι λογικό αν αναλογιστούμε πως θετικά συσχετισμένες μεταβλητές ουσιαστικά περιέχουν την ίδια πληροφορία. Συνεπώς η συνολική πληροφορία που έχουμε δεν είναι τόση όσο θα περιμέναμε γιατί κάποια πληροφορία επαναλαμβάνετε και μας οδηγεί στο να αυξάνει τη διακύμανση της μέσης τιμής.

Μπορεί κάποιος να δει πως ισχύει και το αντίθετο. Δηλαδή ας υποθέσουμε πως ο συντελεστής συσχέτισης είναι αρνητικός. Μπορεί εύκολα κανείς να επιβεβαιώσει πως για να είναι θετικά ορισμένος ο πίνακας $\mathbf{\Sigma}$ θα πρέπει να ισχύει πως $\rho > -\frac{1}{n-1}$. Όχι τυχαία αυτή είναι και η τιμή που μας εξασφαλίζει πως η μέση τιμή έχει θετική διακύμανση. Αν λοιπόν η τιμή του συντελεστή είναι αρνητική τότε η διακύμανση της μέσης τιμής μειώνεται και μάλιστα όσο πιο έντονη (αλλά επιτρεπτή) αρνητική συσχέτιση έχουμε τόσο μεγαλύτερη μείωση της διακύμανσης πετυχαίνουμε.

Αυτό συμβαίνει γιατί αρνητικά συσχετισμένες μεταβλητές περιέχουν περισσότερη πληροφορία. Επομένως μας συμφέρει να έχουμε αρνητικά συσχετισμένες μεταβλητές.

Σημείωση: Η τεχνική να χρησιμοποιούμε αρνητικά συσχετισμένες μεταβλητές χρησιμοποιείται στην προσομοίωση για να μειώσει κανείς την διακύμανση. Η μέθοδος ονομάζεται 'μέθοδος αντιθετικών μεταβλητών' (antithetic variables). □

Εφαρμογή : Γραμμικό μοντέλο.

Έστω το απλό γραμμικό μοντέλο $y_i = \alpha + \beta x_i + \varepsilon_i$. Εκτιμούμε τις παραμέτρους α και β και συνήθως στη συνέχεια θέλουμε να βρούμε την κατανομή της πρόβλεψης σε ένα συγκεκριμένο σημείο, έστω x. Αυτό γίνεται συνήθως με τη χρήση του τύπου

$$\hat{y} = \hat{\alpha} + \hat{\beta} x,$$

$\hat{\alpha}, \hat{\beta}$ είναι οι εκτιμήσεις των α και β αντίστοιχα. Επομένως η πρόβλεψη είναι ένας γραμμικός συνδυασμός των τυχαίων μεταβλητών $\hat{\alpha}, \hat{\beta}$. Οι υποθέσεις του γραμμικού μοντέλου είναι πως

$$\varepsilon_i \sim N(0, \sigma^2) \qquad \text{και πως}$$

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$$

Οι υποθέσεις αυτές μπορούν να περιγραφούν ισοδύναμα ως εξής. Αν συμβολίσουμε με **ε** το τυχαίο διάνυσμα $\boldsymbol{\varepsilon}' = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$. Οι παραπάνω υποθέσεις γράφονται ισοδύναμα ως

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$$

Παρατηρείστε πως ο πίνακας διακύμανσης είναι διαγώνιος, άρα οι τμ ανεξάρτητες, η διακύμανση σταθερή και όλοι οι μέσοι ίσοι με 0.

Γνωρίζουμε πως αν ισχύουν οι υποθέσεις του γραμμικού μοντέλου τότε το διάνυσμα

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \sim N_2\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \sigma^2 \left(X'X \right)^{-1} \right)$$

δηλαδή η από κοινού κατανομή των παραμέτρων είναι η διδιάστατη κανονική κατανομή, όπου σ^2 είναι η (ας υποθέσουμε γνωστή) διακύμανση του μοντέλου και X ο γνωστός πίνακας σχεδιασμού. □

Επομένως η κατανομή του $\hat{y}$ θα είναι κανονική κατανομή με μέσο $\begin{bmatrix} 1 & x \end{bmatrix} \begin{bmatrix} a \\ \beta \end{bmatrix} = a + \beta x$ και διακύμανση $\begin{bmatrix} 1 & x \end{bmatrix} \sigma^2 (X'X)^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}$.

Θεώρημα Έστω $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και έστω $\mathbf{A}_{q \times p}$ ένας πίνακας διαστάσεων ($q \times p$) από σταθερούς αριθμούς (όχι τυχαίες μεταβλητές) και b ένα διάνυσμα (q×1). Τότε για το τυχαίο διάνυσμα $\mathbf{y} = \mathbf{A}\mathbf{x} + b$ (q×1) ισχύει πως: $\mathbf{y} \sim N_q(\mathbf{A}\boldsymbol{\mu} + b, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$

Εφαρμογή: Έστω οι τ.μ. X_1, X_2, X_3 που αποτελούν το τυχαίο διάνυσμα $\mathbf{x}' = (X_1, X_2, X_3)$ και έστω πως $\mathbf{x} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Ορίζουμε νέες μεταβλητές $Y_1 = X_1 - X_2$ και $Y_2 = X_2 - X_3$. Θέλουμε να βρούμε τη συνδιακύμανση των Y_1, Y_2.

Έστω το διάνυσμα $\mathbf{y} = \mathbf{A}\mathbf{x}$, όπου $\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ και $\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$. Μπορεί κανείς εύκολα να δει πως η μορφή του **A** είναι η κατάλληλη για να μετασχηματίσουμε το αρχικό διάνυσμα **x** στο διάνυσμα **y**.

Ο πίνακας συνδιακύμανσης θα είναι ο

$$Cov(\mathbf{y}) = \mathbf{A}Cov(\mathbf{x})\mathbf{A}' = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 - \sigma_{12} & \sigma_{12} - \sigma_2^2 & \sigma_{13} - \sigma_{23} \\ \sigma_{12} - \sigma_{13} & \sigma_2^2 - \sigma_{23} & \sigma_{23} - \sigma_3^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} =$$

$$\begin{bmatrix} \sigma_1^2 - 2\sigma_{12} + \sigma_2^2 & \sigma_{12} + \sigma_{23} - \sigma_{13} - \sigma_2^2 \\ \sigma_{12} + \sigma_{23} - \sigma_{13} - \sigma_2^2 & \sigma_{21}^2 - 2\sigma_{23} + \sigma_3^2 \end{bmatrix}$$

Συνεπώς η συνδιακύμανση θα είναι $\sigma_{12} + \sigma_{23} - \sigma_{13} - \sigma_2^2$ όπου σ_i^2 είναι η διακύμανση της τμ X_i.

Ένα ακόμα ενδιαφέρον αποτέλεσμα είναι το εξής:

Γνωρίζουμε από τη θεωρία πως η κατανομή χ^2 προκύπτει αν υψώσουμε μια τυποποιημένη κανονική τμ στο τετράγωνο, ή στη γενική μορφή

Αν $X_i \sim N(0,1)$, $i = 1, ..., n$ και τα X_i είναι ανεξάρτητες τ.μ., τότε ισχύει πως $\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$.

Αν κανείς ορίσει το διάνυσμα $\mathbf{x}' = [X_1, X_2, ..., X_p]$ τότε $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} X_i^2$, δηλαδή η κατανομή χ^2 προκύπτει από ένα τυχαίο διάνυσμα από ανεξάρτητες τυχαίες μεταβλητές.

Θα δείξουμε πως

Θεώρημα: Αν $\mathbf{x}_{p\times 1} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, τότε $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_p^2$.

Απόδειξη:

Επειδή $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, μπορεί κανείς εύκολα να δει από το προηγούμενο θεώρημα πως

$$\mathbf{z} = \Sigma^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I}),$$

και επομένως όλα τα στοιχεία του $\mathbf{z}$ είναι ανεξάρτητα μεταξύ τους.

Όμως

$$\begin{aligned}
&(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) = \\
&(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} (\mathbf{x}-\boldsymbol{\mu}) = \\
&\left[\boldsymbol{\Sigma}^{-1/2} (\mathbf{x}-\boldsymbol{\mu})\right]' \left[\boldsymbol{\Sigma}^{-1/2} (\mathbf{x}-\boldsymbol{\mu})\right] = \\
&\mathbf{z}'\mathbf{z} = \sum_{i=1}^{p} Z_i^2
\end{aligned}$$

και επομένως αφού το διάνυσμα $\mathbf{z}$ αποτελείται από ανεξάρτητες τυποποιημένες κανονικές τμ (προσέξτε πως ο πίνακας διακύμανσης είναι ο μοναδιαίος πίνακας) προκύπτει πως

$$(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \sim \chi_p^2$$

Ελλειψοειδή ίσης πιθανότητας

Η εξίσωση $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$ είναι μια εξίσωση έλλειψης. Μάλιστα η ποσότητα αυτή καθορίζει τις υψομετρικές καμπύλες σταθερής πυκνότητας, δηλαδή κάθε σημείο πάνω σε αυτή την έλλειψη (ή το ελλειψοειδές για περισσότερες διαστάσεις) έχει την ίδια ακριβώς πυκνότητα.

Επομένως μπορούμε να ορίσουμε πως

$$P\left[(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \le \chi_{p,a}^2\right] = a$$

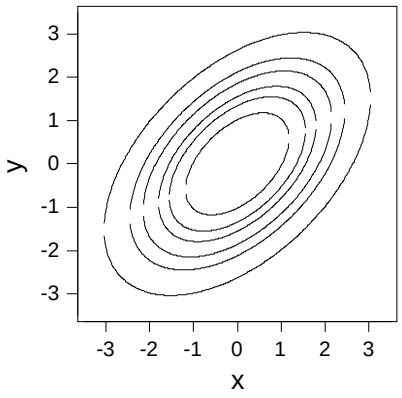
όπου $\chi_{p,a}^2$ είναι το α- ποσοστιαίο σημείο της κατανομής χ^2 με p βαθμούς ελευθερίας.

Αυτό μας επιτρέπει να κατασκευάζουμε διαστήματα εμπιστοσύνης, ή καλύτερα περιοχές εμπιστοσύνης (confidence regions) για τα τυχαία διανύσματα. Οι περιοχές εμπιστοσύνης είναι τα αντίστοιχα των διαστημάτων εμπιστοσύνης για τη μονομεταβλητή περίπτωση. Το διάνυσμα $\boldsymbol{\mu}$ είναι το κέντρο όλων των ισοϋψών καμπυλών. Βέβαια για περισσότερες από 2 μεταβλητές δεν είναι καθόλου εύκολο να αναπαραστήσουμε γραφικά την περιοχή εμπιστοσύνης. Στις 2 διαστάσεις αυτή θα είναι μια έλλειψη, στις 3 διαστάσεις θα μοιάζει με μπάλα του ράγκμπι ενώ για περισσότερες δεν μπορεί να αναπαρασταθεί με κάποιο σχήμα.

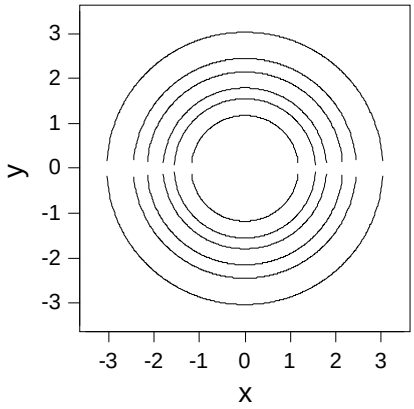
Βέβαια κανείς μπορεί υπολογίζοντας την εξίσωση της έλλειψης να δει αν το σημείο που τον ενδιαφέρει είναι μέσα ή έξω από την περιοχή αυτή. Επίσης ο παραπάνω ορισμός έχει περιορισμένη χρησιμότητα καθώς υποθέτει πως το διάνυσμα των μέσων και ο πίνακας

διακύμανσης είναι γνωστά. Θα δούμε αργότερα τι γίνεται με την περίπτωση που αυτά είναι άγνωστα και πρέπει να εκτιμηθούν.

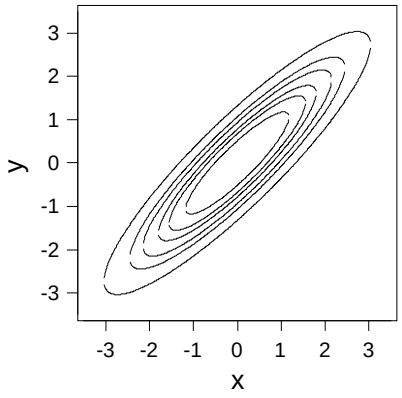
Στο γράφημα 3.5 μπορεί να δει κανείς τα ελλειψοειδή για μια σειρά από τιμές για την πιθανότητα. Έτσι η εσωτερική έλλειψη αντιστοιχεί σε πιθανότητα 0.50 (με απλά λόγια η πιθανότητα μια παρατήρηση να είναι μέσα σε αυτή την έλλειψη είναι 0.5, ενώ με σειρά εμφάνισης από μέσα προς τα έξω οι ελλείψεις αντιστοιχούν σε πιθανότητα 0.5, 0.7, 0.8, 0.9, 0.95, 0.99. Μπορεί να παρατηρήσει κανείς πως αν τα δεδομένα είναι ασυσχέτιστα ($\rho=0$) τότε η έλλειψη εκφυλίζεται σε κύκλο.



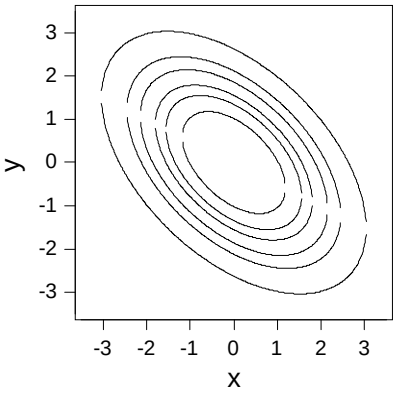
$\rho=0.5$



$\rho=0$



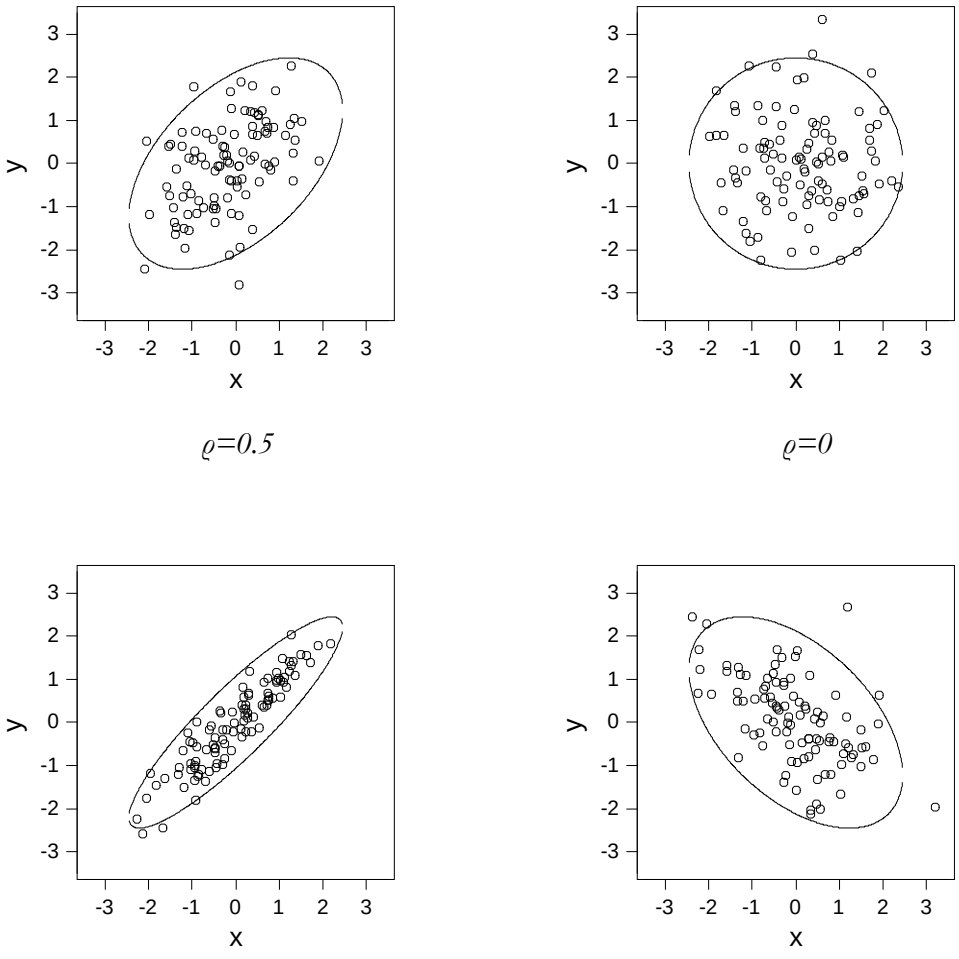
$\rho=09$



$\rho=-0.5$

Γράφημα 3.5. Ελλειψοειδή ίσης πυκνότητας για διάφορες τιμές του ρ. Το διάνυσμα των μέσων ήταν [0 0] για όλα τα γραφήματα, το ίδιο και οι διακυμάνσεις, ήταν όλες ίσες με 1

Στις περισσότερες διαστάσεις όταν οι μεταβλητές είναι ασυσχέτιστες προκύπτει πως το υπερελλειψοειδές γίνεται υπερσφαίρα, για αυτό και πολλές φορές την έλλειψη συσχέτισης την ονομάζουμε και σφαιρικότητα των δεδομένων. Το πρόσημο του συντελεστή συσχέτισης καθορίζει την κλίση των ελλείψεων. Τέλος η απόλυτη τιμή του συντελεστή καθορίζει τα χαρακτηριστικά της έλλειψης. Για την τιμή *ρ=0.9* η έλλειψη είναι πολύ στενή, όσο η τιμή τείνει στο 1 η έλλειψη τείνει να εκφυλιστεί σε γραμμή.



Γράφημα 3.6. 95% περιοχή εμπιστοσύνης και διάγραμμα σημείων για προσομοιωμένα δείγματα από την διμεταβλητή κανονική κατανομή που υποθέσαμε. Το διάνυσμα των μέσων ήταν [0 0] για όλα τα γραφήματα, το ίδιο και οι διακυμάνσεις, ήταν όλες ίσες με 1. Από τον τρόπο κατασκευής της έλλειψης περιμένουμε περίπου το 95% των παρατηρήσεων να είναι μέσα στην έλλειψη

Για να εξετάσουμε λίγο την ιδέα των περιοχών εμπιστοσύνης για ένα τυχαίο διάνυσμα, προσομοιώσαμε δείγματα μεγέθους 100 από συγκεκριμένες διμεταβλητές κανονικές κατανομές. Για τον τρόπο προσομοίωσης θα μιλήσουμε αργότερα. Στο γράφημα 3.6 μπορεί κανείς να δει το διάγραμμα σημείων για το προσομοιωμένο δείγμα καθώς και μια 95% περιοχή

εμπιστοσύνης για κάθε παρατήρηση. Οι διμεταβλητές κατανομές είχαν διάνυσμα μέσων $\mu' = [0 \quad 0]$ και πίνακα διακύμανσης $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ για διάφορες τιμές του *ρ*. Θα πρέπει να τονιστεί πως η επιλογή αυτή των παραμέτρων δεν επηρεάζει τα όποια συμπεράσματα καθώς το διάνυσμα των μέσων απλά μετακινεί το κέντρο της έλλειψης, ενώ ο πίνακας διακύμανσης απλά καθορίζει το μήκος των αξόνων. Επειδή οι περιοχές εμπιστοσύνης ήταν 95% αυτό σημαίνει πως περιμένουμε περίπου το 95% των παρατηρήσεων να είναι μέσα στην έλλειψη. Παρατηρείστε πως γενικά αυτό συμβαίνει. (Για λόγους τυχαιότητας σε κάποιες περιπτώσεις υπάρχουν λιγότερα ή περισσότερα από5 σημεία έξω από τις ελλείψεις, αυτό ήταν αναμενόμενο)

3.5 Περιθώριες και δεσμευμένες κατανομές

Περιθώριες κατανομές

Έστω $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και $\mathbf{x}' = \left(X_1, X_2, \ldots, X_p\right)$. Έστω πως θέλουμε να βρούμε την από κοινού περιθώρια κατανομή των q πρώτων μεταβλητών, δηλαδή του $\mathbf{y}$ όπου $\mathbf{y}' = \left(X_1, X_2, \ldots, X_q\right)$, q<p. Θα πρέπει να πούμε πως η σειρά των μεταβλητών δεν παίζει καμιά σημασία και απλά θεωρούμε τις *q* πρώτες για λόγους ευκολίας.

$$\mathbf{y} = \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_q \end{bmatrix} = \begin{bmatrix} X_1 + 0 \cdot X_2 + \ldots 0 \cdot X_q \ldots + 0 \cdot X_p \\ 0 \cdot X_1 + X_2 + \ldots 0 \cdot X_q \ldots + 0 \cdot X_p \\ \ldots \\ 0 \cdot X_1 + 0 \cdot X_2 + \ldots X_q + \ldots + 0 \cdot X_p \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 0 & \ldots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \ldots & 0 & 0 & \cdots & & 0 \\ \ldots & \ldots & \ldots & \ldots & & \ddots & \ddots & \\ \underbrace{0 \quad \ldots \quad \ldots \quad 1}_{q \times q} & & & & \underbrace{0 \quad \ldots \quad \ldots \quad 0}_{q \times (p-q)} & & & \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_p \end{bmatrix} =$$

$$\begin{bmatrix} \underset{q \times q}{\mathbf{I}} & \vdots & \underset{q \times (p-q)}{\mathbf{0}} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_p \end{bmatrix}$$

όπου $\mathbf{0}_{\mathbf{m \times n}}$ είναι ένας πίνακας με όλα του τα στοιχεία 0 και $\mathbf{I}$ είναι ο μοναδιαίος πίνακας. Δηλαδή $\mathbf{y}_{q \times 1} = \mathbf{A}_{q \times p} \mathbf{x}_{p \times 1}$, όπου $\mathbf{A} = \left[\mathbf{I}_{q \times q} \;\vdots\; \mathbf{0}_{q \times (p-q)}\right]$ και επομένως κάνοντας χρήση του προηγούμενου θεωρήματος προκύπτει πως $\mathbf{y} \sim N_q(\mathbf{A}\boldsymbol{\mu}_x, \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}')$

Χρησιμοποιώντας διαμερισμένους (ή μερικούς) πίνακες (partitioned matrices) και διανύσματα μπορούμε να γράψουμε

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \dots \\ x_q \\ \hline x_{q+1} \\ \dots \\ x_p \end{bmatrix} = \begin{bmatrix} \underset{q\times 1}{\mathbf{x}_1} \\ \hline \underset{(p-q)\times 1}{\mathbf{x}_2} \end{bmatrix} \qquad \text{οπότε } \mathbf{y} = \mathbf{x}_1 \text{ και}$$

$$\text{Όπου } \boldsymbol{\mu}_x = \begin{bmatrix} \mu_1 \\ \dots \\ \mu_q \\ \hline \mu_{q+1} \\ \dots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \underset{q\times 1}{\boldsymbol{\mu}_1} \\ \hline \underset{(p-q)\times 1}{\boldsymbol{\mu}_2} \end{bmatrix}, \ \boldsymbol{\Sigma}_x = \left[\begin{array}{c|c} \underset{q\times q}{\boldsymbol{\Sigma}_{11}} & \underset{q\times(p-q)}{\boldsymbol{\Sigma}_{12}} \\ \hline \underset{(p-q)\times q}{\boldsymbol{\Sigma}_{21}} & \underset{(p-q)\times(p-q)}{\boldsymbol{\Sigma}_{22}} \end{array}\right]$$

Δηλαδή δημιουργούμε τους μερικούς πίνακες παίρνοντας συγκεκριμένα κομμάτια των πινάκων. Παρατηρείστε πως έτσι όπως έχουμε διαμερίσει τον πίνακα συνδιακύμανσης ο υποπίνακας $\boldsymbol{\Sigma}_{11}$ περιέχει τις διακυμάνσεις και τις συνδιακυμάνσεις όλων των μεταβλητών που περιέχονται στο διάνυσμα $\mathbf{x}_1$ και αντίστοιχη ερμηνεία μπορεί να δοθεί στον πίνακα $\boldsymbol{\Sigma}_{22}$. Μόνο αυτοί οι δύο πίνακες από τη διαμέριση του $\boldsymbol{\Sigma}$ είναι πίνακες διακυμάνσεις καθώς οι πίνακες $\boldsymbol{\Sigma}_{12}$ και $\boldsymbol{\Sigma}_{21}$ δεν είναι συμμετρικοί και απλά περιέχουν τις συνδιακυμάνσεις των μεταβλητών του ενός διανύσματος με τις μεταβλητές του άλλου διανύσματος. Τότε βρίσκουμε πως

$$\mathbf{A}\boldsymbol{\mu}_x = \left[\begin{array}{cccc|ccc} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 & \dots & 0 \\ \underbrace{0}_{q\times q} & \dots & \dots & 1 & \underbrace{0}_{q\times(p-q)} & \dots & 0 \end{array}\right] \left[\begin{array}{c} \underset{q\times 1}{\boldsymbol{\mu}_1} \\ \hline \underset{(p-q)\times 1}{\boldsymbol{\mu}_2} \end{array}\right] = \begin{bmatrix} \underset{q\times 1}{\boldsymbol{\mu}_1} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_q \end{bmatrix}$$

και

$$\mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}' = \left[\begin{array}{c|c} \mathbf{I}_{q\times q} & \mathbf{0}_{q\times(p-q)} \end{array}\right] \left[\begin{array}{c|c} \underset{q\times q}{\boldsymbol{\Sigma}_{11}} & \underset{q\times(p-q)}{\boldsymbol{\Sigma}_{12}} \\ \hline \underset{(q-p)\times q}{\boldsymbol{\Sigma}_{21}} & \underset{(q-p)\times(q-p)}{\boldsymbol{\Sigma}_{22}} \end{array}\right] \left[\begin{array}{c} \mathbf{I}_{q\times q} \\ \hline \mathbf{0}_{q\times(p-q)} \end{array}\right] = \begin{bmatrix} \underset{q\times q}{\boldsymbol{\Sigma}_{11}} & \underset{q\times(p-q)}{\boldsymbol{\Sigma}_{12}} \end{bmatrix} \left[\begin{array}{c} \mathbf{I}_{q\times q} \\ \hline \mathbf{0}_{q\times(p-q)} \end{array}\right] = \underset{q\times q}{\boldsymbol{\Sigma}_{11}}$$

Επομένως

$$\mathbf{y} \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) .$$

Δηλαδή κάθε περιθώρια (είτε απλή περιθώρια, δηλαδή για καθεμία μεταβλητή είτε από κοινού περιθώρια, δηλαδή για ζεύγη, τριάδες κλπ μεταβλητών) είναι πολυμεταβλητή κανονική κατανομή. Επομένως και κάθε απλή περιθώρια κατανομή είναι μια απλή κανονική κατανομή.

Δεσμευμένες κατανομές

Θα δούμε τώρα πως και όλες οι δεσμευμένες κατανομές είναι και αυτές πολυμεταβλητές κανονικές κατανομές.

Ας υποθέσουμε τη διαμέριση των πινάκων και των διανυσμάτων ως εξής:

$$\mathbf{x}_{p\times 1}:\ \mathbf{x}=\begin{bmatrix}\mathbf{x}_1\\ \underset{q\times 1}{}\\ \mathbf{x}_2\\ \underset{m\times 1}{}\end{bmatrix},\ \text{όπου q+m=p},\ \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_1\\ \underset{q\times 1}{}\\ \boldsymbol{\mu}_2\\ \underset{m\times 1}{}\end{bmatrix}\ \text{και}\ \boldsymbol{\Sigma}=\left[\begin{array}{c:c}\underset{q\times q}{\boldsymbol{\Sigma}_{11}} & \underset{q\times m}{\boldsymbol{\Sigma}_{12}}\\ \hdashline \underset{m\times q}{\boldsymbol{\Sigma}_{21}} & \underset{m\times m}{\boldsymbol{\Sigma}_{22}}\end{array}\right],\ \boldsymbol{\Sigma}_{12}=\boldsymbol{\Sigma}_{21}'$$

όπου $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ είναι πίνακες διακύμανσης-συνδιακύμανσης αλλά οι πίνακες $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$ δεν είναι πίνακες διακύμανσης-συνδιακύμανσης. Επίσης αν όλα τα στοιχεία του πίνακα $\boldsymbol{\Sigma}_{12}$ είναι 0, δηλαδή αν $\boldsymbol{\Sigma}_{12}=\mathbf{0}$ τότε τα διανύσματα $\mathbf{x}_1$ και $\mathbf{x}_2$ είναι ανεξάρτητα.

Για να γίνει πιο κατανοητή η διαμέριση ας δούμε ένα παράδειγμα.

Παράδειγμα 3.4: Έστω το διάνυσμα $\mathbf{x}=\begin{bmatrix}Z_1\\ Z_2\\ Z_3\\ Z_4\end{bmatrix}$ το οποίο διαμερίζουμε σε δύο μικρότερα διανύσματα $\mathbf{x}_1=\begin{bmatrix}Z_1\\ Z_2\end{bmatrix}$ και $\mathbf{x}_2=\begin{bmatrix}Z_3\\ Z_4\end{bmatrix}$. Επομένως διαμερίζουμε και το διάνυσμα των μέσων και τον πίνακα διακύμανσης ως

$$\boldsymbol{\mu}_1=\begin{bmatrix}E(Z_1)\\ E(Z_2)\end{bmatrix},\ \boldsymbol{\mu}_2=\begin{bmatrix}E(Z_3)\\ E(Z_4)\end{bmatrix}$$

$$\boldsymbol{\Sigma}_{11}=\begin{bmatrix}Var(Z_1) & Cov(Z_1,Z_2)\\ Cov(Z_1,Z_2) & Var(Z_2)\end{bmatrix},\ \boldsymbol{\Sigma}_{22}=\begin{bmatrix}Var(Z_3) & Cov(Z_3,Z_4)\\ Cov(Z_3,Z_4) & Var(Z_4)\end{bmatrix}$$

$$\boldsymbol{\Sigma}_{12}=\begin{bmatrix}Cov(Z_1,Z_3) & Cov(Z_1,Z_4)\\ Cov(Z_2,Z_3) & Cov(Z_2,Z_4)\end{bmatrix}$$

Στην περίπτωση που οι μεταβλητές δεν είναι στη σειρά πρέπει να προσέξουμε πως θα κατασκευάσουμε τους υποπίνακες που θέλουμε ώστε να είμαστε σίγουροι ότι αντιστοιχούν στην διαμέριση που ζητάμε

Μπορούμε τώρα να δούμε το θεώρημα που αφορά τις δεσμευμένες κατανομές

Θεώρημα

Έστω ένα τυχαίο διάνυσμα $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ το οποίο μπορούμε να διαμερίσουμε όπως δείξαμε πριν. Τότε για τη δεσμευμένη κατανομή θα έχουμε πως

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N_q(\boldsymbol{\mu}', \boldsymbol{\Sigma}'),$$

όπου $\boldsymbol{\mu}' = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$, $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$

Παρατήρηση: Προφανώς μπορούμε να χρησιμοποιήσουμε το παραπάνω αποτέλεσμα για να βρούμε την απλή δεσμευμένη κατανομή μιας μόνο τυχαίας μεταβλητής, όπως επίσης και να βρούμε δεσμευμένη κατανομή αγνοώντας κάποιες άλλες μεταβλητές. Για παράδειγμα αν έχουμε ένα αρχικό διάνυσμα $\mathbf{x}' = (X_1, X_2, X_3)$ μπορούμε να βρούμε τις δεσμευμένες κατανομές των $X_1, X_2 \mid X_3$, $X_1 \mid X_2, X_3$ ή ακόμα και την κατανομή του $X_1 \mid X_3$.

Αυτό που είναι πολύ ενδιαφέρον στο παραπάνω αποτέλεσμα είναι πως ο δεσμευμένος πίνακας διακύμανσης-συνδιακύμανσης δεν εξαρτάται από τις τιμές των μεταβλητών ως προς τις οποίες έχουμε δεσμεύσει. Αυτό σημαίνει πως οποιαδήποτε και αν είναι η τιμή τους η δεσμευμένη διακύμανση είναι η ίδια.

Παράδειγμα 3.5: Έστω πως έχουμε μόλις 2 τμ Χ και Υ και πως η από κοινού κατανομή τους είναι διμεταβλητή κανονική με διάνυσμα μέσων $\mu' = \begin{bmatrix} \mu_x & \mu_y \end{bmatrix}$ και πίνακα διακύμανσης $\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$. Επομένως με χρήση του παραπάνω θεωρήματος προκύπτει πως η δεσμευμένη κατανομή της τμ Υ δοθέντος ότι Χ=x θα είναι κανονική κατανομή με

$$E(Y \mid X = x) = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x) = \left[\mu_y - \frac{\sigma_{xy}}{\sigma_x^2}\mu_x\right] + \frac{\sigma_{xy}}{\sigma_x^2}x \text{ και}$$

$$Var(Y \mid X = x) = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}$$

Αν χρησιμοποιήσουμε το συντελεστή συσχέτισης $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ προκύπτει πως

$$Var(Y \mid X = x) = \sigma_y^2 (1 - \rho^2)$$

Από τους παραπάνω τύπους υπάρχουν μερικά πολύ χρήσιμα συμπεράσματα:

- Κατ' αρχάς είναι προφανές πως η δεσμευμένη διακύμανση δεν εξαρτάται από την τιμή x της τμ X που έχουμε ως δέσμευση, δηλαδή την πληροφορία. Γενικά αυτό δεν συμβαίνει συχνά, δηλαδή η δεσμευμένη διακύμανση να μην εξαρτάται από τη δέσμευση.

- Επίσης μπορεί κανείς να δει πως η δεσμευμένη αναμενόμενη τιμή είναι γραμμική ως προς την τιμή της τμ X, δηλαδή ως προς τη δέσμευση. Στη στατιστική η δεσμευμένη αναμενόμενη τιμή ονομάζεται παλινδρόμηση (regression). Ο όρος παλινδρόμηση είναι γνωστός από το γραμμικό μοντέλο αλλά αυτό είναι λογικό αφού το γραμμικό μοντέλο στηρίζεται στις υποθέσεις που μόλις τώρα είδαμε, δηλαδή πως η αναμενόμενη τιμή του Υ δοθέντος του $X=x$ είναι γραμμική ως προς το x και πως η διακύμανση είναι σταθερή και ανεξάρτητη από το x. Επίσης η κατανομή του Υ είναι η κανονική. Οι υποθέσεις της γραμμικής παλινδρόμησης είναι αυτές που βλέπουμε για τη δεσμευμένη κατανομή (Παρατήρηση: στο γραμμικό μοντέλο το X δεν είναι τμ αλλά γνωστό εκ των προτέρων, αυτό δηλαδή που ουσιαστικά δηλώνει μια δεσμευμένη κατανομή)).

- Οι ομοιότητες δεν σταματούν εδώ καθώς αν κοιτάξει κανείς τη δεσμευμένη αναμενόμενη τιμή θα αναγνωρίζει πως τα α και β της γραμμικής παλινδρόμησης εκτιμούνται με τα δειγματικά αντίστοιχα των τύπων που βλέπετε στον τύπο της. Ανατρέχοντας δηλαδή σε οποιοδήποτε βιβλίο γραμμικής παλινδρόμησης θα δει κανείς πω; Οι εκτιμήτριες ελαχίστων τετραγώνων για τα α και β δίνονται από τις σχέσεις $\hat{a} = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}$, $\hat{\beta} = \frac{s_{xy}}{s_x^2}$ οι οποίες είναι τα δειγματικά αντίστοιχα αυτών που βλέπουμε στη δεσμευμένη αναμενόμενη τιμή.

- Επίσης πολύ ενδιαφέρον είναι να δει κανείς πως η δεσμευμένη διακύμανση του Υ είναι σε κάθε περίπτωση μικρότερη ή ίση της διακύμανσης του Υ, δηλαδή $Var(Y \mid X = x) \le Var(Y)$ με την ισότητα να ισχύει μόνο αν $\rho=0$, δηλαδή οι δύο μεταβλητές είναι ασυσχέτιστες. Αυτό σημαίνει πως αν υπάρχει έστω και μικρή συσχέτιση ανάμεσα σε δύο μεταβλητές μπορούμε να μειώσουμε τη διακύμανση της μιας έχοντας κάποια πληροφορία για την άλλη. Ουσιαστικά αυτό είναι το βασικό συστατικό κάθε πολυμεταβλητής ανάλυσης, να χρησιμοποιήσουμε δηλαδή πληροφορίες από άλλες μεταβλητές ώστε να μειώσουμε τη διακύμανση (και άρα να βελτιώσουμε τη δυνατότητα μας να ελέγξουμε την αβεβαιότητα) της μεταβλητής που μας ενδιαφέρει.

- Το πρόσημο της συσχέτισης ανάμεσα σε δύο μεταβλητές δεν παίζει ρόλο.

Παράδειγμα 3.6: Έστω το τυχαίο διάνυσμα $\mathbf{x}' = (X_1, X_2, X_3, X_4)$ και έστω πως $\mathbf{x} \sim N_4(\mu, \Sigma)$ όπου

$\mu' = \begin{bmatrix} 0 & 1 & 0 & 2 \end{bmatrix}$ και $\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0.5 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 1 \end{bmatrix}$. Να βρεθεί η δεσμευμένη κατανομή των $X_1, X_2 \mid X_3 = x, X_4 = y$.

Γνωρίζουμε πως η δεσμευμένη κατανομή θα είναι πολυμεταβλητή κανονική. Για να βρούμε τις παραμέτρους χρειάζεται να διαμερίσουμε τόσο το διάνυσμα των μέσων όσο και τον πίνακα διακύμανσης. Έτσι θα έχουμε πως

$\mu_1' = \begin{bmatrix} 0 & 1 \end{bmatrix}, \ \mu_2' = \begin{bmatrix} 0 & 2 \end{bmatrix}$

$\Sigma_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \ \Sigma_{22} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \ \Sigma_{12} = \Sigma_{21}' = \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix}$. Παρατηρείστε πως καθώς δεν είναι όλα τα στοιχεία του πίνακα Σ_{12} ίσα με 0, και άρα τα δύο διανύσματα (X_1, X_2) και (X_3, X_4) δεν είναι ανεξάρτητα .

Χρησιμοποιώντας τους τύπους βρίσκουμε πως η δεσμευμένη αναμενόμενη τιμή θα είναι

$$(X_1, X_2 \mid X_3 = x, X_4 = y) \sim N_2(\mu^*, \Sigma^*)$$

όπου

$$\mu^* = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} 0 \\ 2 \end{bmatrix} \right) =$$
$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y-2 \end{bmatrix} =$$
$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} y/2-1 \\ 0 \end{bmatrix} = \begin{bmatrix} y/2-1 \\ 1 \end{bmatrix}$$

και

$$\Sigma^* = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 0.5 & 0 \end{bmatrix} =$$
$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 0.25 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.75 & 0 \\ 0 & 2 \end{bmatrix}$$

Συνεπώς αν x=1 και y=1 προκύπτει πως

$$x_1 \mid x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \sim N_2\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.75 & 0 \\ 0 & 2 \end{bmatrix} \right)$$

ενώ στη γενική μορφή η δεσμευμένη κατανομή θα είναι η

$$x_1 \mid x_2 = \begin{bmatrix} x \\ y \end{bmatrix} \sim N_2\left(\begin{bmatrix} y/2-1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.75 & 0 \\ 0 & 2 \end{bmatrix} \right)$$

Αν δούμε τα αποτελέσματα παρατηρούμε πως τα X_1 και X_2 είναι και πάλι ανεξάρτητα αλλά τώρα η διακύμανση της τμ X_1 έχει μικρύνει.

3.6 Εκτίμηση παραμέτρων

Οι άγνωστες παράμετροι της πολυμεταβλητής κανονικής κατανομής είναι το διάνυσμα των μέσων $\boldsymbol{\mu}$ (διάνυσμα p×1) και ο πίνακας $\boldsymbol{\Sigma}$ διαστάσεων p× p. Δεδομένου πως ο πίνακας $\boldsymbol{\Sigma}$ είναι συμμετρικός, δεν χρειάζεται να εκτιμήσει κανείς όλα του τα στοιχεία. Στην πραγματικότητα χρειάζεται να εκτιμήσει p(p+1)/2 παραμέτρους (διακυμάνσεις και συνδιακυμάνσεις) και άλλες p παραμέτρους από το διάνυσμα των μέσων, άρα συνολικά p(p+1)/2+p άγνωστες παραμέτρους. Καταλαβαίνει κανείς πως το πρόβλημα της εκτίμησης είναι αρκετά πολύπλοκο καθώς χρειαζόμαστε p(p+3)/2 εξισώσεις.

Μέθοδος των ροπών

Οι εκτιμήτριες με την μέθοδο των ροπών είναι:

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} \text{ και } \hat{\boldsymbol{\Sigma}} = \mathbf{S}$$

και προκύπτουν εύκολα αν αναλογιστούμε την ερμηνεία των παραμέτρων της πολυμεταβλητής κανονικής κατανομής, (δειγματικό διάνυσμα μέσων και πίνακας διακύμανσης-συνδιακύμανσης).

Μέθοδος μέγιστης πιθανοφάνειας

Έστω πως έχουμε ένα δείγμα μεγέθους n από πολυμεταβλητή κανονική κατανομή, δηλαδή υποθέτουμε πως $X_i \sim N_p(\mu, \Sigma)$, *i=1,…,n.*

Τότε η πιθανοφάνειας του δείγματος δίνεται ως

$$L = f(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)$$

και άρα

$$L = (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left[(\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \right\}.$$

Λογαριθμίζοντας βρίσκουμε πως

$$l = \ln L = -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln|\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} \left[(\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] =$$
$$= \text{constant} - \frac{n}{2} \ln|\boldsymbol{\Sigma}| - \frac{n}{2} tr\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) - \frac{n}{2} (\bar{\mathbf{x}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})$$

γιατί

$$\sum_{i=1}^{n} \left[(\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] = \sum_{i=1}^{n} \left[((\mathbf{x}_i - \bar{x}) + (\bar{x} - \boldsymbol{\mu}))' \boldsymbol{\Sigma}^{-1} ((\mathbf{x}_i - \bar{x}) + (\bar{x} - \boldsymbol{\mu})) \right]$$
$$\sum_{i=1}^{n} \left[(\mathbf{x}_i - \bar{x})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{x}) + (\bar{x} - \mu)' \boldsymbol{\Sigma}^{-1} (\bar{x} - \mu) + 2(\bar{x} - \mu)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{x}) \right] =$$
$$\sum_{i=1}^{n} \left[(\mathbf{x}_i - \bar{x})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{x}) \right] + n(\bar{x} - \mu)' \boldsymbol{\Sigma}^{-1} (\bar{x} - \mu)$$

επειδή

$$\sum_{i=1}^{n} \left[2(\bar{x} - \mu)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{x}) \right] = 2(\bar{x} - \mu)' \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{x}) = 0$$

Όμως από τη γραμμική άλγεβρα ξέρουμε ότι αν τα x_i είναι διανύσματα $p \times 1$ και Α πίνακας $p \times p$ τότε ισχύει πως

$$\sum_{i=1}^{n} x_i' A x_i = tr(AT)$$

όπου $T = \sum_{i=1}^{n} x_i x_i'$ είναι ένας πίνακας διαστάσεων $p \times p$ και άρα

$$\sum_{i=1}^{n} \left[(\mathbf{x}_i - \bar{x})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{x}) \right] = \sum_{i=1}^{n} tr\left[\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{x})' (\mathbf{x}_i - \bar{x}) \right] =$$
$$tr\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \left[(\mathbf{x}_i - \bar{x})' (\mathbf{x}_i - \bar{x}) \right] = tr\boldsymbol{\Sigma}^{-1}(nS) = n\ tr\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)$$

Έτσι,

$$\sum_{i=1}^{n}\left[\left(\mathbf{x}_i-\bar{x}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\bar{x}\right)\right]+n\left(\bar{x}-\mu\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{x}-\mu\right)=$$

$$=n\ tr\boldsymbol{\Sigma}^{-1}S+n\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)$$

Επομένως για να μεγιστοποιήσει κανείς την πιθανοφάνεια ως προς το μ ουσιαστικά ενδιαφέρεται για την ποσότητα

$$-\frac{n}{2}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)$$

η οποία είναι οπωσδήποτε αρνητικός αριθμός ως το αρνητικό μιας τετραγωνικής μορφής, και επομένως για $\boldsymbol{\mu}=\bar{\mathbf{x}}$ η συνάρτηση μεγιστοποιείται.

Για να δείξουμε πως προκύπτει η αντίστοιχη εκτιμήτρια μεγίστης πιθανοφάνειας για το $\boldsymbol{\Sigma}$ δεν είναι τόσο απλό και για αυτό η απόδειξη δεν δίνεται σε αυτές τις σημειώσεις. Προκύπτει όμως πως $\hat{\boldsymbol{\Sigma}}=\mathbf{S}$, δηλαδή οι εκτιμήτριες μεγίστης πιθανοφάνειας ταυτίζονται με τις εκτιμήτριες της μεθόδου των ροπών. Όπως θα δούμε αργότερα η εκτιμήτρια για τη διακύμανση είναι μεροληπτική.

3.7 Προσομοίωση δεδομένων από πολυμεταβλητή κανονική κατανομή

Όπως είδαμε προηγουμένως ισχύει πως αν $\mathbf{x}\sim N_p(\mathbf{0},\mathbf{I})$ τότε ο μετασχηματισμός $\mathbf{y}=\boldsymbol{\Sigma}^{1/2}\mathbf{x}+\boldsymbol{\mu}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Επομένως μπορεί κανείς να προσομοιώσει από τη ζητούμενη πολυμεταβλητή κανονική κατανομή ξεκινώντας με ένα διάνυσμα ανεξάρτητων τυποποιημένων κανονικών μεταβλητών και χρησιμοποιώντας στη συνέχεια το μετασχηματισμό που περιγράψαμε. Ουσιαστικά το πρόβλημα είναι η εύρεση του πίνακα $\boldsymbol{\Sigma}^{1/2}$, ο οποίος μπορεί να βρεθεί με την ανάλυση Cholesky (Cholesky decomposition) ενός συμμετρικού πίνακα. Επομένως ο αλγόριθμος προσομοίωσης είναι ο εξής

Βήμα 1: Προσομοίωσε p τυποποιημένες κανονικές τυχαίες μεταβλητές, $X_1,...,X_p$. Με αυτές όρισε το διάνυσμα $\mathbf{x}'=(X_1,...,X_p)$

Βήμα 2: Θέσε $\mathbf{y}=\boldsymbol{\Sigma}^{1/2}\mathbf{x}+\boldsymbol{\mu}$

Το **y** είναι ένα τυχαίο διάνυσμα που ακολουθεί τη ζητούμενη $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ κατανομή

Εναλλακτικά μπορεί κανείς να προσομοιώσει από μια πολυμεταβλητή κανονική κατανομή χρησιμοποιώντας την ιδιότητα πως όλες οι περιθώριες και όλες οι δεσμευμένες κατανομές είναι και αυτές πολυμεταβλητές κανονικές. Δηλαδή ένας αλγόριθμος είναι ο εξής:

- Παρήγαγε μια τμ X_1 από την περιθώρια κατανομή της X_1
- Παρήγαγε μια τμ από την κατανομή της $X_2 \mid X_1 = x_1$ (η οποία είναι μια κανονική κατανομή με κάποιες παραμέτρους που μπορεί κανείς σχετικά εύκολα να υπολογίσει)
- Παρήγαγε μια τμ από την κατανομή της $X_3 \mid X_1 = x_1, X_2 = x_2$ (η οποία και πάλι είναι μια κανονική κατανομή με κάποιες παραμέτρους που μπορεί κανείς σχετικά εύκολα να υπολογίσει)
- ….
- Παρήγαγε μια τμ από την κατανομή της $X_p \mid X_1 = x_1, X_2 = x_2, \ldots, X_{p-1} = x_{p-1}$ (η οποία και πάλι είναι μια κανονική κατανομή με κάποιες παραμέτρους που μπορεί κανείς σχετικά εύκολα να υπολογίσει)

Στην πραγματικότητα αυτός ο αλγόριθμος δεν είναι ιδιαίτερα αποδοτικός καθώς χρειάζεται να υπολογίζεται σε κάθε βήμα η δεσμευμένη μέση τιμή και η αντίστοιχη διακύμανση. Παρόλα αυτά στη διμεταβλητή περίπτωση τα πράγματα είναι πιο εύκολα και ο αλγόριθμος αρκετά γρήγορος. Έτσι για αυτή την περίπτωση έχουμε:

Έστω ότι θέλουμε να γεννήσουμε ένα τυχαίο διάνυσμα από μια διμεταβλητή κανονική κατανομή με διάνυσμα μέσων τιμών $\boldsymbol{\mu}' = (\mu_1, \mu_2)$ και πίνακας διακύμανσης $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$.

Η συσχέτιση επομένως είναι $\rho = \dfrac{\sigma_{12}}{\sigma_1 \sigma_2}$. Έτσι για να προσομοιώσουμε από αυτή την κατανομή

Βήμα 1: Προσομοίωσε μια τμ $x_1 \sim N(\mu_1, \sigma_1^2)$

Βήμα 2: Προσομοίωσε μια τμ $x_2 \sim N\left(\mu_2 + \rho \dfrac{\sigma_2}{\sigma_1}(x_1 - \mu_1),\ \sigma_2^2(1-\rho^2) \right)$

4 ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Πριν ξεκινήσουμε την περιγραφή δειγματοληπτικών κατανομών, καθώς και πολυμεταβλητές κατανομές σχετικά με αυτές θα παρουσιάσουμε μερικές ενδιαφέρουσες μονοδιάστατες κατανομές με σκοπό να δούμε αμέσως μετά τη γενίκευση τους στην πολυμεταβλητή περίπτωση.

4.1 Μη κεντρικές κατανομές

Οι μη κεντρικές κατανομές είναι γενικεύσεις γνωστών κατανομών και προκύπτουν με ανάλογο τρόπο. Συνήθως η διαφορά τους είναι πως περιέχουν άλλη μια παράμετρο που συνήθως ονομάζεται παράμετρος μη κεντρικότητας. Συνήθως οι περισσότερες κατανομές ξεκινούν από τυποποιημένες κανονικές τυχαίες μεταβλητές. Οι μη κεντρικές κατανομές προκύπτουν συνήθως όταν οι τυχαίες μεταβλητές δεν είναι τυποποιημένες δηλαδή δεν έχουν μέση τιμή 0.

Μη κεντρική χ^2 κατανομή

Από τη θεωρία κατανομών είναι γνωστό πως η κατανομή χ^2 προκύπτει ως το άθροισμα τετραγώνων ανεξάρτητων κανονικών κατανομών. Δηλαδή

$$X_i \sim N(0,1), \quad i = 1,...,n$$
$$Z = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

ή ισοδύναμα

$$X_i \sim N(0,\sigma_i^2), \quad i = 1,...,n$$
$$Z = \sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2} \sim \chi_n^2$$

Στην περίπτωση που οι κανονικές τμ δεν ακολουθούν κανονική κατανομή με μέση τιμή 0, τότε οδηγούμαστε στην μη κεντρική χ^2 κατανομή, και συγκεκριμένα

$$X_i \sim N(\mu_i, \sigma_i^2), \quad i = 1,\ldots,n$$
$$Z = \sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2} \sim \chi_n^2(\delta)$$

όπου $\chi_n^2(\delta)$ είναι η μη κεντρική χ^2 κατανομή με παράμετρο μη κεντρικότητας $\delta = \sum_{i=1}^{n} \mu_i^2$. Μπορεί εύκολα κανείς να δει πως αν $\mu_i = 0$ τότε και $\delta=0$ και επομένως οδηγούμαστε στη γνωστή χ^2 κατανομή. Είναι ενδιαφέρον πως αν $Z \sim \chi_n^2(\delta)$ τότε $E(Z)= n + \delta$, $Var(Z) = 2n + 4\delta$.

Μη κεντρική F κατανομή

Η κατανομή F προκύπτει ως ο λόγος δύο χ^2 τυχαίων μεταβλητών. Δηλαδή

$$X \sim \chi_n^2, \; Y \sim \chi_m^2$$
$$Z = \frac{m}{n}\frac{X}{Y} \sim F(n,m)$$

Αν τώρα η τυχαία μεταβλητή του αριθμητή δεν είναι κεντρική χ^2 αλλά μη κεντρική προκύπτει η μη κεντρική κατανομή F, δηλαδή

$$X \sim \chi_n^2(\delta), \; Y \sim \chi_m^2$$
$$Z = \frac{m}{n}\frac{X}{Y} \sim F(n,m;\delta)$$

όπου και πάλι δ είναι η παράμετρος μη κεντρικότητας

Μη κεντρική t κατανομή

Η κατανομή t-student (ή απλά κατανομή t) προκύπτει ως ο λόγος μιας τυποποιημένης κανονικής μεταβλητής και της τετραγωνικής ρίζας μιας χ^2 τυχαίας μεταβλητής. Δηλαδή

$$X \sim N(0,1), \; Y \sim \chi_n^2$$
$$Z = \frac{X}{\sqrt{Y/n}} = \frac{\sqrt{n}X}{\sqrt{Y}} \sim t_n$$

Η μη κεντρική κατανομή t προκύπτει αν η κανονική τυχαία μεταβλητή δεν έχει μέσο 0, δηλαδή προκύπτει ως

$$X \sim N(\mu,1),\ Y \sim \chi^2_n$$

$$Z = \frac{X}{\sqrt{Y/n}} = \frac{\sqrt{n}X}{\sqrt{Y}} \sim t_n(\mu)$$

όπου το μ είναι η παράμετρος μη κεντρικότητας

Όλες αυτές οι μη κεντρικές κατανομές προκύπτουν ως οι κατανομές γνωστών μας ελεγχοσυναρτήσεων κάτω από την εναλλακτική υπόθεση. Θυμηθείτε πως συνήθως για να κάνουμε τους ελέγχους μας αρκεί η κατανομή της ελεγχοσυνάρτησης κάτω από τη μηδενική υπόθεση και για αυτό συνήθως δεν ασχολούμαστε με το τι γίνεται όταν δεν ισχύει η μηδενική υπόθεση.

4.2 Η κατανομή Wishart

Έστω τα ανεξάρτητα τυχαία διανύσματα $(\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_n)$ μεγέθους p όπου $\mathbf{x}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}),\ i = 1,...,n$. Ο τυχαίος πίνακας $\mathbf{W}$ διαστάσεων $p \times p$ που προκύπτει ως $\mathbf{W} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^{n} \mathbf{W}_i$ λέμε πως ακολουθεί την κατανομή Wishart με παραμέτρους n και $\boldsymbol{\Sigma}$.

Θα συμβολίζουμε την κατανομή Wishart με παραμέτρους n και $\boldsymbol{\Sigma}$ ως $W_p(n, \Sigma)$. Μερικά σημαντικά σημεία για την κατανομή αυτή είναι τα εξής

- Αν $\mathbf{X} \sim$ Wishart (m, $\boldsymbol{\Sigma}$), η συνάρτηση πυκνότητας πιθανότητας του πίνακα $\mathbf{X}$ είναι η

$$f\left(\mathbf{X}\right) = \frac{|\mathbf{X}|^{(m-p-1)/2} \exp\left\{-\frac{1}{2} tr\left(\Sigma^{-1}\mathbf{X}\right)\right\}}{\pi^{p(p-1)/4} 2^{mp/2} |\Sigma|^{m/2} \prod_{i=1}^{p} \Gamma\left(\frac{m-i+1}{2}\right)}.$$

- Για p=1, $|\boldsymbol{\Sigma}|$=1 η κατανομή είναι η χ^2 με m βαθμούς ελευθερίας. Μπορεί κανείς να παρατηρήσει ότι και από τον τρόπο που προκύπτει η κατανομή Wishart είναι γενίκευση της χ^2 κατανομής
- Αν στο μοντέλο γέννησης της κατανομής αφήσουμε τα τυχαία διανύσματα να μην έχουν μέσα διανύσματα 0, δηλαδή αν $\mathbf{x}_i \sim N_p(\mu_i, \boldsymbol{\Sigma}),\ i = 1,...,n$ τότε προκύπτει η μη κεντρική

κατανομή Wishart, (συμβολισμός Wishart (n, $\mathbf{\Sigma}$, $\boldsymbol{\mu}$)) όπου $\mu = \sum_{i=1}^{n} \mu_i$ είναι η παράμετρος μη κεντρικότητας.

- Ισχύει πως $E(\mathbf{W}) = n\mathbf{\Sigma}$ ενώ για τη διακύμανση ενός τυχαίου πίνακα η σχέση είναι ιδιαίτερα πολύπλοκη και δεν θα δοθεί.
- Ενδιαφέρον παρουσιάζει ο αντίστροφος ενός πίνακα **W** που ακολουθεί την κατανομή Wishart. Η κατανομή λοιπόν του αντιστρόφου ακολουθεί την αντίστροφη Wishart κατανομή (inverted Wishart distribution) και έχει αρκετή χρησιμότητα στην Μπευζιανή στατιστική.
- Η κατανομή Wishart ορίζεται μόνο για συμμετρικούς πίνακες. Επομένως και ο πίνακας **Σ** πρέπει να είναι συμμετρικός

Η μεγάλη χρησιμότητα της κατανομής Wishart στην πολυμεταβλητή στατιστική στηρίζεται στο γεγονός πως είναι η κατανομή δειγματοληψίας του δειγματικού πίνακα διακύμανσης. Συγκεκριμένα ισχύει πως

Θεώρημα: Έστω ένα τυχαίο δείγμα από ανεξάρτητα τυχαία διανύσματα $\mathbf{x}_1, ..., \mathbf{x}_n$ (διαστάσεων $p\times1$) . Αν $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$, τότε $n\mathbf{S} = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' \sim W_p(n-1, \mathbf{\Sigma})$.

Το παραπάνω θεώρημα είναι και η αιτία για να χρησιμοποιούμε τον αμερόληπτο δειγματικό πίνακα διακύμανσης. Παρατηρείστε πως η αναμενόμενη τιμή του δειγματικού πίνακα διακύμανσης δεν είναι αμερόληπτη και για αυτό εμφανίζεται το *(n-1)* στον παρονομαστή ώστε η εκτιμήτρια να είναι αμερόληπτη.

Επίσης γνωρίζουμε από πριν πως για το διάνυσμα των μέσων ισχύει πως $\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\mathbf{\Sigma})$. Είναι ενδιαφέρον ότι ισχύει και το επόμενο θεώρημα:

Θεώρημα: Το δειγματικό διάνυσμα των μέσων ($\bar{\mathbf{x}}$) είναι ανεξάρτητο από τον δειγματικό πίνακα διακύμανσης-συνδιακύμανσης (**S**).

Το παραπάνω θεώρημα γενικεύει την ιδιότητα που ξέρουμε από την μονομεταβλητή περίπτωση πως ο δειγματικός μέσος είναι ανεξάρτητος από τη δειγματική διακύμανση.

4.3 Ιδιότητες της κατανομής Wishart

1) Αν $\mathbf{M}_1 \sim W(m_1, \mathbf{\Sigma})$, $\mathbf{M}_2 \sim W(m_2, \mathbf{\Sigma})$ και οι δύο πίνακες είναι ανεξάρτητοι μεταξύ τους τότε $\mathbf{M}_1 + \mathbf{M}_2 \sim W(m_1 + m_2, \mathbf{\Sigma})$

Παρατηρείστε πως οι δύο πίνακες πρέπει να έχουν τον ίδιο πίνακα $\mathbf{\Sigma}$ για να ισχύει το αποτέλεσμα. Προφανώς το αποτέλεσμα γενικεύεται και για περισσότερους από 2 πίνακες

2) Αν $\mathbf{M} \sim W(m, \mathbf{\Sigma})$ και πίνακας $\mathbf{C}_{p\times p}$, τότε $\mathbf{CMC}' \sim W(m, \mathbf{C\Sigma C}')$

Το αποτέλεσμα έχει το εξής ενδιαφέρον. Αν ξεκινήσουμε από ένα τυχαίο διάνυσμα το οποίο μετασχηματίσουμε πολλαπλασιάζοντας το με έναν κατάλληλο πίνακα τότε η διακύμανση του καινούριου διανύσματος θα δίνεται από μια έκφραση αντίστοιχη με αυτές που βλέπουμε στο δεξί μέλος της παραπάνω ιδιότητας. Αυτό σημαίνει πως η διακύμανση του καινούριου διανύσματος θα ακολουθεί και αυτή κατανομή Wishart αν αυτό συνέβαινε για τη διακύμανση του αρχικού διανύσματος.

3) Αν $\mathbf{M} \sim W(m, \mathbf{\Sigma})$ και διάνυσμα $\mathbf{C}_{p\times 1}$, τότε $\dfrac{\mathbf{c}'\mathbf{Mc}}{\sigma^2} \sim \chi^2_m$ όπου $\sigma^2 = \mathbf{c}'\mathbf{\Sigma c}$

Η παραπάνω ιδιότητα μας επιτρέπει να δούμε πως κάθε στοιχείο του πίνακα $\mathbf{M}$ ακολουθεί μια χ^2 κατανομή. Αυτό προκύπτει εύκολα αν για παράδειγμα ορίσει κανείς το διάνυσμα $\mathbf{c}$ να έχει όλο μηδενικά και μόνο μια μονάδα. Για παράδειγμα, αν $c' = [1,0,0,...,0]$ τότε θα ισχύει για το στοιχείο M_{11} του πίνακα $\mathbf{M}$ πως

$\dfrac{M_{11}}{\sigma_1^2} \sim \chi^2_m$ όπου σ_1^2 είναι το (1,1) στοιχείο του πίνακα $\mathbf{\Sigma}$.

4.4 Η κατανομή T^2 του Hotelling

Θεώρημα: Έστω ένα τυχαίο διάνυσμα **x** και ένας τυχαίος πίνακας **M** που ακολουθούν $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και $W_p(m,\boldsymbol{\Sigma})$ αντίστοιχα τότε η ποσότητα $m(x-\mu)'M^{-1}(x-\mu)$ ακολουθεί την κατανομή T^2 του Hotelling με παραμέτρους p και m (συμβολικά $T^2(p,m)$) □

Το παραπάνω θεώρημα έχει μεγάλη σημασία αν θυμηθούμε πως το διάνυσμα των δειγματικών μέσων και ο δειγματικός πίνακας διακύμανσης ικανοποιούν τις συνθήκες του παραπάνω θεωρήματος. Συγκεκριμένα έχουμε πως

$$\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}) \text{ και } n\mathbf{S} = \sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})' \sim W_p(n-1,\boldsymbol{\Sigma})$$

και συνδυάζοντας τα με το παραπάνω θεώρημα προκύπτει πως

$$T^2 = (n-1)(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu) \sim T^2(p,n-1)$$

ή ισοδύναμα αν χρησιμοποιήσουμε τον αμερόληπτο πίνακα διακύμανσης $\mathbf{S}_*$

$$T^2 = n(\bar{x}-\mu)'S_*^{-1}(\bar{x}-\mu)$$

Η κατανομή T^2 συνδέεται με την κατανομή F και επομένως ποσοστιαία σημεία της μπορούν να υπολογιστούν πολύ εύκολα. Συγκεκριμένα για μια τμ που ακολουθεί την κατανομή $T^2(p,m)$ ισχύει πως

$$\frac{m-p+1}{mp}T^2(p,m) \sim F(p,m-p+1)$$

και άρα μπορεί κανείς να χρησιμοποιήσει τους πίνακες της F κατανομής.

Μερικές ενδιαφέρουσες παρατηρήσεις είναι οι εξής:

- Η συνάρτηση T^2 έχει την πολύ χρήσιμη ιδιότητα πως είναι αμετάβλητη (invariant) κάτω από γραμμικούς μετασχηματισμούς, δηλαδή αν μετασχηματίσουμε τα δεδομένα η τιμή της δεν αλλάζει
- Στην ουσία η συνάρτηση T^2 είναι μια γενίκευση του τρόπου που προκύπτει η κατανομή *t* στη μονομεταβλητή περίπτωση. Αν μάλιστα θυμηθούμε πως το τετράγωνο μιας t_n τυχαίας μεταβλητής ακολουθεί $F_{n,1}$ κατανομή η κατανομή T^2 είναι μια γενίκευση της κατανομής *t*.

- Αν p=1 τότε η συνάρτηση T^2 δεν είναι τίποτα άλλο παρά η συνάρτηση $t=\frac{\bar{x}-\mu}{S/\sqrt{n}}$ που τόσο πολύ χρησιμοποιούμε στη μονομεταβλητή περίπτωση.

Το γεγονός πως η ποσότητα

$$\frac{n-p}{(n-1)p}n(\bar{x}-\mu)'S_*^{-1}(\bar{x}-\mu)\sim F(p,n-p)$$

μπορεί να χρησιμοποιηθεί για την κατασκευή διαστημάτων εμπιστοσύνης για τη μέση τιμή του πληθυσμού. Τα διαστήματα εμπιστοσύνης θα είναι υπερελλειψοειδή όπως και στην περίπτωση της γνωστής διακύμανσης που είδαμε προηγούμενα.

Γενικεύοντας η σχέση που το θεώρημα αποδεικνύει μπορεί να χρησιμοποιηθεί για διαστήματα εμπιστοσύνης σε πολλές περιπτώσεις κατά αναλογία των ελλειψοειδών που κατασκευάσαμε όταν ο πίνακας διακύμανσης ήταν γνωστός.

Με τα δεδομένα του θεωρήματος η ποσότητα $m(x-\mu_0)'M^{-1}(x-\mu_0)$ θα ακολουθεί μια μη κεντρική $T^2(p,m,\delta)$ κατανομή όπου δ είναι μια παράμετρος μη κεντρικότητας και δίνεται από τον τύπο

$$\delta=m(\mu-\mu_0)'M^{-1}(\mu-\mu_0).$$

Η μη κεντρική κατανομή T^2 συνδέεται με τη μη κεντρική κατανομή F ως εξής

$$\frac{m-p+1}{mp}T^2(p,m,\delta)\sim F(p,m-p+1,\delta)$$

4.5 Η κατανομή Λάμδα του Wilks

Αν **A,B** δύο ανεξάρτητοι πίνακες, $\mathbf{A}\sim W_p(m,\Sigma)$ και $\mathbf{B}\sim W_p(n,\Sigma)$ $(m\geq p)$, τότε η ποσότητα $\Lambda=\frac{|A|}{|A+B|}$ ακολουθεί την κατανομή Λάμδα του Wilks με παραμέτρους p,m και n (συμβολισμός $\Lambda\sim\Lambda(p,m,n)$).

Η κατανομή Λάμδα είναι μονοδιάστατη. Η συνάρτηση πυκνότητας πιθανότητας της είναι αρκετά πολύπλοκη. Εναλλακτικά προκύπτει ως το γινόμενο τυχαίων μεταβλητών από τη Βήτα κατανομή, δηλαδή

$$\Lambda(p,m,n)=\prod_{i=1}^{n}u_i$$
$$u_i \sim B\big((m+i-p)/2,\ p/2\big),\quad i=1,\ldots,n$$

Παρατηρείστε πως η κατανομή Λ είναι ανεξάρτητη από τον κοινό πίνακα **Σ**.

Ποσοστιαία σημεία της κατανομής έχουν πινακοποιηθεί. Παρόλα αυτά σε κάποιες περιπτώσεις, για συγκεκριμένες τιμές των παραμέτρων, μπορούμε με μετασχηματισμούς να καταλήξουμε σε μια F κατανομή. Συγκεκριμένα ισχύουν τα ακόλουθα

Η κατανομή Λ(p,m,n) είναι ισοδύναμη με την κατανομή Λ(n,m+n-p,p)

Για συγκεκριμένες τιμές των παραμέτρων ισχύουν τα ακόλουθα

$$\left(\frac{m-p+1}{p}\right)\frac{1-\Lambda(p,m,1)}{\Lambda(p,m,1)}\sim F_{p,m-p+1}$$
$$\left(\frac{m}{n}\right)\frac{1-\Lambda(1,m,n)}{\Lambda(1,m,n)}\sim F_{n,m}$$

$$\left(\frac{m-p+1}{p}\right)\frac{1-\sqrt{\Lambda(p,m,2)}}{\sqrt{\Lambda(p,m,2)}}\sim F_{2p,2(m-p+1)}$$
$$\left(\frac{m-1}{n}\right)\frac{1-\sqrt{\Lambda(2,m,n)}}{\sqrt{\Lambda(2,m,n)}}\sim F_{2n,2(m-1)}$$

Μπορεί να δει κανείς πως επειδή η κατανομή προκύπτει ως το γινόμενο βήτα τυχαίων μεταβλητών, ο λογάριθμος μια τυχαίας μεταβλητής από τη Λ κατανομή θα είναι το άθροισμα ανεξάρτητων τυχαίων μεταβλητών από βήτα κατανομές μετά από λογαριθμικό μετασχηματισμό των τυχαίων μεταβλητών. Αυτό διευκολύνει κάπως τον υπολογισμό της συνάρτησης πυκνότητας πιθανότητας σε περιπτώσεις όπου οι παραπάνω τύποι δεν μπορούν να χρησιμοποιηθούν. Μια προσέγγιση για το λογάριθμο μιας Λ τυχαίας μεταβλητής είναι η εξής

$$\left[\frac{p-n+1}{2}-m\right]\ln\Lambda(p,m,n)\sim\chi^2_{np}$$

Θα πρέπει να παρατηρήσουμε πως αυτό το αποτέλεσμα στηρίζεται στο γεγονός πως η κατανομή του Λ προκύπτει ως η κατανομή μιας ελεγχοσυνάρτησης που είναι λόγος πιθανοφανειών και άρα ασυμπτωτικά ακολουθεί τη χ^2 κατανομή.

Ενδιαφέρον είναι και το εξής. Αν p=1 δηλαδή έχουμε απλές τυχαίες μεταβλητές, τότε με βάση όσα είπαμε πριν αυτές θα ακολουθούν χ^2 κατανομή και επομένως ο λόγος που παίρνουμε θα είναι μια F κατανομή εκτός ίσως από μια σταθερά που θα είναι οι βαθμοί ελευθερίας (αφού θα έχουμε μια χ^2 στον αριθμητή και μια χ^2 στον παρονομαστή).

Τέλος πρέπει να σημειωθεί πως η συνάρτηση Λ μπορεί να γραφεί μέσω των ιδιοτιμών του πίνακα $\mathbf{A^{-1}B}$. Αν λοιπόν συμβολίσουμε με λ_i τις ιδιοτιμές του πίνακα αυτού τότε

$$\Lambda = \prod_{i=1}^{p} \frac{1}{1+\lambda_i}$$

Από αυτό τον τύπο διαπιστώνει κανείς εύκολα πως η τιμή του Λ ανήκει αναγκαστικά στο διάστημα (0,1).

Αν $\mathbf{A,B}$ δύο ανεξάρτητοι πίνακες, $\mathbf{A}\sim W_p(m,\mathbf{\Sigma})$ και $\mathbf{B}\sim W_p(n,\mathbf{\Sigma})$ $(m\geq p)$, τότε η μεγαλύτερη ιδιοτιμή θ_1 του πίνακα $\mathbf{(A+B)^{-1}B}$ ακολουθεί την κατανομή Θήτα με παραμέτρους p,m και n (συμβολισμός $\theta_1\sim\Theta(p,m,n)$).

Η κατανομή είναι και πάλι ανεξάρτητη του πίνακα $\mathbf{\Sigma}$. Η συνάρτηση πυκνότητας πιθανότητας είναι αρκετά πολύπλοκη. Και πάλι για συγκεκριμένες τιμές των παραμέτρων η κατανομή μπορεί να μετασχηματιστεί σε μια F κατανομή. Ισχύει ότι

$$\frac{m-p+1}{p}\frac{\Theta(p,m,1)}{1-\Theta(p,m,1)} \sim F(p, m-p+1)$$

Είναι απλό να δει κανείς πως αν έχω p=1 τότε η Θήτα κατανομή ταυτίζεται με την Λ και επομένως και με την F.

5 ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

5.1 Εισαγωγή

Όπως είπαμε και στην αρχή αυτών των σημειώσεων το μεγάλο πλεονέκτημα της πολυμεταβλητής ανάλυσης είναι πως μας επιτρέπει να εξετάσουμε πολλές μεταβλητές συγχρόνως λαμβάνοντας υπόψη μας και τις όποιες συσχετίσεις υπάρχουν μεταξύ τους. Για παράδειγμα έχουμε 5 μεταβλητές οι οποίες μας περιγράφουν κάποιο χαρακτηριστικό. Αν θέλουμε να εξετάσουμε αν οι παρατηρήσεις μας δεν συμφωνούν με κάποια εκ των προτέρων γνώση για τα 5 αυτά χαρακτηριστικά θα μπορούσε κανείς να δουλέψει ξεχωριστά με τις πέντε μεταβλητές, να κάνει δηλαδή έναν έλεγχο για κάθε μεταβλητή και με βάση αυτούς τους ελέγχους να αποφανθεί για το σύνολο.

Αυτή η προσέγγιση έχει δύο μειονεκτήματα:

- αν ο έλεγχος για κάθε χαρακτηριστικό γίνει σε επίπεδο 5%, τότε δεδομένου πως κάνουμε 5 ελέγχους ξεχωριστά το συνολικό επίπεδο του πολύπλοκου (πενταπλού στην περίπτωση μας) ελέγχου δεν θα είναι 5% αλλά $1-0.95^5$ δηλαδή μόλις 22%, δηλαδή η πιθανότητα σφάλματος πολύ μεγάλη.
- επειδή είναι πολύ πιθανό οι μεταβλητές μας να είναι μεταξύ τους συσχετισμένες κάνοντας πέντε ξεχωριστούς ελέγχους αγνοούμε αυτή τη συσχέτιση, δηλαδή αγνοούμε την πληροφορία που έχουμε για κάθε μεταβλητή από τις υπόλοιπες. Επομένως χρειαζόμαστε πολυμεταβλητούς ελέγχους υποθέσεων οι οποίοι θα εξετάζουν όλο το διάνυσμα κάθε παρατήρησης και όχι μεμονωμένες μεταβλητές.

Οι έλεγχοι που θα περιγράψουμε αφορούν παρατηρήσεις από πολυμεταβλητούς κανονικούς πληθυσμούς, αντίστοιχα με την περίπτωση της υπόθεσης της κανονικότητας που είχαμε πίσω από τους περισσότερους μονομεταβλητούς ελέγχους.

5.2 Έλεγχοι για ένα διάνυσμα μέσων τιμών

5.2.1 Γνωστός πίνακας διακύμανσης

Ας ξεκινήσουμε με την πιο απλή περίπτωση του ελέγχου για ένα διάνυσμα μέσων με γνωστή διακύμανση. Στη μονομεταβλητή περίπτωση αυτός είναι ο πιο απλός έλεγχος και γίνεται μέσω της ελεγχοσυνάρτησης *Z* και της κανονικής κατανομής, δηλαδή έχουμε

H_0: μ=μ_0, έναντι της

H_1: μ≠μ_0

Η ελεγχοσυνάρτηση είναι η $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$.

Θυμηθείτε πως για τον έλεγχο υποθέτουμε πως οι μεταβλητές μας είναι κανονικές αλλά αν δεν ισχύει η υπόθεση της κανονικότητας με επίκληση του κεντρικού οριακού θεωρήματος το αποτέλεσμα ισχύει για μεγάλα δείγματα.

Στην πολυμεταβλητή περίπτωση υποθέτουμε πως τα δεδομένα προέρχονται από πολυμεταβλητή κανονική κατανομή. Τώρα έχουμε να ελέγξουμε την υπόθεση

$H_0: \underset{p\times 1}{\boldsymbol{\mu}} = \underset{p\times 1}{\boldsymbol{\mu}_0}$, έναντι της εναλλακτικής

$H_1: \underset{p\times 1}{\boldsymbol{\mu}} \neq \underset{p\times 1}{\boldsymbol{\mu}_0}$

Ο πίνακας διακύμανσης $\boldsymbol{\Sigma}$ είναι γνωστός

Η ελεγχοσυνάρτηση που θα χρησιμοποιήσουμε είναι η

$Z^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$ η οποία με βάση αυτά που είδαμε νωρίτερα ακολουθεί χ^2_p κατανομή. Επομένως ο έλεγχος μπορεί να γίνει είτε υπολογίζοντας το p-value είτε χρησιμοποιώντας την κριτική τιμή της κατανομής αυτής.

Μερικά ενδιαφέροντα σημεία είναι τα εξής

- Η εναλλακτική είναι δίπλευρή καθώς σε ένα διάνυσμα δεν είναι εύκολο να ορίσουμε μονόπλευρους ελέγχους. Παρόλα αυτά μονόπλευροι έλεγχοι θα μπορούσαν να οριστούν χρησιμοποιώντας ελέγχους πιθανοφάνειας.
- Αν p=1 τότε ο έλεγχος ταυτίζεται με τον απλό έλεγχο Z, η ελεγχοσυνάρτηση είναι απλά η ελεγχοσυνάρτηση Z^2 και για αυτό η κατανομή είναι πια χ^2 με έναν βαθμό ελευθερίας.
- Μια μορφή της ελεγχοσυνάρτησης Z^2 είχαμε πριν χρησιμοποιήσει για να δημιουργήσουμε ελλειψοειδή ίσης πιθανότητας. Επομένως μπορούμε να χρησιμοποιήσουμε την

ελεγχοσυνάρτηση για να κατασκευάσουμε από κοινού διαστήματα εμπιστοσύνης για όλο το διάνυσμα των μέσων. Πράγματι θα ισχύει πως

$$P(Z^2 < \chi^2_{1-a,p}) =$$
$$= P(n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0) < \chi^2_{1-a,p}) = 1-a$$

όπου $\chi^2_{1-a,p}$ είναι το (1-α) ποσοστιαίο σημείο της χ^2 κατανομής με *p* βαθμούς ελευθερίας.

Συνεπώς η εξίσωση $n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0) = \chi^2_{1-a,p}$ ορίζει ένα υπερελλειψοειδές το οποίο με πιθανότητα 1-α περιλαμβάνει το άγνωστο διάνυσμα μέσων τιμών του πληθυσμού.

Αν θυμηθούμε τη σχέση δίπλευρων ελέγχων και διαστημάτων εμπιστοσύνης μπορούμε να κάνουμε έλεγχο κατασκευάζοντας το διάστημα εμπιστοσύνης και ελέγχοντας αν η τιμή κάτω από τη μηδενική υπόθεση περιέχεται σε αυτό

5.2.2 Άγνωστος πίνακας διακύμανσης

Ας υποθέσουμε τώρα πως ο πίνακας διακύμανσης δεν είναι γνωστός. Στην μονομεταβλητή περίπτωση χρησιμοποιούσαμε τη δειγματική διακύμανση αντί για τη διακύμανση του πληθυσμού και ο έλεγχος γινόταν με τη βοήθεια της κατανομής t, δηλαδή η ελεγχοσυνάρτηση ήταν η

$$t = \frac{\bar{x}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

Στην πολυμεταβλητή περίπτωση θα χρησιμοποιήσουμε παρόμοια τεχνική. Αντικαθιστούμε τον πίνακα διακύμανσης του πληθυσμού με το δειγματικό πίνακα διακύμανσης και προκύπτει η συνάρτηση του Hotelling

$$T^2 = n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)$$

για την οποία μιλήσαμε πριν και γνωρίζουμε πως ακολουθεί την T^2 κατανομή. Χρησιμοποιώντας τη σχέση της T^2 με την F μπορούμε να χρησιμοποιήσουμε την ελεγχοσυνάρτηση

$$\frac{n}{n-1}\frac{n-p}{p}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0) \sim F_{p,n-p}$$

Και πάλι αν p=1 ο έλεγχος του Hotelling συμπίπτει με τον απλό μονομεταβλητό έλεγχο.

Διαστήματα εμπιστοσύνης μπορούν να κατασκευαστούν και πάλι αλλά τώρα με τη χρήση κριτικών τιμών της κατανομής F. Συγκεκριμένα το υπερελλειψοειδές θα σχηματίζεται από την ανισότητα

$$n(\boldsymbol{\mu}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{x}}) < \frac{p(n-1)}{n-p}F_{a(p,n-p)}$$

Ο έλεγχος που μόλις περιγράψαμε συνήθως ονομάζεται έλεγχος T^2 για μια μέση τιμή (one sample T^2 test).

5.3 Έλεγχος για διαφορά δύο μέσων

Ας δούμε τώρα την περίπτωση που έχουμε δύο δείγματα και άρα θέλουμε να ελέγξουμε αν υπάρχουν διαφορές ανάμεσα στα διανύσματα των μέσων για τις δύο ομάδες. Θα δούμε απευθείας την ρεαλιστική περίπτωση της άγνωστης διακύμανσης. Παράλληλα υποθέτουμε πως η διακύμανση είναι κοινή και επομένως αυτό είναι κάτι που θα πρέπει ουσιαστικά να ελέγξουμε. Θα δούμε αργότερα πως.

Στη μονομεταβλητή περίπτωση ο έλεγχος είχε τις υποθέσεις

H_0: $\mu_1=\mu_2$, έναντι της εναλλακτικής

H_1: $\mu_1\neq\mu_2$

Η ελεγχοσυνάρτηση είναι η

$t=\dfrac{\bar{x}_1-\bar{x}_2}{S_{pooled}\Big/\sqrt{n_1+n_2-2}}\sim t_{n_1+n_2-2}$ που ακολουθεί κάτω από τη μηδενική υπόθεση την κατανομή t.

Η ποσότητα s^2_{pooled} είναι μια εκτίμηση της κοινής διακύμανσης και υπολογίζεται ως σταθμικός μέσος των διακυμάνσεων των δύο ομάδων.

Στην πολυμεταβλητή περίπτωση τα πράγματα είναι ανάλογα.

Οι υποθέσεις που θέλουμε να ελέγξουμε είναι οι

$H_0:\underset{p\times 1}{\boldsymbol{\mu}_1}=\underset{p\times 1}{\boldsymbol{\mu}_2}$, $\boldsymbol{\Sigma}$ κοινός και άγνωστος έναντι της

$H_1:\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$

Θυμηθείτε πως θα ισχύει πως

$$(n_1-1)S_1\sim W_p(n_1-1,\Sigma)$$
$$(n_2-1)S_2\sim W_p(n_2-1,\Sigma)$$

όπου S_i , i=1,2 είναι οι αμερόληπτοι πίνακες διακύμανσης για τα δύο γκρουπ. Επομένως από τις ιδιότητες της κατανομής Wishart

$$(n_1-1)S_1+(n_2-1)S_2\sim W_p(n_1+n_2-2,\Sigma)\,.$$

Επομένως οι προϋποθέσεις για να δημιουργήσουμε μια ανάλογη της T^2 συνάρτησης υπάρχουν (έχουμε ένα διάνυσμα από πολυμεταβλητή κανονική κατανομή και έναν πίνακα από Wishart) και άρα ορίζουμε την ελεγχοσυνάρτηση

$$\frac{n_1+n_2-p-1}{p(n_1+n_2-2)}\frac{n_1 n_2}{n_1+n_2}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)'\mathbf{S}_{pooled}^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)\sim F_{p,n_1+n_2-p-1}$$

με $\mathbf{S}_{pooled}=\frac{(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2}{n_1+n_2-2}$ η οποία έχει μετασχηματιστεί ώστε να ακολουθεί την κατανομή F.

Και σε αυτή την περίπτωση μπορούμε να φτιάξουμε από κοινού διαστήματα εμπιστοσύνης καθώς και να προκύψει η ειδική περίπτωση για p=1.

Τελειώνοντας με τους ελέγχους για μέσες τιμές πρέπει να τονιστεί πως όταν απορριφθεί η μηδενική υπόθεση, μπορεί ο ερευνητής να εστιάσει το ενδιαφέρον του στο ποια(ες) μεταβλητή(ές) είναι υπεύθυνη(ες) για την απόρριψη της υπόθεσης. Δεν θα ασχοληθούμε εδώ με αυτό το πρόβλημα.

Αν θέλουμε να ελέγξουμε την ισότητα περισσοτέρων από δύο δειγμάτων τότε υπάρχει μια αντίστοιχη μεθοδολογία με αυτή της ανάλυσης διακύμανσης την οποία θα δούμε σε λίγο.

5.4 Έλεγχος Ισότητας Πινάκων Διακύμανσης

Στην περίπτωση του ελέγχου για δύο διανύσματα μέσων χρειαστήκαμε την υπόθεση ότι οι πίνακες διακύμανσης στον πληθυσμό είναι ίσοι. Αυτή την υπόθεση μπορούμε να την ελέγξουμε με μια διαδικασία που θα περιγράψουμε στην πιο γενική της μορφή ώστε να μας φανεί χρήσιμη και στην πολυμεταβλητή ανάλυση διακύμανσης που θα ακολουθήσει.

Ας ξεκινήσουμε και πάλι από την περίπτωση μιας μεταβλητής. Έλεγχοι περί ισότητας διακυμάνσεων (συχνά αποκαλούνται έλεγχοι ομοιογένειας διακύμανσης, homogeneity of variance tests) υπάρχουν αρκετοί. Οι πιο γνωστοί είναι ο έλεγχος του Levene και ο έλεγχος του Bartlett και φυσικά για 2 δείγματα ο κλασσικός έλεγχος *F*. Θα περιγράψουμε την πολυμεταβλητή γενίκευση του ελέγχου του Bartlett στις πολλές διαστάσεις. Ο έλεγχος δεν είναι τίποτα διαφορετικό από έλεγχο λόγου πιθανοφανειών.

Συγκεκριμένα οι υποθέσεις που θα εξετάσουμε είναι

H_0: $\mathbf{\Sigma}_1=\mathbf{\Sigma}_2=...=\mathbf{\Sigma}_k$

H_1: Τουλάχιστον 2 διαφέρουν

όπου k ο αριθμός των ομάδων που έχουμε. Ο έλεγχος, που ονομάζεται Box-M, χρησιμοποιεί την ελεγχοσυνάρτηση

$$M = \phi\sum_{i=1}^{k}\left[(n_i-1)\ln\left|\mathbf{S}_i^{-1}\mathbf{S}_{pooled}\right|\right],$$

όπου $\mathbf{S}_{pooled} = \dfrac{\sum_{i=1}^{k}(n_i-1)\mathbf{S}_i}{n-k}$,

$$\phi = 1-\frac{2p^2+3p-1}{6(p+1)(k-1)}\sum_{i=1}^{k}\frac{1}{(n_i-k)(n-k)},$$

$n=\sum_{i=1}^{k} n_i$, και

p είναι ο αριθμός των μεταβλητών (επομένως οι πίνακες διακύμανσης είναι $\mathrm{p}\times\mathrm{p}$).

Ο έλεγχος βασίζεται στο λόγο πιθανοφάνειας κάτω από τις δύο υποθέσεις. Μελετώντας την ελεγχοσυνάρτηση παρατηρούμε πως αν όλοι οι πίνακες διακύμανσης είναι σχεδόν ίσοι, δηλαδή ισχύει η μηδενική υπόθεση, τότε και ο $\mathbf{S}_{pooled}$ είναι σχεδόν ίσος με αυτούς και επομένως ο πίνακας $\mathbf{S}_i^{-1}\mathbf{S}_{pooled}$ θα μοιάζει πολύ με τον μοναδιαίο πίνακα. Συνεπώς η ορίζουσα του θα είναι κοντά στο 1, αυτό με τη σειρά του σημαίνει πως ο λογάριθμος θα είναι σχεδόν 0 και άρα η τιμή της ελεγχοσυνάρτησης θα είναι πολύ μικρή.

Δεδομένου πως πρόκειται για έλεγχο πιθανοφανειών η κατανομή της ελεγχοσυνάρτησης κάτω από τη μηδενική υπόθεση και ασυμπτωτικά είναι η $\chi^2_{p(p+1)(k-1)/2}$. Η ποσότητα φ έχει σκοπό να κάνει την ασυμπτωτική προσέγγιση πιο καλή για πεπερασμένα δείγματα.

Παράδειγμα 5.1: Σε ένα δείγμα 15 μαθητών μιας τάξης μετρήθηκαν το ύψος (X_1), η περιφέρεια στήθους (X_2), η περιφέρεια βραχίονα (X_3), και καταγράφηκε και το φύλο (0 αγόρι, 1 κορίτσι). Τα δεδομένα μπορείτε να τα δείτε στον πίνακα που ακολουθεί.

Παρατηρείστε πως με την παραπάνω παρουσίαση των δεδομένων που είναι αυτή που όλα τα στατιστικά πακέτα ακολουθούν κάθε παρατήρηση είναι μια γραμμή. Όμως στη θεωρία χρησιμοποιήσαμε ως παρατήρηση το διάνυσμα στήλη. Επομένως ο παραπάνω πίνακας περιέχει τις παρατηρήσεις μας ως διανύσματα στήλη, δηλαδή ισχύει (λαμβάνοντας τις 3 πρώτες μεταβλητές υπόψη) πως $x_1'=[78,\ 60.6,\ 16.5]$ κλπ

Ύψος	**Περιφέρεια Στήθους**	**Περιφέρεια Βραχίονα**	**Φύλο**
78	60.6	16.5	0
76	58.1	12.5	0
92	63.2	14.5	0
1	59	14	0
81	60.8	15.5	0
84	59.5	14	0
80	58.4	14	1
75	59.2	15	1
78	60.3	15	1
75	57.4	13	1
79	59.5	14	1
78	58.1	14.5	1
75	58	12.5	1
64	55.5	11	1
80	59.2	12.5	1

Πίνακας 5.1 Δεδομένα για το παράδειγμα 5.1 (Πηγή: Chatfield and Colling, 1984.)

Έστω ότι θέλουμε να ελέγξουμε την υπόθεση πως το διάνυσμα των μέσων για τα αγόρια είναι το $\begin{bmatrix} 90 \\ 58 \\ 16 \end{bmatrix}$ δοθέντος πως ο πίνακας διακύμανσης είναι γνωστός και ίσος με $\boldsymbol{\Sigma} = \begin{bmatrix} 30 & 6 & 0.5 \\ 6 & 3 & 1.5 \\ 0.5 & 1.5 & 2 \end{bmatrix}$.

Συνεπώς οι υποθέσεις που μας ενδιαφέρουν είναι οι

H_0: $\boldsymbol{\mu}' = [90 \quad 58 \quad 16]$ έναντι της

H_1: $\boldsymbol{\mu}' \neq [90 \quad 58 \quad 16]$

Για τα αγόρια έχουμε 6 παρατηρήσεις (n=6). Και επειδή ο πίνακας διακύμανσης είναι γνωστός θα κάνουμε έλεγχο για γνωστό πίνακα διακύμανσης. Το διάνυσμα των μέσων είναι $\bar{\mathbf{x}} = \begin{bmatrix} 82 \\ 60.2 \\ 14.5 \end{bmatrix}$ και η ελεγχοσυνάρτηση που θα χρησιμοποιήσουμε είναι η

$$Z^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$$

Ο αντίστροφος του πίνακα διακύμανσης που χρειαζόμαστε είναι ο

$$\boldsymbol{\Sigma}^{-1}=\begin{bmatrix}0.07692 & & \\ -0.23077 & 1.22564 & \\ 0.15385 & -0.86154 & 1.170769\end{bmatrix},$$

και το διάνυσμα με τις αποκλίσεις των δειγματικών μέσων από αυτές της μηδενικής υπόθεσης είναι

$$\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right)=\begin{bmatrix}-8\\ 2.2\\ -1.5\end{bmatrix}.$$

Συνεπώς η τιμή της ελεγχοσυνάρτησης υπολογίζεται ως

$$Z^2=6\begin{bmatrix}-8 & 2.2 & -1.5\end{bmatrix}\begin{bmatrix}0.07692 & -0.23077 & 0.15385\\ -0.23077 & 1.22564 & -0.86154\\ 0.15385 & -0.86154 & 1.17077\end{bmatrix}\begin{bmatrix}-8\\ 2.2\\ -1.5\end{bmatrix}=185.094,$$

Δεδομένου πως p=3, η κατανομή της ελεγχοσυνάρτησης είναι η χ^2 με 3 βαθμούς ελευθερίας και βρίσκουμε πως p-value<0.0001, δηλαδή απορρίπτουμε τη μηδενική υπόθεση σχεδόν σε κάθε επίπεδο στατιστικής σημαντικότητας.

Ας υποθέσουμε τώρα πως η διακύμανση είναι άγνωστη. Σε αυτή την περίπτωση θα πρέπει να χρησιμοποιήσουμε το δειγματικό πίνακα διακύμανσης.

Η ελεγχοσυνάρτηση θα είναι η T^2 του Hotelling δηλαδή η

$$\frac{n}{n-1}\frac{n-p}{p}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right)$$

Ο δειγματικός πίνακας διακύμανσης και ο αντίστροφός του είναι οι

$$\mathbf{S}=\begin{bmatrix}31.6 & 8.04 & 0.5\\ 8.04 & 3.172 & 1.31\\ 0.5 & 1.31 & 1.9\end{bmatrix},\quad \mathbf{S}^{-1}=\begin{bmatrix}0.1836 & -0.63 & 0.38\\ & 2.584 & -1.615\\ & & 1.53\end{bmatrix}$$

και επομένως υπολογίζουμε την ελεγχοσυνάρτηση ως

$$\frac{6}{5}\frac{3}{3}\begin{bmatrix}-8 & 2.2 & -1.5\end{bmatrix}\mathbf{S}^{-1}\begin{bmatrix}-8\\ 2.2\\ -1.5\end{bmatrix}=84.09$$

Επειδή η ελεγχοσυνάρτηση ακολουθεί την κατανομή F με βαθμούς ελευθερίας 3 και 3 βρίσκουμε πως p-value=0.00003 και άρα απορρίπτουμε σχεδόν σε κάθε επίπεδο στατιστικής σημαντικότητας τη μηδενική υπόθεση.

Έστω πως τώρα θέλουμε να ελέγξουμε την υπόθεση πως τα διανύσματα των μέσων θα είναι τα ίδια για τα αγόρια και τα κορίτσια. Δηλαδή η μηδενική μας υπόθεση είναι πως

H_0: $\boldsymbol{\mu}_A=\boldsymbol{\mu}_K$ έναντι της εναλλακτικής

H_1: $\boldsymbol{\mu}_A \neq \boldsymbol{\mu}_K$

Ο πίνακας διακύμανσης Σ υποθέτουμε πως είναι άγνωστος αλλά κοινός. Θα ελέγξουμε αργότερα αν η υπόθεση για ίσες διακυμάνσεις είναι λογική.

Ελεγχοσυνάρτηση $\frac{n_1+n_2-p-1}{p(n_1+n_2-2)}\frac{n_1 n_2}{n_1+n_2}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)' \mathbf{S}_{pooled}^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)$

Είναι $\bar{\mathbf{x}}_1=\begin{bmatrix}82\\60.2\\14.5\end{bmatrix}$, $\bar{\mathbf{x}}_2=\begin{bmatrix}76\\58.4\\13.5\end{bmatrix}$

Με τη βοήθεια των πινάκων $\mathbf{S}_1$ και $\mathbf{S}_2$ θα υπολογίσουμε

$$\mathbf{S}_{pooled}=\frac{(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2}{n_1+n_2-2}=\begin{bmatrix}27.23 & 6.56 & 2.84\\ 6.56 & 2.43 & 1.4\\ 2.84 & 1.4 & 1.84\end{bmatrix}$$

Άρα η τιμή της ελεγχοσυνάρτησης είναι $1.489 < F_{0.95,3,11}$, ενώ p-value=0.24 και συνεπώς δεν απορρίπτουμε H_0.

Όπως είδαμε προηγουμένως, για να κάνουμε τον έλεγχο δύο διανυσμάτων μέσων τιμών χρειαζόμαστε την υπόθεση ισότητας των πινάκων διακυμάνσεων. Θα χρησιμοποιήσουμε τον έλεγχο Box-M.

Συγκεκριμένα θα ελέγξουμε την υπόθεση

H_0: $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$ έναντι της εναλλακτικής

H_1: $\boldsymbol{\Sigma}_1\neq\boldsymbol{\Sigma}_2$

Για τα δεδομένα μας έχουμε πως p=3, k=2, n=15, n_1=6, n_2=9 ενώ βρίσκουμε

$$\mathbf{S}_1=\begin{bmatrix}31.6 & 8.04 & 0.5\\ 8.04 & 3.17 & 1.31\\ 0.5 & 1.31 & 1.9\end{bmatrix}, \mathbf{S}_2=\begin{bmatrix}24.5 & 5.63 & 4.31\\ 5.63 & 1.97 & 1.45\\ 4.31 & 1.45 & 1.31\end{bmatrix}$$

και άρα

$$\mathbf{S}_{pooled}=\frac{5\mathbf{S}_1+8\mathbf{S}_2}{13}=\begin{bmatrix}27.23 & 6.56 & 2.84\\ 6.56 & 2.43 & 1.4\\ 2.84 & 1.4 & 1.84\end{bmatrix}, \left|\mathbf{S}_1^{-1}\mathbf{S}_{pooled}\right|=0.95, \left|\mathbf{S}_2^{-1}\mathbf{S}_{pooled}\right|=1.82$$

$$\phi = 1 - \frac{18+9-1}{24}\left(\frac{1}{5\cdot 13}+\frac{1}{8\cdot 13}\right) = 0.963$$

Χρησιμοποιώντας λοιπόν τον τύπο για την ελεγχοσυνάρτηση του ελέγχου βρίσκουμε πως $M = 0.963\cdot(5\ln 0.95 + 8\ln 1.82) = 4.37 < \chi_{6,0.95}$, ενώ το p-value=0.63 οπότε και δεν απορρίπτουμε την H_0.

5.5 Συμπεράσματα

Στον παρακάτω πίνακα εμφανίζουμε συγκριτικά τους ελέγχους για τη μονοδιάστατη και την πολυδιάστατη περίπτωση για να γίνει σαφής η σχέση τους

Μονοδιάστατη περίπτωση	**Πολυδιάστατη περίπτωση**
$Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	$n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0) \sim \chi^2_p$
$t = \frac{\bar{x}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$	$\frac{n}{n-1}\frac{n-p}{p}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)' \mathbf{S}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0) \sim F_{p,n-p}$
$t = \frac{\bar{x}_1-\bar{x}_2}{S_{pooled}/\sqrt{n_1+n_2-2}} \sim t_{n_1+n_2-2}$	$\frac{n_1+n_2-p-1}{p(n_1+n_2-2)}\frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)' \mathbf{S}^{-1}_{pooled}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \sim F_{p,n_1+n_2-p-1}$
$M = \phi^* \sum_{i=1}^{k}\left[(n_i-1)\ln\left(\frac{s^2_{pooled}}{s_i^2}\right)\right]$	$M = \phi \sum_{i=1}^{k}\left[(n_i-1)\ln\left\|\mathbf{S}_i^{-1}\mathbf{S}_{pooled}\right\|\right]$

Ο μονομεταβλητός έλεγχος ομοιογένειας διακυμάνσεων είναι ο έλεγχος του Bartlett.Η σταθερά φ^* υπολογίζεται με πολύ παρόμοιο τρόπο όπως και η σταθερά φ της πολυμεταβλητής περίπτωσης

Ολοκληρώνοντας σχετικά με τους ελέγχους υποθέσεων θα πρέπει να αναφέρουμε τα εξής:

- Όπως και στη μονομεταβλητή περίπτωση οι έλεγχοι είναι βασισμένοι στην υπόθεση της πολυμεταβλητής κανονικότητας. Επειδή και για τα διανύσματα μέσων υπάρχουν αντίστοιχα κεντρικά οριακά θεωρήματα τα αποτελέσματα μπορούν να χρησιμοποιηθούν αν το μέγεθος του δείγματος είναι σχετικά μεγάλο. Μόνο ο έλεγχος Box-M έχει πρόβλημα και πολύ κακές ιδιότητες αν η υπόθεση της κανονικότητας δεν ισχύει

- Αν στον έλεγχο για μια μέση τιμή (ή για διαφορά δύο μέσων τιμών) απορρίψουμε τη μηδενική υπόθεση, το ενδιαφέρον είναι να βρούμε πια ποια ή ποιες μεταβλητές διαφέρουν και γενικά οδηγούν στην απόρριψη της μηδενικής υπόθεσης. Υπάρχουν στη βιβλιογραφία τέτοιες τεχνικές αλλά δεν θα ασχοληθούμε μαζί τους σε αυτές τις σημειώσεις
- Προφανώς οι έλεγχοι που συζητήσαμε είναι απλά κάποια δείγματα για το πώς ιδέες πολύ γνωστές στη μονοδιάστατη περίπτωση μεταφέρονται ή αν θέλετε γενικεύονται σε περισσότερες διαστάσεις. Υπάρχει μια τεράστια ποικιλία από ελέγχους υποθέσεων για πολυμεταβλητά δεδομένα.

5.6 Έλεγχοι για την πολυμεταβλητή κανονική κατανομή

Ένα πρόβλημα το οποίο εμφανίζεται πολύ συχνά, αλλά δυστυχώς δεν έχει εύκολη και ξεκάθαρη λύση είναι το εξής: πως μπορούμε να ελέγξουμε αν τα δεδομένα μας προέρχονται από μια πολυμεταβλητή κανονική κατανομή. Αυτή η υπόθεση είναι βασική σε αρκετές διαδικασίες στατιστικής συμπερασματολογίας.

Στη μονομεταβλητή περίπτωση για παράδειγμα υπάρχει μια ποικιλία ελέγχων για να ελέγξει κανείς την κανονικότητα των δεδομένων. Αυτό που χρειαζόμαστε είναι μια γενίκευση της ιδέας σε μεγαλύτερες διαστάσεις.

Δυστυχώς τα πράγματα δεν είναι τόσο εύκολα στην πολυμεταβλητή περίπτωση. Στην πράξη οι έλεγχοι για την κανονικότητα βασίζονται είτε σε κάποιες χαρακτηριστικές ιδιότητες της κανονικής κατανομής είτε (οι περισσότερο γνωστοί) στην εμπειρική συνάρτηση κατανομής . Στην πολυμεταβλητή περίπτωση δεν είναι εύκολο να ορίσει κανείς την εμπειρική συνάρτηση κατανομής. Αυτό αποτελεί επομένως ένα πρόσθετο πρόβλημα εκτός από την πολυπλοκότητα των πολλών διαστάσεων των δεδομένων.

Στην πράξη υπάρχουν πολλοί έλεγχοι για την πολυμεταβλητή κανονικότητα των δεδομένων. Οι περισσότεροι από αυτούς είναι ιδιαίτερα πολύπλοκοι και για αυτό δεν θα περιγραφούν σε αυτές τις σημειώσεις. Θα περιγράψουμε απλά κάποιες διαδικασίες περισσότερο διαγνωστικού χαρακτήρα.

Ένας πρώτος τρόπος είναι να δούμε τι γίνεται με τις περιθωριακές κατανομές. Ξέρουμε πως αν τα δεδομένα προέρχονται από πολυμεταβλητή κανονική τότε και κάθε μια μεταβλητή θα ακολουθεί από μόνη της μια κανονική κατανομή. Προσοχή, το αντίστροφο δεν ισχύει δηλαδή αν όλες οι περιθώριες κατανομές είναι κανονικές αυτό δεν μας αποδεικνύει πως η από κοινού είναι πολυμεταβλητή και αυτή κανονική.

Επομένως μια πρώτη καθαρά διαγνωστική μέθοδος είναι να δούμε αν όλες οι περιθώριες κατανομές είναι κανονικές. Αν δεν είναι έχουμε αρκετά στοιχεία να πιστεύουμε πως και όλες μαζί δεν ακολουθούν πολυμεταβλητή κανονική. Αν όμως δεν απορρίψουμε την κανονικότητα για καμία μεταβλητή τότε δεν μπορούμε με σιγουριά να πούμε ότι η από κοινού κατανομή είναι πολυμεταβλητή κανονική.

Παρατηρείστε επίσης πως αν έχουμε πολλές μεταβλητές, δηλαδή η διάσταση της πολυμεταβλητής κανονικής είναι μεγάλη, τότε έχουμε το γνωστό πρόβλημα των πολλαπλών ελέγχων υποθέσεων και έτσι η πιθανότητα σφάλματος τύπου Ι είναι αρκετά μεγάλη. Για αυτό το λόγο αυτή η προσέγγιση καλό είναι να αποφεύγεται.

Για το παράδειγμα με τις 20 αμερικάνικες πόλεις, κάνοντας τον έλεγχο κανονικότητας των Anderson-Darling βρήκαμε τα εξής p-value για τις 5 μεταβλητές: 0.400, 0.633, 0.586, 0.525, 0.165. Δηλαδή σε επίπεδο στατιστικής σημαντικότητας 5% δεν απορρίπτουμε την υπόθεση της κανονικότητας για καμιά μεταβλητή. Δυστυχώς αυτό δεν είναι αρκετό για να μας δείξει και την πολυμεταβλητή κανονικότητα των δεδομένων μας.

Παράδειγμα: Έστω ότι η από κοινού κατανομή των *x,y,* δίνεται από την

$$f(x,y)=\frac{1}{2}f(x,y\mid\rho_1)+\frac{1}{2}f(x,y\mid\rho_2)$$

όπου

$$f(x,y\mid\rho_1)=\frac{1}{2\pi(1-\rho_1^2)^{1/2}}\exp\left\{-\frac{1}{2(1-\rho_1^2)}\left(x^2+y^2-2\rho_1 xy\right)\right\} \quad \text{και}$$

$$f(x,y\mid\rho_2)=\frac{1}{2\pi(1-\rho_2^2)^{1/2}}\exp\left\{-\frac{1}{2(1-\rho_2^2)}\left(x^2+y^2-2\rho_2 xy\right)\right\}$$

δηλαδή διμεταβλητές κανονικές κατανομές με μέσους μηδέν, διακυμάνσεις 1 και διαφορετικές συσχετίσεις. Μπορεί κανείς εύκολα να επαληθεύσει πως οι περιθώριες κατανομές είναι κανονικές αλλά η από κοινού ξεκάθαρα δεν είναι διμεταβλητή κανονική.

Ο δεύτερος έλεγχος βασίζεται στην ιδιότητα που είδαμε σχετικά με την κατανομή T^2 του Hotelling. Συγκεκριμένα επειδή οι παρατηρήσεις προέρχονται από $N_p(\mu,\Sigma)$ και ο δειγματικός πίνακας διακύμανσης ακολουθεί μια Wishart κατανομή μπορεί κανείς να δει πως η ποσότητα

$$\frac{n-p}{np}(x_i-\mu)'S^{-1}(x_i-\mu)\sim F(p,n-p)$$

Συνεπώς το πολυμεταβλητό πρόβλημα έχει μετασχηματισθεί σε ένα πρόβλημα να ελέγξουμε αν οι παρατηρήσεις μας προέρχονται από την κατανομή F. Εδώ θα πρέπει να σημειώσουμε τα εξής:

- Επειδή δεν γνωρίζουμε το διάνυσμα των μέσων χρησιμοποιούμε το διάνυσμα των δειγματικών μέσων. Αυτό οδηγεί σε κάποια απόκλιση από το γενικό αποτέλεσμα, η κατανομή δεν είναι πια F, αλλά, ευτυχώς, πολύ όμοια με την F.
- Μπορεί η παραπάνω σχέση να ισχύει για κάθε παρατήρηση αλλά οι μετασχηματισμένες τιμές δεν είναι ανεξάρτητες. Συνεπώς δεν έχουμε ένα ανεξάρτητο δείγμα από παρατηρήσεις.
- Ακόμα και για ανεξάρτητα δείγματα έλεγχοι για την καλή προσαρμογή της κατανομής F δεν είναι τόσο γνωστοί. Ουσιαστικά κάποιος μπορεί να καταφύγει σε ελέγχους Monte Carlo, είτε απλά σε κάποιο P-P plot για να δει αν οι μετασχηματισμένες τιμές ακολουθούν την κατανομή F

Συνεπώς ο γραφικός τρόπος ελέγχου της πολυμεταβλητής κανονικότητας είναι ο εξής:

Έστω n το μέγεθος του δείγματος και p ο αριθμός των μεταβλητών

- Για κάθε παρατήρηση δημιούργησε την τιμή

$$r_i = \frac{n-p}{np}(x_i - \bar{x})' S^{-1}(x_i - \bar{x}), \; i=1,2,\ldots,n$$

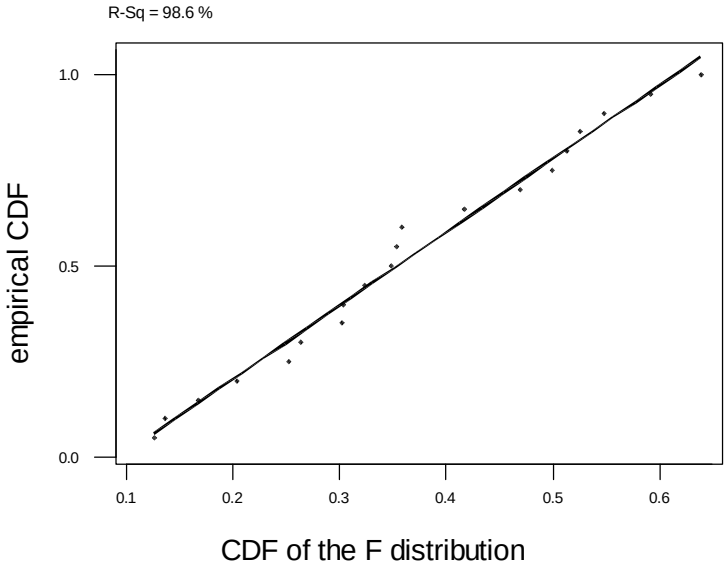
- Φτιάξε ένα P-P plot (ή εναλλακτικά ένα Q-Q plot) για τα r_i χρησιμοποιώντας την κατανομή F με βαθμούς ελευθερίας p και $n-p$. Για να γίνει αυτό αρκεί να βρεθεί η εμπειρική συνάρτηση κατανομής των r_i και οι αντίστοιχες θεωρητικές τιμές της κατανομής F
- Αν τα σημεία είναι κοντά σε μια νοητή ευθεία στη διαγώνιο του γραφήματος (με κάποιες μικροαποκλίσεις φυσικά) τότε υπάρχουν ισχυρές ενδείξεις πως τα δεδομένα ακολουθούν πολυμεταβλητή κανονική κατανομή

Θα πρέπει να σημειωθούν τα εξής:

- Η μέθοδος αυτή δεν δουλεύει καλά για μικρά μεγέθη δείγματος (κυρίως γιατί τότε η εξάρτηση ανάμεσα στις τιμές r_i είναι υψηλή). Επίσης δεν δουλεύει καλά για μεγάλες τιμές του p .
- Επειδή η κατανομή F όταν οι βαθμοί ελευθερίας του παρονομαστή είναι πολύ μεγάλος αριθμός προσεγγίζεται καλά από την κατανομή χ^2 , αυτό σημαίνει πως για μεγάλα δείγματα μπορεί να χρησιμοποιηθεί η κατανομή χ^2 με p βαθμούς ελευθερίας.

Παράδειγμα Ας δούμε πάλι το παράδειγμα των αμερικάνικων πόλεων. Θέλουμε να ελέγξουμε αν τα δεδομένα προέρχονται από πολυμεταβλητή κανονική κατανομή. Στην περίπτωση μας $p=5$, $n=20$. Βρίσκουμε τις τιμές και κατασκευάζουμε το P-P plot που βλέπετε στο γράφημα 5.1

Μπορεί να δει κανείς πως τα σημεία είναι πολύ κοντά στη διαγώνιο και επομένως από το γράφημα μοιάζει λογικό να υποθέσουμε την πολυμεταβλητή κανονικότητα των δεδομένων μας



Γράφημα 5.1. P-P plot για τα μετασχηματισμένα δεδομένα. Η διαγώνιος φαίνεται να προσαρμόζει καλά στα σημεία και αυτό αποτελεί μια καλή ένδειξη πως η υπόθεση της πολυμεταβλητής κανονικότητας είναι λογική

6 ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

6.1 Εισαγωγή

Η μέθοδος της Πολυμεταβλητής Ανάλυσης Διακύμανσης (Multivariate Analysis of Variance, MANOVA) αποτελεί γενίκευση της μονομεταβλητής ανάλυσης διακύμανσης όταν εξετάζουμε συγχρόνως περισσότερες από μια μεταβλητές. Αποτελεί επομένως μια μέθοδο ελέγχου του αν οι μέσοι δύο ή περισσοτέρων ομάδων διαφέρουν και γενικεύοντας σε περιπτώσεις πολλών παραγόντων αν οι παράγοντες αυτοί επιδρούν στη μέση τιμή (μιλάμε πια για διάνυσμα μέσων τιμών). Όπως θα δούμε και στη συνέχεια πολλά πράγματα που ισχύουν στη μονομεταβλητή περίπτωση μεταφέρονται με ανάλογο τρόπο και στην πολυμεταβλητή περίπτωση, όπως για παράδειγμα η διάσπαση της συνολικής διακύμανσης στην μεταξύ των γκρουπ και εντός των γκρουπ διακύμανση (between και within). Επίσης η MANOVA μπορεί να ειδωθεί σαν ένα πολυμεταβλητό γραμμικό μοντέλο. Στις σημειώσεις αυτές θα γίνει προσπάθεια να περιγραφεί η μέθοδος σε ένα επίπεδο εφαρμογών και όχι θεωρίας, αποφεύγοντας πολύπλοκες μαθηματικές αποδείξεις.

6.2 MANOVA ως προς έναν Παράγοντα

Θα αρχίσουμε εξετάζοντας την απλούστερη περίπτωση, ανάλυσης με έναν παράγοντα (one way).

Έστω ότι έχουμε ένα δείγμα $X_{i,j}$, j=1,…,k, και i=1,…,n_j, όπου γενικά j συμβολίζει το γκρουπ και i την παρατήρηση μέσα στο γκρουπ. Η παρατήρηση $X_{i,j}$, είναι πια ένα διάνυσμα με p στοιχεία και όχι μια απλή τιμή. Παρατηρείστε ότι το μέγεθος του δείγματος δεν είναι αναγκαστικά το ίδιο για όλα τα γκρουπ. Υποθέτουμε ότι οι πληθυσμοί για κάθε γκρουπ είναι $N_p(\mu_j, \Sigma)$, όπου p είναι η διάσταση (μιλάμε για μια πολυμεταβλητή κανονική κατανομή p

διαστάσεων), και μ και Σ είναι το διάνυσμα των μέσων και ο πίνακας διακύμανσης συνδιακύμανσης αντίστοιχα. Υποθέτουμε επίσης πως ο πίνακας διακύμανσης συνδιακύμανσης είναι ο ίδιος για όλα τα γκρουπ. Επομένως οι δύο υποθέσεις που χρειάζεται να κάνουμε στην MANOVA είναι

- Κανονικότητα των δεδομένων
- Ίδιος πίνακας διακύμανσης συνδιακύμανσης για όλα τα γκρουπ.

Η μηδενική υπόθεση που θέλουμε να ελέγξουμε είναι

H_0: $\mu_1 = \mu_2 = ... = \mu_k$ έναντι της εναλλακτικής

H_1: τουλάχιστον δύο μέσοι διαφέρουν

Προχωρώντας με το συνηθισμένο τρόπο κατασκευής ελέγχων με τη μέθοδο του λόγου πιθανοφανειών μπορούμε να βρούμε την πιθανοφάνεια κάτω από τις δύο υποθέσεις. Κάτω από τη μηδενική υπόθεση ο λογάριθμος της πιθανοφάνειας των δεδομένων είναι

$$-\frac{nkp}{2}\ln(2\pi) - \frac{nkp}{2}\ln|\Sigma| - \frac{1}{2}\sum_j\sum_i (x_{ij}-\mu_j)'\Sigma^{-1}(x_{ij}-\mu_j)$$

το οποίο αγνοώντας τον πρώτο όρο γράφεται ως

$$-\frac{nkp}{2}\ln|\Sigma| - \frac{1}{2}trace\Sigma^{-1}\left[\sum_j\sum_i (x_{ij}-\bar{x}_{\bullet j})(x_{ij}-\bar{x}_{\bullet j})' + n\sum_j (\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet})(\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet})'\right].$$

Παρατηρείστε ότι τώρα μπορεί να γραφτεί στη μορφή

$-\frac{nkp}{2}\ln|\Sigma| - \frac{1}{2}trace\Sigma^{-1}[B+W]$ όπου

$$W = \sum_j\sum_i (x_{ij}-\bar{x}_{\bullet j})(x_{ij}-\bar{x}_{\bullet j})'$$

$$B = n\sum_j (\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet})(\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet})'.$$

Όπως και στη μονομεταβλητή περίπτωση $\bar{x}_{\bullet j} = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$, $\bar{x}_{\bullet\bullet} = \frac{\sum_j\sum_i x_{ij}}{n}$, δηλαδή ο μέσος του j γκρουπ και ο γενικός μέσος αντίστοιχα.

Δηλαδή παρατηρούμε ότι σπάσαμε σε δύο όρους τις τετραγωνικές αποκλίσεις όπως κάναμε και στην μονομεταβλητή περίπτωση. Ο πρώτος όρος που τον αποτελεί ο πίνακας W μας δίνει τις αποκλίσεις μέσα στο κάθε γκρουπ και αντιστοιχεί στο within sum of squares. Ο

δεύτερος όρος μας δίνει τις αποκλίσεις ανάμεσα στους μέσους των γκρουπ από τον γενικό μέσο και αντιστοιχεί στο between sum of squares. Θα πρέπει να σημειωθεί πως τα στοιχεία της διαγωνίου είναι τα γνωστά από τη μονομεταβλητή ανάλυση διακύμανσης αθροίσματα τετραγώνων που χρησιμοποιούνται για τους μονομεταβλητούς ελέγχους.

Δουλεύοντας ομοίως κάτω από την εναλλακτική υπόθεση βρίσκουμε πως ο λογάριθμος της πιθανοφάνειας είναι

$$-\frac{nkp}{2}\ln|\Sigma|-\frac{1}{2}tr\Sigma^{-1}[W]$$

και αυτό γιατί διασπώντας τις συνολικές τετραγωνικές αποκλίσεις παίρνουμε πως ο λογάριθμος της πιθανοφάνειας είναι

$$-\frac{nkp}{2}\ln|\Sigma|-\frac{1}{2}trace\Sigma^{-1}\left[\sum_j\sum_i(x_{ij}-\bar{x}_{\bullet j})(x_{ij}-\bar{x}_{\bullet j})'+n\sum_j(\bar{x}_{\bullet j}-\bar{x}_{\bullet j})(\bar{x}_{\bullet j}-\bar{x}_{\bullet j})'\right].$$

Ο δεύτερος όρος μέσα στην αγκύλη όμως είναι 0 οπότε παίρνουμε το αποτέλεσμα που θέλουμε.

Παρατήρηση: Στην πρώτη περίπτωση επειδή υποθέτουμε πως όλοι οι μέσοι είναι ίδιοι ο εκτιμητής μεγίστης πιθανοφάνειας του κοινού μέσου είναι ο μέσος όλου του δείγματος. Στην δεύτερη περίπτωση όμως, που τέτοια υπόθεση δεν κάνουμε, χρησιμοποιούμαι τον εκτιμητή για κάθε γκρουπ και έτσι οδηγούμαστε στο αποτέλεσμα που είδαμε.

Έτσι οι μέγιστες τιμές των λογαριθμοποιημένων πιθανοφανειών είναι

$$-\frac{nk}{2}\ln|B+W|-\frac{np}{2} \qquad \text{και} \qquad -\frac{nk}{2}\ln|W|-\frac{np}{2}$$

για τη μηδενική και την εναλλακτική υπόθεση αντίστοιχα, οπότε καταλήγουμε πως ο λόγος των πιθανοφανειών LR είναι

$$LR=kn\ln\frac{|W|}{|B+W|}.$$

Είναι επομένως λογικό να βασίσουμε τη συμπερασματολογία μας στο λόγο των δύο οριζουσών και συγκεκριμένα στην ποσότητα $\Lambda=\dfrac{|W|}{|B+W|}$.

6.3 Έλεγχοι Υποθέσεων

Αν συμβολίσουμε με T=B+W και θυμηθούμε πως η ορίζουσα ενός πίνακας διακύμανσης συνδιακύμανσης είναι ένα μέτρο της γενικής μεταβλητότητας μπορούμε να δούμε πως η ποσότητα αυτή είναι παρόμοια με το λόγο F που χρησιμοποιούμε στη μονομεταβλητή περίπτωση.

Βασισμένοι στην υπόθεση ότι τα δεδομένα μας προέρχονται από μια πολυμεταβλητή κανονική κατανομή και χρησιμοποιώντας τις ιδιότητες της κατανομής Wishart και κυρίως του τρόπου που αυτή προκύπτει από την πολυμεταβλητή κανονική κατανομή, έχει δειχτεί ότι η στατιστική συνάρτηση Λ ακολουθεί την κατανομή Λ-Wilks. Πιο συγκεκριμένα μπορεί να δειχτεί ότι

$$W \sim W_p(\Sigma, n-k)$$
$$B \sim W_p(\Sigma, k-1)$$

Αν οι πίνακες B,W είναι ανεξάρτητοι (κάτι που ισχύει στην περίπτωση μας) τότε η συνάρτηση Λ ακολουθεί την κατανομή $\Lambda(p, n-k, k-1)$, χρησιμοποιώντας τα αποτελέσματα που είδαμε στο κεφάλαιο 4. συνεπώς ο έλεγχος θα στηριχτεί στην κατανομή Λ του Wilks και τις σχέσεις που τη συνδέουν με γνωστές κατανομές.

Στην πράξη, απορρίπτουμε την μηδενική υπόθεση όταν η τιμή του Λ είναι κοντά στο 0. Αυτό είναι λογικό, αν αναλογισθεί κανείς ότι μικρή τιμή Λ συνεπάγεται για τον πίνακα **W** των 'within διαφορών' ότι αυτές είναι μικρές σε σχέση με τις 'between διαφορές' του πίνακα **B** και άρα οι διαφορές ανάμεσα στις ομάδες είναι μεγάλες.

Παρατηρείστε ότι

$$\Lambda = \frac{1}{\left|W^{-1}\right|\left|B+W\right|} = \frac{1}{\left|I+W^{-1}B\right|} = \prod_{i=1}^{p}\frac{1}{\lambda_i}$$

όπου λ_i είναι οι ιδιοτιμές του πίνακα $I+W^{-1}B$.

Όμως

$$\left|I+W^{-1}B-\lambda I\right| = 0 \Rightarrow \left|W^{-1}B-(\lambda-1)I\right| = 0 \Rightarrow$$
$$\Rightarrow l = (\lambda-1) \Rightarrow \lambda = l+1$$

όπου l_i είναι οι ιδιοτιμές του πίνακα $W^{-1}B$. Συνεπώς μπορούμε να γράψουμε πως

$\Lambda = \prod_{i=1}^{p}\frac{1}{l_i+1}$ και να βασιστούμε στις ιδιοτιμές αυτές για την κατασκευή εναλλακτικών ελέγχων.

Άλλοι έλεγχοι που έχουν προταθεί στη βιβλιογραφία είναι

1. Έλεγχος Roy. Θεωρούμε τη στατιστική συνάρτηση

$F = \frac{n-k-p+1}{p} l_1$ όπου l_1 είναι η μεγαλύτερη ιδιοτιμή του πίνακα $W^{-1}B$. Η στατιστική συνάρτηση F ακολουθεί την κατανομή F με βαθμούς ελευθερίας |k-p-1|+1 και n-k-p+1 αντίστοιχα. Απορρίπτουμε για μεγάλες τιμές της F.

2. Έλεγχος Pilai. Η στατιστική συνάρτηση είναι η $tr(B+W)^{-1}B = \sum_{j=1}^{k} \frac{1}{\lambda_j}$.

3. Έλεγχος ίχνους Lawley-Hoteling. Η στατιστική συνάρτηση είναι η $tr(W^{-1}B) = \sum_{j=1}^{k} l_j$.

Ενώ για τον έλεγχο του Roy καταλήγουμε σε γνωστή κατανομή, αυτό δεν ισχύει για τους άλλους δυο ελέγχους όπου η κατανομή δεν είναι γενικά γνωστή αν και υπάρχουν πίνακες.

Θα πρέπει να σημειωθεί πως αν οι υποθέσεις του μοντέλου μας είναι γενικά σωστές (πολυμεταβλητή κανονικότητα και σταθερή διακύμανση) όλοι οι έλεγχοι θα πρέπει να μας δίνουν ίδιο αποτέλεσμα, και p-value τα οποία να είναι κοντά. Για αυτό σε περίπτωση που παρατηρούμε μεγάλες αποκλίσεις στα αποτελέσματα χρησιμοποιώντας τους ελέγχους αυτό είναι μια ισχυρή ένδειξη πως θα πρέπει να κοιτάξουμε τα δεδομένα για το κατά πόσο η MANOVA μπορεί να εφαρμοστεί. Σε σχέση με την ισχύ των ελέγχων το ποιος από αυτούς είναι καλύτερος εξαρτάται από το σχεδιασμό του πειράματος. Σαν συμβουλή έλεγχοι που βασίζονται στη συνάρτηση Λ καλό είναι να αποφεύγονται όταν η υπόθεση της ισότητας των διακυμάνσεων δεν ισχύει. Ο έλεγχος Pilai έχει βρεθεί να είναι πιο ανθεκτικός (robust) όταν οι υποθέσεις δεν τηρούνται.

Στην περίπτωση που έχουμε 2 γκρουπ όλοι οι έλεγχοι είναι ισοδύναμοι με τον έλεγχο του Hotelling.

Θα πρέπει να τονιστεί πως, όπως και στην απλή ANOVA χρειαζόμαστε την υπόθεση ότι οι πίνακες διακυμάνσεων είναι ίσοι για να είναι αξιόπιστα τα αποτελέσματα. Χωρίς αυτή την υπόθεση η μεθοδολογία δεν ισχύει πια. Για να ελέγξουμε την υπόθεση ότι όλοι οι πίνακες διακύμανσης είναι ίσοι μπορούμε να χρησιμοποιήσουμε τον έλεγχο Box-M που περιγράψαμε στο κεφάλαιο 5.

Σε περίπτωση που απορρίψουμε την υπόθεση των ίσων πινάκων διακυμάνσεων, όπως και στην μονομεταβλητή περίπτωση, χρειάζεται να κάνουμε μετασχηματισμούς ώστε να μετασχηματίσουμε τα δεδομένα να έχουν ίσους πίνακες διακυμάνσεων. Αυτό σημαίνει πως πρέπει κανείς πρώτα να εξετάσει ποιες διακυμάνσεις ή συνδιακυμάνσεις του πίνακα είναι αυτές που δημιουργούν το πρόβλημα. Επειδή γενικά πολυμεταβλητοί μετασχηματισμοί δεν είναι καθόλου απλοί προτρέπεται ο ερευνητής να αρχίσει από απλούς μετασχηματισμούς κάποιων από τις μεταβλητές που έχουν διαφορετικές διακυμάνσεις και στη συνέχεια να προσπαθήσει να αντιμετωπίσει το πρόβλημα μα ταυτόχρονους μετασχηματισμούς σε όλον τον πίνακα διακυμάνσεων.

Πάντως ο έλεγχος αυτός είναι πολύ ευαίσθητος σε αποκλίσεις από την κανονικότητα. Δηλαδή όταν τα δεδομένα δεν προέρχονται από κανονικούς πληθυσμούς η ισχύς του ελέγχου είναι πολύ μικρή.

Ολοκληρώνοντας αυτή τη σύντομη παρουσίαση της μεθόδου MANOVA πρέπει να τονίσουμε πως μπορεί να γενικευτεί παρόμοια με την περίπτωση της απλής ANOVA. Δηλαδή μπορούμε να έχουμε δύο παράγοντες ή και περισσότερους, μπορούμε να ορίσουμε αλληλεπιδράσεις, μπορούμε να κάνουμε πολλαπλούς ελέγχους για να βρούμε ποιες ομάδες διαφέρουν μεταξύ τους κλπ. Επίσης μπορούμε να χρησιμοποιήσουμε πειραματικούς σχεδιασμούς σε συνδυασμό με MANOVA.

6.4 Πίνακας Ανάλυσης Διακύμανσης

Από όσα αναφέρθηκαν πιο πάνω η ιδέα της MANOVA είναι παρόμοια με αυτή της μονομεταβλητής περίπτωσης. Δηλαδή τη συνολική μεταβλητότητα την σπάμε σε δύο κομμάτια. Στην απλή ANOVA κυρίαρχο ρόλο την πρακτική εφαρμογή της μεθόδου έπαιζε ο πίνακας ανάλυσης διακύμανσης. Και στη MANOVA μπορούμε να κατασκευάσουμε τον αντίστοιχο πίνακα ανάλυσης διακύμανσης που έχει ως εξής

Πηγή	SSP πίνακας	Βε	Test
Μεταξύ γκρουπ (between)	$B=\sum_{j=1}^{k} n_j(\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet})(\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet})'$	*k-1*	$\Lambda=\frac{\lvert W\rvert}{\lvert W+B\rvert}$
Μεσα στα γκρουπ (within)	$W=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_{\bullet j})(x_{ij}-\bar{x}_{\bullet j})'$	*n-k*	
Συνολική	$T=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_{\bullet\bullet})(x_{ij}-\bar{x}_{\bullet\bullet})'$	*n-1*	

Παρατηρείστε πως οι διαφορές είναι ελάχιστες. Για παράδειγμα αντί να έχουμε αθροίσματα τετραγώνων τώρα πια έχουμε πίνακες με αθροίσματα τετραγώνων. Συνήθως αυτοί οι πίνακες ονομάζονται sum of squared cross products (ή SSCP). Αν κανείς ελέγξει τους τύπους με τους οποίους προκύπτουν οι πίνακες αυτοί θα δει πως είναι ακριβώς αντίστοιχοι με τα αθροίσματα τετραγώνων της ANOVA. Μάλιστα αποδεικνύεται πολύ εύκολα πως αν p=1, δηλαδή έχω μόνο μια μεταβλητή, τότε ο πίνακας ταυτίζεται με τον πίνακα της απλής ANOVA. Ακόμα και η ελεγχοσυνάρτηση Λ του Wilks σε αυτή την περίπτωση είναι ανάλογη με την τιμή της ελεγχοσυνάρτησης F για τον απλό έλεγχο.

6.5 Πολυμεταβλητή Παλινδρόμηση

Η πολυμεταβλητή Παλινδρόμηση αποτελεί μια γενίκευση της απλής παλινδρόμησης όπου εξεταζόταν η σχέση μιας μεταβλητής Υ με μια σειρά επεξηγηματικών μεταβλητών X_1, X_2, …., X_k . Στην πολυμεταβλητή παλινδρόμηση μπορούμε να εξετάσουμε περισσότερες από μια μεταβλητές. Με αυτό κερδίζουμε στο γεγονός ότι μας επιτρέπει να λάβουμε υπόψη τη συνδιακύμανση που έχουν οι μεταβλητές κάτι που αν χρησιμοποιούσαμε πολλά μονομεταβλητά μοντέλα θα το αγνοούσαμε. Το μοντέλο μπορεί να γραφτεί ως

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

όπου

Υ είναι ένας n x m πίνακας που περιέχει τις m μεταβλητές για τις n παρατηρήσεις μας

X είναι ένας n x p πίνακας που περιέχει τις p επεξηγηματικές μεταβλητές για τις n παρατηρήσεις μας (η πρώτη στήλη περιέχει μονάδες που αντιστοιχούν στη σταθερά, άρα έχουμε στην ουσία p-1 μεταβλητές)

B είναι ένας p x m πίνακας με τους συντελεστές της παλινδρόμησης όπου ο ij συντελεστής αναφέρεται στο συντελεστή της j επεξηγηματικής μεταβλητής στην i μεταβλητή και

E είναι ένας πίνακας n x m με τα τυχαία σφάλματα.

Για το μοντέλο αυτό υποθέτουμε πως το διάνυσμα των σφαλμάτων ε_i ακολουθεί πολυμεταβλητή κανονική κατανομή με $E(\varepsilon_i)=0$ και $Var(\varepsilon_i)=\Sigma$.

Είναι ιδιαίτερα σημαντικό να παρατηρήσουμε πως η εκτιμήτρια ελαχίστων τετραγώνων για τον πίνακα των συντελεστών **B** δίνεται και στην πολυμεταβλητή παλινδρόμηση από τον τύπο

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Μπορεί επίσης να δειχτεί ότι οι συντελεστές είναι οι ίδιοι που θα έπαιρνε κάποιος αν έκανε τις παλινδρομήσεις μια προς μια. Εκεί που διαφέρει και αυτό είναι σημαντικό είναι πως επιτρέπει έναν όρο συνδιακύμανσης στα σφάλματα τον οποίο δεν θα μπορούσαμε να πάρουμε κάνοντας μια τις παλινδρομήσεις. Κάτω από αυτό το γενικό μοντέλο μπορεί κανείς να δει τόσο την πολυμεταβλητή ανάλυση διακύμανσης όσο και την πολυμεταβλητή ανάλυση συνδιακύμανσης

6.6 Η MANOVA ως Γραμμικό Μοντέλο

Η MANOVA μπορεί να γραφτεί ως ένα γραμμικό μοντέλο πολυμεταβλητής παλινδρόμησης με όμοιο τρόπο όπως και στην μονομεταβλητή περίπτωση δηλαδή το μοντέλο είναι το

$$x_{ij} = \mu + r_j + \varepsilon_{ij}, \qquad i=1,2,\ldots,n_j,\ j=1,\ldots,k$$

όπου

μ είναι το διάνυσμα των γενικών μέσων

r_j είναι το διάνυσμα με την επίδραση του j παράγοντα και

ε_{ij} είναι τα τυχαία σφάλματα που ακολουθούν $N_p(0,\Sigma)$.

Όπως και στην μονομεταβλητή περίπτωση η MANOVA μπορεί να γίνει με τη χρήση του γραμμικού μοντέλου και τη χρήση ψευδομεταβλητών (dummy variables). Ομοίως μπορούμε να γενικεύσουμε σε μοντέλα με περισσότερους από έναν παράγοντες και με αλληλεπιδράσεις. Βάζοντας στο παραπάνω μοντέλο και άλλες συνεχής μεταβλητές καταλήγουμε σε μοντέλα πολυμεταβλητής ανάλυσης συνδιακύμανσης

6.7 Άλλα Θέματα

Ένα πρόβλημα που μπορεί να εμφανιστεί στην MANOVA και που είναι άγνωστο στη μονομεταβλητή περίπτωση είναι το γεγονός ότι ο πίνακας διακύμανσης συνδιακύμανσης μπορεί να μην είναι πλήρους βαθμού. Αυτό συμβαίνει όταν κάποιες μεταβλητές είναι πολύ ισχυρά συσχετισμένες. Για λόγους τυχαίου σφάλματος, στην πράξη κάποιος δεν θα δει μια μηδενική ορίζουσα αλλά μια ορίζουσα πολύ κοντά στο 0. Για αυτό καλό θα ήταν να εξεταστεί πρώτα η ύπαρξη τέτοιων μεταβλητών. Δουλεύοντας πιο αυστηρά θα μπορούσε κάποιος να εργαστεί με τις λεγόμενες κανονικές μεταβλητές (canonical variates). Η μέθοδος αυτή που μπορεί να ειδωθεί ως μια πολυμεταβλητή γενίκευση της ανάλυσης σε κύριες συνιστώσες μειώνει τις διαστάσεις του προβλήματος δημιουργώντας διανύσματα μεταβλητών (και όχι απλές μεταβλητές όπως η ανάλυση σε κύριες συνιστώσες) που έχουν δομή συνδιακύμανσης όμοια με την αρχική.

6.8 Παράδειγμα (Fisher's Iris Data).

Τα δεδομένα αυτά αποτελούνται από k=3 είδη τριαντάφυλλων (iris setosa, versicolour, virginica). Από κάθε είδος ο ερευνητής πήρε 50 ($n_j = 50$) τριαντάφυλλα και κατέγραψε τα εξής χαρακτηριστικά ως μεταβλητές:

x_1 : το μήκος του σέπαλου, x_2 : το πλάτος του σέπαλου

x_3 : το μήκος του πετάλου, x_4 : το πλάτος του πετάλου

Έχουμε επίσης n=150.

Οι μέσοι για κάθε δείγμα βρέθηκαν

$$\bar{x}_{\bullet 1} = \begin{pmatrix} 5.01 \\ 3.43 \\ 1.46 \\ 0.25 \end{pmatrix}, \quad \bar{x}_{\bullet 2} = \begin{pmatrix} 5.94 \\ 2.77 \\ 4.26 \\ 1.33 \end{pmatrix}, \quad \bar{x}_{\bullet 3} = \begin{pmatrix} 6.59 \\ 2.97 \\ 5.55 \\ 2.03 \end{pmatrix}$$

Δουλεύοντας με τους τύπους βρήκαμε πως οι πίνακες που χρειαζόμαστε είναι οι

$$B = \begin{bmatrix} 63.21 & -19.95 & 165.25 & 71.28 \\ & 11.35 & -57.24 & -22.93 \\ & & 437.11 & 186.78 \\ & & & 80.41 \end{bmatrix} \text{ και}$$

$$W = \begin{bmatrix} 38.96 & 13.63 & 24.62 & 5.64 \\ & 16.96 & 8.12 & 4.81 \\ & & 27.22 & 6.27 \\ & & & 6.16 \end{bmatrix}$$

από τους οποίους βρίσκουμε πως $|W| = 19976.4$ και $|B+W| = 871968$ και επομένως $\Lambda = 0.023$. Δεδομένου πως έχω 3 γκρουπ μπορώ να χρησιμοποιήσω τους ακριβείς τύπους που περιγράφτηκαν. Έτσι ξέρω πως η ποσότητα $\frac{(n-k-p+1)(1-\sqrt{\Lambda})}{p\sqrt{\Lambda}} = \frac{(150-3-4+1)(1-\sqrt{0.023})}{4\sqrt{0.023}} = 201.25$ ακολουθεί την κατανομή F με βαθμούς ελευθερίας 2p και 2(n-k-p+1). Επειδή η κριτική τιμή της κατανομής F με 8 και 288 βαθμούς ελευθερίας αντίστοιχα είναι 1.94 (α=5%) απορρίπτουμε τη μηδενική υπόθεση περί ισότητας των μέσων.

Χρησιμοποιώντας το ασυμπτωτικό αποτέλεσμα του Bartlett (παρατηρείστε πως επειδή στην περίπτωση μας έχουμε κάποιο ακριβές αποτέλεσμα δεν είναι σωστό να χρησιμοποιήσουμε το ασυμπτωτικό), βρίσκουμε πως η ελεγχοσυνάρτηση έχει τιμή 548.86 η οποία συγκρινόμενη με την κριτική τιμή από μια X^2 κατανομή με 8 βαθμούς ελευθερίας οδηγεί και πάλι στην απόρριψη της μηδενικής υπόθεσης.

Παρατηρείστε ότι τα διαγώνια στοιχεία των πινάκων B και W είναι αυτά που χρησιμοποιούνται για τους μονομεταβλητούς ελέγχους.

Έτσι έχουμε

	Variable sum of squares	Error Sum of Squares	F	p-value
x_1 : το μήκος του σέπαλου,	63.21	38.96	119.26	0.000
x_2 : το πλάτος του σέπαλου	11.35	16.96	49.16	0.000
x_3 : το μήκος του πετάλου,	437.11	27.22	1180.16	0.000
x_4 : το πλάτος του πετάλου	80.41	6.16	960.01	0.000
βαθμοί ελευθερίας (για όλα)	2	147		

κάνοντας δηλαδή μονομεταβλητούς ελέγχους θα απορρίπταμε για όλες τις μεταβλητές την ισότητα των μέσων.

Ο πίνακας ανάλυσης διακύμανσης για τα παραπάνω δεδομένα είναι ο

Πηγή	SSP πίνακας	Βε	Test
Ανάμεσα στα γκρουπ (between)	$B=\begin{bmatrix} 63.21 & -19.95 & 165.25 & 71.28 \\ & 11.35 & -57.24 & -22.93 \\ & & 437.11 & 186.78 \\ & & & 80.41 \end{bmatrix}$	*2*	*0.023*
Μεσα στα γκρουπ (within)	$W=\begin{bmatrix} 38.96 & 13.63 & 24.62 & 5.64 \\ & 16.96 & 8.12 & 4.81 \\ & & 27.22 & 6.27 \\ & & & 6.16 \end{bmatrix}$	*147*	
Συνολική	T=W+B	*1491*	

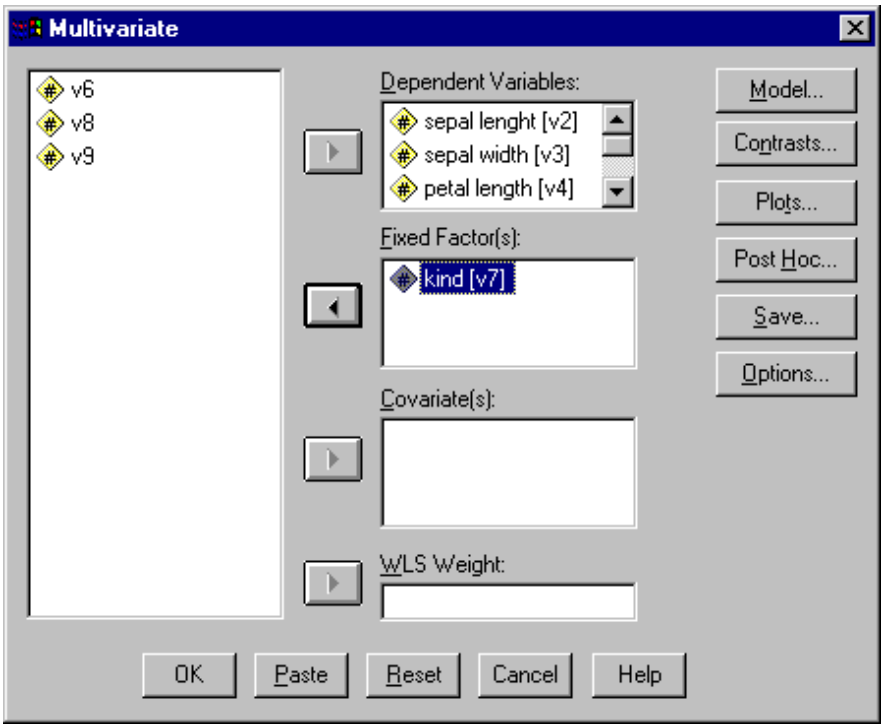
Πίνακας 6.1 Ο πίνακας ανάλυσης διακύμανσης για τα δεδομένα του Fisher

6.9 MANOVA με τη Χρήση του Πακέτου SPSS

Η καινούρια έκδοση του SPSS for Windows δεν προσφέρει την πολυμεταβλητή ανάλυση διακύμανσης ως αυτάρκη μέθοδο παρά μόνο κάτω από το γενικότερο πλαίσιο του γενικού πολυμεταβλητού γραμμικού μοντέλου. Συνεπώς πρέπει να επιλέξουμε

Analyse > General Linear Model > Multivariate

Εμφανίζεται τότε το παράθυρο της εικόνας 6.1. Πρέπει να επιλέξουμε τότε τις μεταβλητές που θέλουμε να χρησιμοποιήσουμε για την ανάλυση βάζοντας αυτές στο παράθυρο Dependent variables και τη μεταβλητή που μας δείχνει σε πιο γκρουπ ανήκει η παρατήρηση στο παράθυρο Fixed Factors.Οι υπόλοιπες επιλογές δεν μας αφορούν αλλά γενικά μιλώντας επιτρέπουν να χρησιμοποιήσουμε άλλες επεξηγηματικές μεταβλητές (covariates) να επιλέξουμε επιδράσεις για το μοντέλο μας (αν έχουμε περισσότερους από έναν παράγοντες κλπ.



Εικόνα 6.1. Το βασικό παράθυρο του πολυμεταβλητού γενικού γραμμικού μοντέλου

Εναλλακτικά κάποιος μπορεί να γράψει απευθείας στο παράθυρο Syntax την εντολή που κάνει MANOVA. Η απλούστερη μορφή της εντολής είναι η

MANOVA dependent variables BY factor variables (min, max).

όπου dependent variables είναι οι μεταβλητές που θα χρησιμοποιήσουμε και factor variables η μεταβλητή (ες) που περιέχουν την ομάδα που ανήκει κάθε παρατήρηση. Έτσι μια εντολή που θα έκανε MANOVA για την περίπτωση μας είναι

MANOVA SEPLEN SEPWID PETLEN PETWIDTH BY KIND(1,3)

όπου

SEPLEN, SEPWID, PETLEN, PETWIDTH οι μεταβλητές μας και KIND μια μεταβλητή με το είδος του τριαντάφυλλου.

Δουλεύοντας με το Syntax έχουμε πολύ περισσότερες επιλογές ως προς την ανάλυση μας. Επίσης τα outputs που παίρνουμε είναι πιο εύκολο να διαβαστούν αφού αφορούν την MANOVA και όχι τη γενικότερη περίπτωση του γενικού γραμμικού μοντέλου.

7 ΑΝΑΛΥΣΗ ΣΕ ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ

7.1 Εισαγωγή

Η μέθοδος των κυρίων συνιστωσών (Principal Components Analysis) είναι μια μέθοδος η οποία έχει σκοπό να δημιουργήσει γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε οι γραμμικοί αυτοί συνδυασμοί να είναι ασυσχέτιστοι μεταξύ τους αλλά να περιέχουν όσο γίνεται μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών. Το κέρδος από μια τέτοια διαδικασία είναι πως:

- Από ένα σύνολο συσχετισμένων μεταβλητών καταλήγουμε σε ένα σύνολο ασυσχέτιστων μεταβλητών, κάτι το οποίο για ορισμένες στατιστικές μεθόδους είναι περισσότερο χρήσιμο. Για παράδειγμα θυμηθείτε το πρόβλημα της πολυσυγγραμικότητας στην παλινδρόμηση, όπου αν χρησιμοποιήσουμε τις συσχετισμένες μεταβλητές οι εκτιμήσεις που θα πάρουμε δεν θα είναι συνεπείς ενώ αν χρησιμοποιούσαμε ασυσχέτιστες μεταβλητές το πρόβλημα αυτό δεν θα υπήρχε.

- Αν οι κύριες συνιστώσες που θα προκύψουν μπορούν να ερμηνεύσουν ένα μεγάλο ποσοστό της διακύμανσης τότε αυτό σημαίνει πως αντί να έχουμε p μεταβλητές όπως είχαμε αρχικά, έχουμε λιγότερες, με κόστος βέβαια ότι χάνουμε κάποιο (ελπίζουμε μικρό) ποσοστό της συνολικής μεταβλητότητας. Σε μερικές εφαρμογές αυτό είναι ζωτικής σημασίας. Για παράδειγμα σε μια τεράστια βάση δεδομένων αντί να αποθηκεύουμε όλες τις μεταβλητές μπορούμε να αποθηκεύουμε μόνο κάποιον αριθμό κυρίων συνιστωσών. Σίγουρα χάνουμε κάποιο μέρος της πληροφορίας αλλά το κέρδος σε χώρο αλλά και ταχύτητα επεξεργασίας μπορεί να είναι τεράστιο. Από την άλλη πλευρά πολλές φορές συμβαίνει να έχουμε λίγες παρατηρήσεις αλλά πολλές μεταβλητές. Τέτοια προβλήματα για παράδειγμα εμφανίζονται στην αρχαιομετρία ένα πεδίο εφαρμογής στατιστικών μεθόδων στην αρχαιολογία, όπου τα αντικείμενα που θέλει κάποιος να μελετήσει είναι συνήθως λίγα (π.χ. αμφορείς της κλασικής περιόδου) αλλά τα στοιχεία και οι μεταβλητές που έχει είναι πάρα πολλά. Η μείωση των διαστάσεων του προβλήματος φαντάζει η μόνη λύση για να προχωρήσει κανείς σε στατιστική επεξεργασία.

- Ένα άλλο μεγάλο πλεονέκτημα (το οποίο από την άλλη ίσως είναι και μειονέκτημα για πολλούς) είναι πως με τη μέθοδο των κυριών συνιστωσών μπορούμε να εξετάσουμε τις συσχετίσεις ανάμεσα στις μεταβλητές και να διαπιστώσουμε πόσο οι μεταβλητές μοιάζουν ή όχι. Επίσης η μέθοδος μας επιτρέπει να αναγνωρίσουμε δίνοντας ονόματα στις καινούριες μεταβλητές (τις συνιστώσες) παρατηρώντας ποιες από τις αρχικές μεταβλητές έχουν μεγάλη επίδραση σε αυτές. Αυτό είναι πολύ χρήσιμο σε κάποιες επιστήμες καθώς μας επιτρέπουν να ποσοτικοποιήσουμε μη μετρήσιμες ποσότητες, όπως η αγάπη, η ευφυΐα, η ικανότητα ενός μπασκετμπολίστα, η εμπορευσιμότητα ενός προϊόντος κλπ αφηρημένες έννοιες. Το γεγονός βέβαια πως τέτοιες ερμηνείες εμπεριέχουν σε μεγάλο βαθμό υποκειμενικά κριτήρια έχει οδηγήσει πολλούς στο να κατηγορούν τη μέθοδο και να μην την εμπιστεύονται.

Στα πλεονεκτήματα και τα μειονεκτήματα όμως της μεθόδου θα επανέλθουμε προς το τέλος του κεφαλαίου όταν πια θα έχουμε αποκτήσει μια πιο καλή εικόνα για την μέθοδο.

7.2 Η Βασική Ιδέα

Πριν ξεκινήσουμε την περιγραφή της μεθόδου των κυρίων συνιστωσών είναι χρήσιμο να δούμε κάποια πράγματα από τη γραμμική άλγεβρα τα οποία και αποτέλεσαν τη βασική ιδέα πάνω στην οποία αναπτύχθηκε η μέθοδος.

Έστω ένας τετραγωνικός συμμετρικός πίνακας $\mathbf{A}$ διαστάσεων $p \times p$. Ο πίνακας αυτός μπορεί να αναπαρασταθεί ως

$$\mathbf{A} = \mathbf{P}\ \mathbf{\Lambda}\ \mathbf{P}'$$

όπου $\mathbf{\Lambda}$ είναι ένας $p \times p$ διαγώνιος πίνακας όπου τα στοιχεία της διαγωνίου είναι οι ιδιοτιμές του πίνακα $\mathbf{A}$, δηλαδή

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & \\ \dots & & & \\ 0 & \dots & \dots & \lambda_p \end{bmatrix}$$

και $\mathbf{P}$ είναι ένας ορθογώνιος $p \times p$ πίνακας (δηλαδή ισχύει $\mathbf{P}'\ \mathbf{P} = \mathbf{I}$) ο οποίος αποτελείται από τα κανονικοποιημένα ιδιοδιανύσματα των αντίστοιχων ιδιοτιμών. Η παραπάνω αναπαράσταση του πίνακα Α ονομάζεται φασματική ανάλυση του πίνακα **A.** Επομένως αφού ο πίνακας είναι ορθογώνιος θα ισχύει πως $\mathbf{P}^{-1} = \mathbf{P}'$

Μπορεί κάποιος να δείξει με βάση τις παραπάνω ιδιότητες πως ισχύει

$$\mathbf{\Lambda = P' A\ P} \tag{7.1}$$

καθώς

$$\mathbf{A = P \Lambda P' \Leftrightarrow P^{-1} A = P^{-1} P \Lambda P' \Leftrightarrow}$$

$$\mathbf{\Leftrightarrow P^{-1} A P = \Lambda P' P = \Lambda}$$

Δηλαδή αυτό που με απλά λόγια είδαμε είναι πως αν ξεκινήσουμε από έναν τετραγωνικό πίνακα **A** μπορούμε να καταλήξουμε σε έναν διαγώνιο πίνακα **Λ.**

Γιατί αυτό όμως μας είναι τόσο χρήσιμο; Θυμηθείτε πως αν έχουμε ένα τυχαίο διάνυσμα **X** το οποίο έχει πίνακα διακύμανσης **Σ** τότε το διάνυσμα **Υ=BX** έχει πίνακα διακύμανσης **BΣB'.** Αν τώρα κοιτάξουμε την σχέση (7.1) βλέπουμε πως από έναν τετραγωνικό πίνακα μπορώ να οδηγηθώ σε έναν διαγώνιο πίνακα, πολλαπλασιάζοντας με έναν κατάλληλο πίνακα **P** και άρα αν ο τετραγωνικός πίνακας είναι πίνακας διακύμανσης καταλήγουμε σε έναν διαγώνιο πίνακα διακύμανσης. Δηλαδή το τυχαίο διάνυσμα που αντιστοιχεί στον πίνακα αυτόν είναι ασυσχέτιστο. Δηλαδή αυτό που μου δίνει η φασματική ανάλυση ενός πίνακα διακύμανσης είναι πως αν πολλαπλασιάσω το αρχικό διάνυσμα με έναν κατάλληλο πίνακα μπορώ να δημιουργήσω έναν νέο διάνυσμα το οποίο να είναι ασυσχέτιστο, να έχει δηλαδή διαγώνιο πίνακα διακύμανσης.

7.3 Εύρεση των Κυριών Συνιστωσών

Όπως είπαμε προηγούμενα η μέθοδος στηρίζεται στη φασματική ανάλυση ενός τετραγωνικού πίνακα. Αυτό σημαίνει πως μπορούμε να χρησιμοποιήσουμε είτε τον πίνακα διακυμάνσεων είτε τον πίνακα συσχετίσεων που είναι στην ουσία ο πίνακας διακυμάνσεων των τυποποιημένων δεδομένων.

Έστω λοιπόν πως έχουμε ένα σύνολο από k μεταβλητές $(X_1, X_2, ..., X_k)$ και θέλουμε να δημιουργήσουμε τις κύριες συνιστώσες $(Y_1, Y_2, ..., Y_k)$ οι οποίες να είναι γραμμικός συνδυασμός των αρχικών μεταβλητών , δηλαδή

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + ... + a_{1k}X_k \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + ... + a_{2k}X_k \\ &..... \\ Y_k &= a_{k1}X_1 + a_{k2}X_2 + ... + a_{kk}X_k \end{aligned}$$

Υπό μορφή πινάκων μπορεί να γραφτεί ως $\mathbf{Y = AX}$ όπου **Υ, X** είναι διανύσματα $k \times 1$ και **A** είναι $k \times k$ πίνακας με στοιχεία

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \dots & & \dots & \\ a_{k1} & a_{k2} & \dots & a_{kk} \end{bmatrix}=\begin{bmatrix} a_1 & a_2 & \dots & a_k \end{bmatrix}$$

όπου a_j είναι το διάνυσμα στήλη με στοιχεία $a_j'=\begin{bmatrix} a_{j1} & a_{j2} & \dots & a_{jk} \end{bmatrix}$, $j=1,\dots,k$, και για να μην έχουμε προβλήματα ταυτοποίησης θέτουμε $\sum_{i=1}^{k} {a_{ji}}^2 = a_j' a_j = 1$.

Επομένως το πρόβλημα εύρεσης των κυρίων συνιστωσών είναι το πρόβλημα της εύρεσης των στοιχείων του πίνακα **A**. Έχουμε όμως έναν επιπλέον περιορισμό, ότι δηλαδή οι κύριες συνιστώσες πρέπει να είναι σε φθίνουσα σειρά ως προς τη διακύμανση τους, δηλαδή η πρώτη να έχει τη μεγαλύτερη διακύμανση, η δεύτερη τη δεύτερη μεγαλύτερη και ούτω καθεξής. Παρατηρείστε πως έχουμε ήδη δει (από την (7.1)) ότι τα ιδιοδιανύσματα αποτελούν μια λύση στο πρόβλημα αν εξαιρέσουμε την τελευταία υπόθεση για τη φθίνουσα σειρά της διακύμανσης.

Ας δουλέψουμε για την πρώτη κύρια συνιστώσα $Y_1 = a_1' X$. Είναι σαφές πως $Var(Y_1) = a_1' \mathbf{\Sigma} a_1$ όπου **Σ** ο πίνακας διακυμάνσεων του τυχαίου διανύσματος X. Επομένως για να βρούμε το a_1 θα πρέπει να μεγιστοποιήσουμε την $Var(Y_1)$ με τον περιορισμό πως $a_1' a_1 = 1$ δηλαδή θα μεγιστοποιήσουμε τη συνάρτηση

$$L(a_1) = a_1' \Sigma a_1 - \lambda (a_1' a_1 - 1),$$

όπου λ είναι ο πολλαπλασιαστής Lagrange.

Χρησιμοποιώντας παραγώγους διανυσμάτων βρίσκουμε πως

$$\frac{\partial L(a_1)}{\partial a_1} = 2(\mathbf{\Sigma} - \lambda \mathbf{I}) a_1 = 0$$

και επομένως αντιστοιχεί στο να λύσουμε την εξίσωση

$$\mathbf{\Sigma} a_1 = \lambda a_1$$

η οποία είναι η εξίσωση των ιδιοδιανυσμάτων του πίνακα **Σ** όπου λ είναι η ιδιοτιμή. Δηλαδή κάθε ζεύγος ιδιοτιμής και του ιδιοδιανύσματος που τη συνοδεύει είναι λύση της εξίσωσης, και άρα έχουμε k δυνατές λύσεις. Από αυτές πρέπει να διαλέξουμε ποια οδηγεί σε μεγαλύτερη διακύμανση. Η διακύμανση του Y_1 θα είναι ίση με λ, και επομένως αρκεί να διαλέξουμε το ζεύγος ιδιοτιμής και ιδιοδιανύσματος που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή.

Με παρόμοια επιχειρήματα μπορούμε να δούμε πως για όλες τις κύριες συνιστώσες τα διανύσματα a_j που χρειαζόμαστε θα αντιστοιχούν στα ιδιοδιανύσματα της j σε φθίνουσα σειρά ιδιοτιμής. Φυσικά για την εύρεση των υπόλοιπων κυρίων συνιστωσών χρειάζεται να προσθέσουμε έναν ακόμη περιορισμό: ότι οι κύριες συνιστώσες είναι ασυσχέτιστες με τις προηγούμενες τους.

Επομένως :

- Για να κατασκευάσουμε τις κύριες συνιστώσες χρειάζεται να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα $\mathbf{\Sigma}$ που χρησιμοποιούμε.

- Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμα της αντιστοιχούν στην πρώτη κύρια συνιστώσα, η δεύτερη μεγαλύτερη ιδιοτιμή στη δεύτερη κύρια συνιστώσα κλπ.

- Η διακύμανση της κάθε κύριας συνιστώσας είναι ίση με την ιδιοτιμή που της αντιστοιχεί. Έτσι αν συμβολίσουμε με λ_j την j μεγαλύτερη ιδιοτιμή τότε έχουμε πως $Var(Y_j) = \lambda_j$.

- Όπως είπαμε και πριν οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους και άρα ο πίνακας διακύμανσης τους είναι ο διαγώνιος με διαγώνια στοιχεία τις ιδιοτιμές λ_j

- Η συνολική διακύμανση των κύριων συνιστωσών θα είναι η ίδια με τη συνολική διακύμανση των αρχικών μεταβλητών εξαιτίας των ιδιοτήτων του ίχνους συμμετρικού και τετραγωνικού πίνακα. Δηλαδή θα ισχύει $tr(\mathbf{\Sigma}) = tr(\Lambda)$ και άρα η συνολική διακύμανση διατηρείται.

- Επίσης η γενικευμένη διακύμανση των κυριών συνιστωσών είναι η ίδια με τη γενικευμένη διακύμανση των αρχικών μεταβλητών. Αυτό προκύπτει εύκολα καθώς η ορίζουσα ενός τετραγωνικού πίνακα είναι το γινόμενο των ιδιοτιμών της και άρα ισχύει $|\mathbf{\Sigma}| = \prod_{i=1}^{p} \lambda_i = |\Lambda|$.

- Η ποσότητα $\dfrac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$ μας δείχνει το ποσοστό της συνολικής διακύμανσης που εξηγεί η j συνιστώσα. Είναι ευνόητο πως αν κάποιος πάρει όλες τις συνιστώσες τότε θα διατηρήσει όλη τη διακύμανση, ενώ αν τελικά παραλείψει κάποιες συνιστώσες κάποιο ποσοστό της διακύμανσης θα χαθεί. Προφανώς συμφέρει να διατηρούμε τις πρώτες συνιστώσες που εξηγούν μεγαλύτερο κομμάτι της διακύμανσης.

Παράδειγμα. Έστω ο παρακάτω πίνακας διακύμανσης $\mathbf{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$. Οι ιδιοτιμές του πίνακα είναι (σε φθίνουσα σειρά) $\lambda_1 = 5.83,\ \lambda_2 = 2,\ \lambda_3 = 0.17$ καθώς και τα ιδιοδιανύσματα που τους αντιστοιχούν είναι τα

$$a_1' = (-0.383 \quad 0.924 \quad 0),\ a_2' = (0 \quad 0 \quad 1),\ a_3' = (0.924 \quad 0.383 \quad 0).$$

Χρησιμοποιώντας λοιπόν αυτά βρίσκουμε πως οι κύριες συνιστώσες είναι:

$$\begin{aligned} Y_1 &= -0.383X_1 + \quad 0.924X_2 \\ Y_2 &= \qquad X_3 \\ Y_3 &= \ 0.924X_1 + \quad 0.383X_2 \end{aligned}$$

Παρατηρείστε τα εξής:

- Προφανώς δεν υπάρχει περίπτωση μια ιδιοτιμή να είναι αρνητική αφού ένας πίνακας διακύμανσης είναι πάντα θετικά ορισμένος και άρα οι ιδιοτιμές του είναι θετικές ή μηδέν.

- Η μεταβλητή X_3 η οποία ήταν ασυσχέτιστη με τις υπόλοιπες είναι η δεύτερη κύρια συνιστώσα. Συνεπώς αν χρησιμοποιήσω ασυσχέτιστες μεταβλητές στην ανάλυση σε κύριες συνιστώσες αυτές θα ταυτιστούν με κάποια συνιστώσα και επομένως δεν κερδίζω τίποτα με το να τις χρησιμοποιήσω στην ανάλυση. Θυμηθείτε πως ένας σκοπός της ανάλυσης σε κύριες συνιστώσες ήταν να οδηγηθώ σε ασυσχέτιστες μεταβλητές. Αν ξεκινήσω από ασυσχέτιστες δεν έχει νόημα η ανάλυση

- Αν αλλάξω τα πρόσημα στις τιμές των ιδιοδιανυσμάτων αυτά συνεχίζουν να είναι λύσεις με την έννοια ότι ικανοποιούν όλες τις συνθήκες. Επομένως οι κύριες συνιστώσες δεν είναι μοναδικές καθώς μπορώ να αλλάξω τα πρόσημα. Αυτό έχει σαν αποτέλεσμα να μην είναι ξεκάθαρη η ερμηνεία τους. Παρόλα αυτά καθώς οι τιμές των συντελεστών δεν θα αλλάξουν σε απόλυτη τιμή μπορώ να 'αναγνωρίσω' κάπως τις συνιστώσες και την επίδραση των αρχικών μεταβλητών σε αυτές άσχετα με το πρόσημο.

- Οι διακυμάνσεις των τριών κυρίων συνιστωσών είναι

$$Var(Y_1) = 5.83$$
$$Var(Y_2) = 2$$
$$Var(Y_3) = 0.17$$

και άρα η συνολική διακύμανση είναι 8 (όπως και στα αρχικά δεδομένα). Επομένως η πρώτη κύρια συνιστώσα εξηγεί το 5.83/8 = 72.8% της συνολικής διακύμανσης, ενώ η 2η κύρια συνιστώσα το 25%. Και οι δύο μαζί εξηγούν το 97.8% και άρα αν αποφασίσει κάποιος να μην διατηρήσει την τρίτη (πιθανότατα για να περιορίσει τον αριθμό των μεταβλητών) του θα χάσει μόλις το 2.2% της πληροφορίας που είχαν τα αρχικά δεδομένα.

7.4 Αλλαγή Κλίμακας

Ένα από τα μειονεκτήματα της ανάλυσης σε κύριες συνιστώσες χρησιμοποιώντας τον πίνακα διακύμανσης είναι πως αν αλλάξει η κλίμακα μέτρησης των δεδομένων μας τότε αλλάζουν και οι κύριες συνιστώσες και η ερμηνεία τους. Για να το δούμε αυτό έστω ο πίνακας διακύμανσης που ακολουθεί και αφορά την ηλικία X_1 σε χρόνια και το βάρος X_2 σε κιλά.

$\mathbf{\Sigma} = \begin{bmatrix} 10 & 5 \\ 5 & 3 \end{bmatrix}$. Αν αντί για το βάρος σε κιλά χρησιμοποιήσουμε το βάρος σε τόνους (X_2^*) τότε ο πίνακας διακύμανσης γίνεται $\mathbf{\Sigma}^* = \begin{bmatrix} 10 & 0.005 \\ 0.005 & 0.0000003 \end{bmatrix}$. Οι ιδιοτιμές του $\mathbf{\Sigma}$ είναι 12.6033 και 0.3967 ενώ του $\mathbf{\Sigma}^*$ είναι 10.00000250 και 0.00000050. Είναι ξεκάθαρο πως και η όποια

ερμηνεία έχει αλλάξει δραματικά. Επομένως η ανάλυση σε κύριες συνιστώσες επηρεάζεται από τις μονάδες μέτρησης των μεταβλητών.

Ένα ακόμα μειονέκτημα είναι πως αν κάποια μεταβλητή έχει πολύ μεγαλύτερη διακύμανση από τις υπόλοιπες, αυτή θα τείνει να ταυτιστεί με την πρώτη κύρια συνιστώσα. Φανταστείτε τον πίνακα $\mathbf{\Sigma} = \begin{bmatrix} 50 & -1 & 0.1 \\ -1 & 1 & -0.5 \\ 0.1 & -0.5 & 0.3 \end{bmatrix}$. Οι ιδιοτιμές του πίνακα είναι 50.0206, 1.2425 και 0.0368 ενώ οι κύριες συνιστώσες που προκύπτουν είναι οι

$$\begin{aligned} Y_1 &= 0.9997X_1 - 0.0190X_2 + 0.0075X_3 \\ Y_2 &= -0.0204X_1 - 0.8840X_2 + 0.4669X_3 \\ Y_3 &= 0.0022X_1 + 0.4669X_2 + 0.8842X_3 \end{aligned}$$

Παρατηρείστε πως η πρώτη κύρια συνιστώσα σχεδόν ταυτίζεται με την πρώτη μεταβλητή η οποία είχε πολύ μεγαλύτερη διακύμανση από ότι οι υπόλοιπες.

Από τα παραπάνω εύκολα προκύπτει πως ένα τρόπος να ξεπεράσουμε τις κακές αυτές ιδιότητες της ανάλυσης σε κύριες συνιστώσες στον πίνακα διακύμανσης είναι να χρησιμοποιήσουμε τον πίνακα συσχετίσεων. Οι συσχετίσεις δεν αλλάζουν όταν αλλάξουν οι μονάδες μέτρησης ή η κλίμακα. Επίσης στην ουσία δίνουν ίδιο βάρος σε όλες τις μεταβλητές καθώς όλα τα στοιχεία της διαγωνίου είναι 1, και άρα τα προβλήματα που δημιουργούσε ο πίνακας διακύμανσης μπορούν να ξεπεραστούν.

Από την άλλη πλευρά πάντως, η γενικευμένη χρήση του πίνακα συσχετίσεων δεν ενδείκνυται καθώς η διαφορά στις διακυμάνσεις ενδέχεται να περιέχει πληροφορία πολύτιμη για το θέμα που εξετάζουμε. Ίσως δηλαδή κάποιες μεταβλητές να πρέπει να θεωρηθούν πως έχουν μεγαλύτερο βάρος εξαιτίας της και επομένως θέτοντας όλες τις μεταβλητές να έχουν το ίδιο βάρος χάνουμε χρήσιμη πληροφορία.

Κατά συνέπεια στην πράξη δεν είναι ξεκάθαρο ποιόν από τους δύο πίνακες πρέπει να χρησιμοποιούμε. Μια καλή στρατηγική είναι να αποφεύγουμε τον πίνακα διακύμανσης όταν υπάρχουν κάποιες μεταβλητές με πολύ μεγαλύτερη διακύμανση από ότι οι υπόλοιπες. Αν οι διακυμάνσεις διαφέρουν μεν αλλά είναι συγκρίσιμες (π.χ. αναφέρονται σε ίδιες μονάδες) τότε καλό είναι να χρησιμοποιούμε αυτή την πληροφορία. Εναλλακτικά θα μπορούσε κανείς να μετασχηματίσει τα δεδομένα του ώστε να κάνει τις διακυμάνσεις συγκρίσιμες.

Η συσχέτιση ανάμεσα στην *i* κύρια συνιστώσα Y_i και την *j* αρχική μεταβλητή X_j δίνεται από τον τύπο

$$\rho(Y_i, X_j) = \frac{a_{ij}\sqrt{\lambda_i}}{s_j^2} \ ,$$

όπου όπως πριν a_{ij} είναι ο συντελεστής της μεταβλητής X_j στην κύρια συνιστώσα Y_i και s_j^2 είναι η διακύμανση της μεταβλητής X_j. Η συσχέτιση αυτή αποτελεί ένα μέτρο του κατά πόσο

η συνιστώσα που προέκυψε σχετίζεται με τη μεταβλητή. Μπορεί κανείς εύκολα να δει πως αν $a_{ij}=0$ τότε δεν υπάρχει συσχέτιση, ενώ αν $a_{ij}=\pm 1$, τότε η συσχέτιση γίνεται ± 1

Συνήθως στην πράξη δεν διατηρούμε όλες τις κύριες συνιστώσες αλλά τις πρώτες m από αυτές και αγνοούμε τις υπόλοιπες. Σε μια τέτοια περίπτωση χάνουμε πληροφορία. Μπορούμε να βρούμε το ποσοστό της διακύμανσης της αρχικής μεταβλητής X_i που εξηγούμε με τη χρήση των πρώτων m κυρίων συνιστωσών ως $\frac{1}{s_i^2}\sum_{j=1}^{m}\lambda_j a_{ji}^{\ 2}$. Είναι ευνόητο πως το ποσοστό της διακύμανσης που εξηγούμε για κάθε μεταβλητή πρέπει να είναι σχετικά μεγάλο γιατί αλλιώς σημαίνει πως χάνουμε πληροφορία για τη μεταβλητή αυτή.

Ως προς την αναμενόμενη τιμή των κυρίων συνιστωσών παρατηρείστε πως

$$E(Y_i) = E(a_{i1}X_1 + a_{i2}X_2 + ... + a_{ik}X_k) =$$
$$= a_{i1}E(X_1) + a_{i2}E(X_2) + ... + a_{ik}E(X_k) = \mu_i$$

Γενικά δηλαδή η αναμενόμενη τιμή είναι διαφορετική του 0. Για αυτό πολλές φορές κεντροποιούμε τις αρχικές μεταβλητές να έχουν μέση τιμή 0 απλά αφαιρώντας τη μέση τιμή. Αυτό δεν έχει καμιά επίδραση στη διακύμανση απλά οι προκύπτουσες κύριες συνιστώσες έχουν μέση τιμή 0.

Σε αυτό το σημείο θα πρέπει να παρατηρήσουμε τα εξής

- Μέχρι τώρα μιλάμε γενικά για μεταβλητές και πουθενά δεν έχουμε μιλήσει για δεδομένα και τυχαία δείγματα.
- Δεν έχουμε κάνει καμιά υπόθεση για τον πληθυσμό και δεν υπάρχει κανένα μοντέλο. Η ανάλυση σε κύριες συνιστώσες είναι ένας μαθηματικός μετασχηματισμός των δεδομένων μας και τίποτα άλλο.
- Δεν υπάρχει επομένως κάποια στατιστική συμπερασματολογία μέχρι τώρα.

Ας υποθέσουμε λοιπόν πως έχουμε πια δεδομένα και συγκεκριμένα n παρατηρήσεις σε k μεταβλητές $(X_1, X_2, ..., X_k)$. Από αυτά τα δεδομένα υπολογίζουμε τον πίνακα διακύμανσης ή τον πίνακα συσχετίσεων με βάση όσα είπαμε προηγουμένως. Στη συνέχεια για τον επιλεγμένο πίνακα βρίσκουμε τις ιδιοτιμές και τα ιδιοδιανύσματα και επομένως βρίσκουμε τις κύριες συνιστώσες.

Μέχρι τώρα τα δεδομένα εισήλθαν στην ανάλυση μόνο για τον καθορισμό του πίνακα διακύμανσης (συσχέτισης) και πουθενά αλλού. Αφού μιλάμε πια για δειγματικό πίνακα διακύμανσης (συσχέτισης) αυτός εμπεριέχει κάποια μεταβλητότητα λόγω του δείγματος. Το ίδιο συμβαίνει και για τις ιδιοτιμές και τα ιδιοδιανύσματά του. Για να μπορέσει κανείς να προχωρήσει σε στατιστική συμπερασματολογία χρειάζεται να κάνει κάποιες υποθέσεις σχετικά με τον πληθυσμό από όπου προήλθε το δείγμα. Με αυτά τα προβλήματα θα ασχοληθούμε σε λίγο.

Παρατηρείστε πως αν δεν ενδιαφερόμαστε για στατιστική συμπερασματολογία η ανάλυση σε κύριες συνιστώσες είναι απλά ένας μετασχηματισμός των δεδομένων. Πως όμως θα μετασχηματίσουμε τα δεδομένα; Για κάθε μια παρατήρηση θα δημιουργήσουμε τόσες καινούριες μεταβλητές όσες και οι κύριες συνιστώσες που αποφασίσαμε να διατηρήσουμε. Για κάθε κύρια συνιστώσα θα υπολογίσουμε την τιμή της για την παρατήρηση χρησιμοποιώντας τις αντίστοιχες τιμές των αρχικών μεταβλητών και τους συντελεστές που έχουμε βρει. Θα δούμε σε λίγο αναλυτικά πως γίνεται αυτό στην πράξη.

7.5 Βήματα της Ανάλυσης Σε Κύριες Συνιστώσες

7.5.1 Έλεγχος συσχετίσεων

Άσχετα με το αν θα χρησιμοποιήσουμε τον πίνακα διακύμανσης ή τον πίνακα συσχετίσεων είναι σκόπιμο να ρίξουμε μια ματιά στον πίνακα συσχετίσεων και να δούμε αν οι αρχικές μας μεταβλητές έχουν συσχετίσεις ή όχι (αυτό γίνεται κυρίως γιατί από τον πίνακα διακύμανσης δεν είναι εύκολο να δούμε την ύπαρξη συσχετίσεων). Αν δεν υπάρχουν συσχετίσεις είναι άσκοπο να συνεχίσουμε. Μεταβλητές που εμφανίζονται ασυσχέτιστες με τις υπόλοιπες πρέπει να τις διώξουμε από την ανάλυση.

Τι εννοούμε όμως όταν λέμε να υπάρχουν συσχετίσεις; Εννοούμε πως η απόλυτη τιμή της συσχέτισης είναι μεγάλη. Αυτό δεν σημαίνει απαραίτητα πως είναι στατιστικά σημαντική, σύμφωνα με το αποτέλεσμα κάποιου ελέγχου υποθέσεων. Ακόμα και συσχετίσεις της τάξης του 0.10 τείνουν να είναι στατιστικά σημαντικές για μέτριου μεγέθους δείγματα (π.χ. 300 παρατηρήσεις). Για να είναι όμως οι συσχετίσεις ικανοποιητικές για να προχωρήσουμε σε ανάλυση σε κύριες συνιστώσες, θέλουμε να είναι της τάξης του 0.4 ή και μεγαλύτερες σε απόλυτη τιμή. Ένα μέτρο που μας επιτρέπει καλύτερα να συγκρίνουμε δύο σετ δεδομένων αλλά και να αξιολογήσουμε αν οι συσχετίσεις είναι 'ενδιαφέρουσες' είναι το

$$\phi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^{\,2} - p}{p(p-1)}},$$

όπου r_{ij} είναι το ij στοιχείο του πίνακα συσχετίσεων δηλαδή η συσχέτιση της X_i με τη X_j μεταβλητή. Το στατιστικό φ παίρνει τιμές κοντά στο 1 αν υπάρχουν μεγάλες συσχετίσεις, καθώς όλα τα r_{ij} πλησιάζουν σε απόλυτη τιμή τη μονάδα και άρα το άθροισμα των τετραγώνων τους είναι κοντά στο p^2 και άρα ο αριθμητής τείνει να είναι ίσος με τον παρονομαστή. Αν δεν υπάρχουν συσχετίσεις η τιμή θα είναι κοντά στο 0, καθώς μόνο τα p διαγώνια στοιχεία θα είναι 1, άρα το άθροισμα τετραγώνων θα είναι p και άρα ο αριθμητής θα μηδενιστεί. Στην πράξη τιμές πάνω από 0.4 θεωρούνται ικανοποιητικές.

Το αντίστοιχο μέτρο, στην περίπτωση που δουλεύουμε με τον πίνακα διακύμανσης, είναι το

$$\phi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} s_{ij}^{\ 2} - \sum_{j=1}^{p} s_{jj}^{\ 2}}{\sum_{i=1}^{p}\sum_{j\neq 1}^{p} s_{ii}s_{jj}}}$$

για το οποίο ισχύουν παρόμοια πράγματα.

Επομένως, ξεκινώντας την ανάλυση, θα ήταν χρήσιμο κανείς όχι απλά να δει αν οι συσχετίσεις είναι στατιστικά σημαντικά διάφορες του 0 αλλά αν είναι επαρκώς μεγάλες σε απόλυτη τιμή για να προχωρήσει

7.5.2 Επιλογή πίνακα που θα δουλέψουμε

Όπως είδαμε μπορούμε να χρησιμοποιήσουμε τον πίνακα διακύμανσης ή τον πίνακα συσχετίσεων. Μιλήσαμε προηγουμένως πως επιλέγουμε και με ποια κριτήρια. Πρέπει να γίνει σαφές ότι τα αποτελέσματα θα διαφέρουν ανάλογα με τον πίνακα που θα επιλέξουμε για αυτό η επιλογή είναι βασική για την αξιοποίηση των αποτελεσμάτων που θα προκύψουν.

7.5.3 Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων

Ανάλογα με τον πίνακα που διαλέξαμε να στηρίξουμε την ανάλυση υπολογίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματα. Κρατήστε στο νου σας πως τα ιδιοδιανύσματα που δίνουν τα στατιστικά πακέτα είναι κανονικοποιημένα, δηλαδή το άθροισμα τετραγώνων του είναι 1 και πως δεν είναι μοναδικά από την άποψη πως μπορούμε να τους αλλάξουμε πρόσημο σε όλα τα στοιχεία τους. Συνεπώς η λύση από στατιστικό πακέτο σε στατιστικό πακέτο μπορεί να διαφέρει ως προς τα πρόσημα.

7.5.4 Απόφαση για τον αριθμό των συνιστωσών που θα κρατήσουμε

Ίσως το πιο σημαντικό κομμάτι της ανάλυσης το οποίο δυστυχώς δεν έχει εύκολη και κοινώς αποδεκτή απάντηση. Κατ' αρχάς να διευκρινίσουμε πως επιλέγοντας λιγότερες κύριες συνιστώσες από όσες μεταβλητές είχαμε αρχικά, χάνουμε αναγκαστικά πληροφορία. Αυτό είναι το κόστος για το κέρδος μας να μειώσουμε τις διαστάσεις του προβλήματος. Συνήθως λοιπόν ενδιαφερόμαστε για κάποιον μικρότερο αριθμό συνιστωσών. Πόσες όμως; Στη βιβλιογραφία υπάρχουν πολλά κριτήρια τα οποία θα προσπαθήσουμε να περιγράψουμε. Αυτά είναι:

Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες. Σύμφωνα με αυτό το κριτήριο βάζουμε κάποιο όριο (π.χ. 80%) και διαλέγουμε τόσες συνιστώσες ώστε αθροιστικά να εξηγούν μεγαλύτερο ποσοστό από το στόχο που βάλαμε. Είναι πολύ απλό και εύκολο να το χρησιμοποιήσουμε αλλά δυστυχώς στην πράξη δεν δίνει τα καλύτερα αποτελέσματα, ιδίως αν

ο στόχος είναι αρκετά υψηλός. Επίσης δεν είναι ξεκάθαρο ποιο ποσοστό της διακύμανσης πρέπει να βάλουμε σαν στόχο.

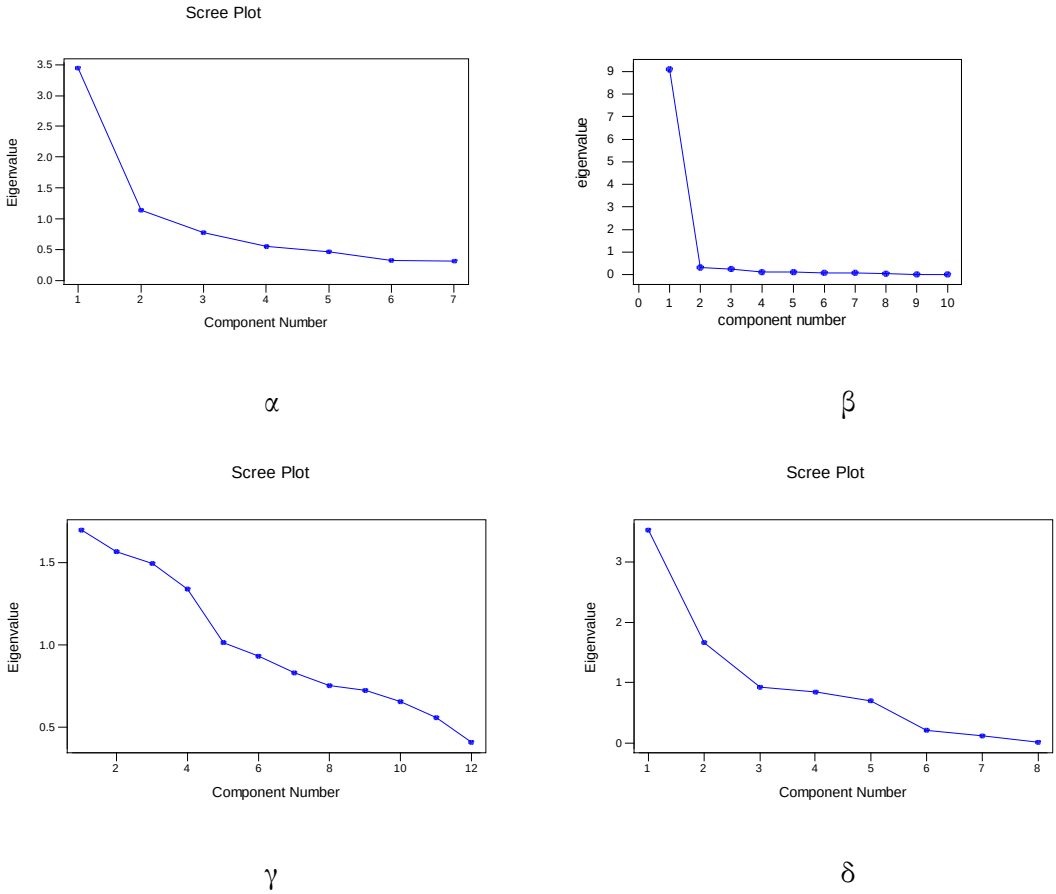
Κριτήριο του Kaiser. Έστω λ_j οι ιδιοτιμές μας . Το κριτήριο αυτό λεει να πάρουμε τόσες ιδιοτιμές όσες είναι μεγαλύτερες από $\overline{\lambda} = \sum_{j=1}^{k} \lambda_j$ δηλαδή μεγαλύτερες από τη μέση τιμή των ιδιοτιμών. Στην περίπτωση που δουλεύουμε με πίνακα συσχετίσεων ισχύει $\overline{\lambda} = 1$ και άρα διαλέγουμε τόσες συνιστώσες όσες και οι ιδιοτιμές μεγαλύτερες της μονάδας. Το κριτήριο στηρίζεται στην εξής απλή υπόθεση. Αν οι μεταβλητές είναι ασυσχέτιστες και άρα δεν υπάρχει καμιά δομή στα δεδομένα, τότε ο πίνακας συσχετίσεων είναι ο μοναδιαίος και όλες οι ιδιοτιμές είναι ίσες με 1 (δουλεύουμε με πίνακα συσχέτισης). Επομένως κάθε ιδιοτιμή μεγαλύτερη της μονάδας δείχνει την παρουσία κάποιας δομής στα δεδομένα μας.

Στην πράξη η υπόθεση αυτή είναι απλοϊκή καθώς ακόμα και αν δεν υπάρχει δομή και όλες οι ιδιοτιμές είναι 1 όταν δουλέψουμε με ένα δείγμα σίγουρα κάποιες από αυτές θα είναι μεγαλύτερες από 1 αφού το άθροισμά τους πρέπει να είναι *p*. Το κριτήριο συνήθως υπερεκτιμά τον αριθμό των συνιστωσών που χρειάζονται.

Ποσοστό της διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται. Όπως είδαμε πριν αν διατηρήσουμε *k* συνιστώσες χάνουμε κάποιο μέρος από την πληροφορία κάθε μεταβλητής και μπορούμε να βρούμε και το ποσοστό της διακύμανσης που ερμηνεύσουμε τελικά. Το κριτήριο αυτό διαλέγει τόσες συνιστώσες ώστε να ερμηνεύεται για κάθε μεταβλητή ένα υψηλό ποσοστό τουλάχιστον. Και πάλι το ποιο είναι αυτό το ποσοστό είναι υποκειμενικό. Επίσης μπορεί κάποια μεταβλητή να μην ερμηνεύεται σωστά και αυτό να οδηγήσει σε μεγάλο αριθμό συνιστωσών

Scree plot. Το scree plot είναι ένα γράφημα που έχει στον οριζόντιο άξονα των x τη σειρά και στον κάθετο άξονα των y την τιμή της κάθε ιδιοτιμής. Το κριτήριο αυτό προτείνει να πάρουμε τόσες συνιστώσες μέχρι το γράφημα να αρχίσει να γίνεται περίπου επίπεδο, στην ουσία μέχρι να διαπιστώσουμε ότι αρχίζει να αλλάζει η κλίση. Στα scree plot που ακολουθούν Γράφημα 7.1) μπορεί κανείς να δει τα προβλήματα που παρουσιάζει αυτή η μέθοδος. Στο γράφημα 7.1β είναι ξεκάθαρο πως θα διαλέξουμε μια μόνο συνιστώσα. Στο γράφημα 7.1α φαίνεται να διαλέγουμε μια συνιστώσα αλλά κάποιοι θα μπορούσαν να ισχυριστούν ότι πρέπει να πάρουμε 2. Στο γράφημα 7.1γ τα πράγματα φαίνονται να μην είναι καθόλου καθαρά, ενώ στο 7.1δ φαίνεται να έχουμε 2 φορές αλλαγή κλίσης. Από αυτά τα γραφήματα γίνεται σαφές πως δεν είναι καθόλου εύκολο να χρησιμοποιήσουμε το scree plot για να επιλέξουμε αριθμό συνιστωσών.

Κατ' αρχάς υπάρχει ένα υποκειμενικό κριτήριο για το που και αν αλλάζει η κλίση. Αφετέρου μερικές φορές δεν είναι καθόλου εύκολο να διακρίνει κανείς κάτι τέτοιο για αυτό το scree plot πρέπει να χρησιμοποιείται με προσοχή.



Γράφημα 7.1 Διάφορα scree plots.

Παραλλαγές του Scree plot. Μερικοί συγγραφείς με σκοπό να αποφύγουν το μειονέκτημα του scree plot ως προς την εύρεση του σημείου αλλαγής της κλίσης πρότειναν διάφορες μεθόδους για να βρει κανείς την αλλαγή κλίσης ξεκινώντας από εμπειρικές παρατηρήσεις φτάνοντας μέχρι τη χρήση γραμμικών μοντέλων. Δεν θα μπούμε σε λεπτομερή περιγραφή τέτοιων μεθόδων.

Η μέθοδος του σπασμένου ραβδιού (Broken Stick). Η μέθοδος αυτή στηρίζεται στην απλή παρατήρηση πως αν πάρουμε ένα ραβδί μεγέθους 1 μονάδας και το σπάσουμε τυχαία σε p κομμάτια τότε το k μεγαλύτερο από αυτά θα έχει αναμενόμενο μήκος $g_k = \frac{1}{p}\sum_{i=k}^{p}\left(\frac{1}{i}\right)$. Επομένως συγκρίνοντας την k ιδιοτιμή με αυτή την ποσότητα μπορεί κανείς να έχει μια εικόνα για τον αν οι ιδιοτιμές προήλθαν από έναν μοναδιαίο πίνακα συσχετίσεων ή

όχι. Το κριτήριο λοιπόν επιλέγει τόσες συνιστώσες όσο ισχύει $\frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i} > g_k$. Δεν μας ενδιαφέρει αν αργότερα ισχύσει ξανά η ανισότητα.

Η μέθοδος του Velicer. Η μέθοδος στηρίζεται στους συντελεστές μερικής συσχέτισης ανάμεσα στις αρχικές μεταβλητές όταν παραλείψουμε κάποιες συνιστώσες. Αν παραλείψουμε κάποια συνιστώσα που είναι χρήσιμη θα πρέπει οι συντελεστές αυτοί να αυξηθούν απότομα και επομένως καταλαβαίνουμε πως η συνιστώσα χρειάζεται. Με τη μέθοδο αυτή αρχίζουμε να 'διώχνουμε' μια τις συνιστώσες μέχρι να βρούμε πως δεν πρέπει να διώξουμε άλλη.

Κανονική προσέγγιση. Όπως θα δούμε στη συνέχεια αν μπορούμε να υποθέσουμε πως ο πληθυσμός μας ακολουθεί πολυμεταβλητή κανονική κατανομή, μπορούμε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης για ιδιοτιμές βασισμένοι στις δειγματικές ιδιοτιμές. Η ιδέα είναι πως δεν εμπιστευόμαστε το κριτήριο του Kaiser γιατί κάποιες ιδιοτιμές για λόγους τυχαίων κυμάνσεων μπορεί να εμφανιστούν μεγαλύτερες της μονάδας ενώ δεν είναι. Έτσι προσπαθούμε να διαχειριστούμε τη μεταβλητότητα φτιάχνοντας διαστήματα εμπιστοσύνης για τις ιδιοτιμές στηριζόμενοι στα ασυμπτωτικά αποτελέσματα της κανονικής κατανομής. Αν το 95% διάστημα εμπιστοσύνης για την *i* ιδιοτιμή δεν περιέχει το 1 και είναι μεγαλύτερο από αυτή την τιμή κρατάμε την αντίστοιχη κύρια συνιστώσα (υποθέτουμε πως δουλεύουμε με τον πίνακα συσχετίσεων).

Bootstrap. Η μεθοδολογία bootstrap βρίσκει ολοένα και περισσότερες εφαρμογές στη στατιστική καθώς μας επιτρέπει με επαναληπτική δειγματοληψία με επανάθεση να εκτιμήσουμε ποσότητες του πληθυσμού και κυρίως τα τυπικά σφάλματα των εκτιμητριών τους. Αν η προηγούμενη μέθοδος προσπαθούσε να φτιάξει διαστήματα εμπιστοσύνης για τις ιδιοτιμές στηριζόμενη σε ασυμπτωτικά αποτελέσματα από την κανονική κατανομή, η μέθοδος bootstrap φτιάχνει τα διαστήματα χωρίς να χρειάζεται να κάνει τέτοια υπόθεση. Για να γίνει αυτό δουλεύουμε ως εξής. Παίρνουμε ένα δείγμα ίσου μεγέθους με το πραγματικό από τα δεδομένα μας κάνοντας δειγματοληψία με επανάθεση ανάμεσα στις παρατηρήσεις μας (αυτό σημαίνει πως στο δείγμα που παίρνουμε κάποια παρατήρηση μπορεί να εμφανιστεί παραπάνω από μια φορά). Στη συνέχεια για αυτό το δείγμα φτιάχνουμε τον πίνακα διακύμανσης (συσχέτισης) και βρίσκουμε τις ιδιοτιμές. Αν επαναλάβουμε τη διαδικασία πολλές φορές έχουμε σχηματίσει μια σειρά από τιμές της κατανομής των ιδιοτιμών και άρα μπορούμε να εκτιμήσουμε από αυτές τις τιμές την τυπική απόκλιση των ιδιοτιμών. Έτσι κατασκευάζουμε διαστήματα εμπιστοσύνης και ελέγχουμε αν η τιμή που το κριτήριο του Kaiser θέτει σαν όριο ανήκει στο διάστημα. Κρατάμε μόνο τις ιδιοτιμές για τις οποίες όλο το διάστημα εμπιστοσύνης είναι πάνω από το όριο που το κριτήριο του Kaiser ορίζει. Είναι κατανοητό πως η μέθοδος απαιτεί μεγάλο υπολογιστικό φόρτο.

Cross Validation. Η μέθοδος αυτή στηρίζεται σε επαναληπτικούς υπολογισμούς, όπου κάθε φορά αγνοούμε κάποιες τιμές των δεδομένων μας και εξετάζουμε τη συμπεριφορά των συνιστωσών προσπαθώντας να προβλέψουμε τα δεδομένα που δεν χρησιμοποιήσαμε στην ανάλυση. Επαναλαμβάνοντας τη διαδικασία αυτή πολλές φορές, έχουμε ένα σκορ που μας δείχνει αν το μοντέλο με *k* συνιστώσες δίνει καλά αποτελέσματα. Έτσι συγκρίνοντας τα αποτελέσματα για διάφορες τιμές του *k* βρίσκουμε την τιμή για την οποία τα αποτελέσματα είναι τα καλύτερα

Έλεγχοι υποθέσεων. Αν μπορούμε να υποθέσουμε κανονικότητα του πληθυσμού ο Bartlett περιέγραψε έναν έλεγχο υπόθεσης για να ελέγξουμε αν οι τελευταίες *p-k* ιδιοτιμές είναι ίσες (και επομένως δεν πρέπει να τις χρησιμοποιήσουμε). Ο έλεγχος ελέγχει τη μηδενική υπόθεση

H_0: οι τελευταίες *p-k* ιδιοτιμές είναι ίσες έναντι της

H_1 : δεν είναι ίσες

Για αυτό το σκοπό χρησιμοποιεί την ελεγχοσυνάρτηση

$$\chi = -(n-1)\sum_{j=k+1}^{p}\ln(\lambda_j) + (n-1)(p-k)\ln\left(\frac{\sum_{j=k+1}^{p}\lambda_j}{p-k}\right),$$

η οποία ακολουθεί την κατανομή χ^2 με $\frac{1}{2}(p-k-1)(p-k+2)$ βαθμούς ελευθερίας.

Παρατηρείστε ότι καθώς οι έλεγχοι θα πρέπει να γίνουν ακολουθιακά αυτό έχει σαν αποτέλεσμα το επίπεδο σημαντικότητας να διαφέρει από το *a* που χρησιμοποιούμε για κάθε έλεγχο ξεχωριστά.

7.5.5 Εύρεση των συνιστωσών

Αυτό αποτελεί το πιο εύκολο ίσως κομμάτι, ιδιαίτερα στις μέρες μας που όλη τη δουλεία την κάνει ο υπολογιστής. Αρκεί να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα που επιλέξαμε για την ανάλυση, σύμφωνα με τη φασματική ανάλυση που είδαμε προηγουμένως.

7.5.6 Ερμηνεία των συνιστωσών

Αυτό το κομμάτι ίσως είναι από τα πιο δύσκολα της ανάλυσης και έχει κατηγορηθεί από πολλούς συγγραφείς. Αφού λοιπόν έχουμε κατασκευάσει τις συνιστώσες πρέπει να προσπαθήσουμε να τους δώσουμε κάποια ερμηνεία, ιδιαίτερα στις πρώτες. Αυτό εξυπηρετεί

τους σκοπούς της ανάλυσης καθώς ερμηνεύει τις συσχετίσεις ανάμεσα στις μεταβλητές μας αλλά και αν όλα πάνε καλά μπορούμε να ποσοτικοποιήσουμε κάποιες μη ποσοτικές μεταβλητές. Το τελευταίο είναι ιδιαίτερα χρήσιμο σε διάφορες επιστήμες όπως την ψυχολογία και το marketing.

Στα πλαίσια της ερμηνευτικότητας των συνιστωσών μπορεί κανείς να καταφύγει στην περιστροφή των αξόνων, τεχνική πιο γνωστή από την παραγοντική ανάλυση που θα συζητήσουμε αργότερα. Η περιστροφή δεν είναι παρά ο πολλαπλασιασμός του πίνακα των συντελεστών που βρήκαμε με έναν ορθογώνιο πίνακα. Από τους άπειρους ορθογώνιους πίνακες μπορούμε να διαλέξουμε κάποιον με βάση κριτήρια βελτιστοποίησης, όπως για παράδειγμα κάθε συνιστώσα να έχει όσο γίνεται λιγότερες μεταβλητές με μεγάλους συντελεστές. Η περιστροφή συνήθως καταλήγει σε κάθε συνιστώσα, οι μεταβλητές να χωρίζονται πιο έντονα σε σχέση με το πρόσημο τους, δηλαδή να υπάρχουν λίγες με μεγάλες απόλυτες τιμές ενώ οι υπόλοιπες να τείνουν να έχουν συντελεστή κοντά στο μηδέν. Αυτό βοηθά να αναγνωρίζουμε πιο εύκολα τη συνιστώσα, δηλαδή στην ευκολότερη ερμηνεία της.

7.5.7 Δημιουργία νέων μεταβλητών

Όπως είπαμε οι κύριες συνιστώσες είναι καινούριες μεταβλητές με κάποιες καλές ιδιότητες. Το ενδιαφέρον είναι πως μπορούμε για κάθε παρατήρηση να δημιουργήσουμε τόσες νέες μεταβλητές όσες και οι κύριες συνιστώσες που αποφασίσαμε να διατηρήσουμε, με σκοπό να χρησιμοποιήσουμε τις κύριες συνιστώσες γα περαιτέρω στατιστική ανάλυση. Για να γίνει αυτό αρκεί να αντικαταστήσουμε στον τύπο της κάθε συνιστώσας τις τιμές που η παρατήρηση είχε για κάθε μεταβλητή.

7.6 Αποτελέσματα για Ανάλυση σε Κύριες Συνιστώσες από Δείγμα

Όπως είπαμε και προηγουμένως συνήθως στην ανάλυση σε συνιστώσες περιοριζόμαστε σε απλό μαθηματικό μετασχηματισμό των δεδομένων. Αν όμως μπορούμε να υποθέσουμε πως ο πληθυσμός ακολουθεί την πολυμεταβλητή κανονική κατανομή τότε προκύπτουν μερικά αποτελέσματα σχετικά με τις ποσότητες του δείγματος που μας ενδιαφέρουν. Κατ' αρχάς ξέρουμε πως ο δειγματικός πίνακας διακύμανσης δεν είναι αμερόληπτος εκτός αν έχουμε διαιρέσει κάθε στοιχείο με *n-1* και όχι με *n*. Αν **S** είναι ο μεροληπτικός πίνακας διακύμανσης τότε και οι ιδιοτιμές και τα ιδιοδιανύσματα είναι μεροληπτικά, ενώ αν έχουμε χρησιμοποιήσει τον αμερόληπτο πίνακα, έστω $\mathbf{S}^*$, τότε είναι αμερόληπτες εκτιμήτριες των αντίστοιχων ποσοτήτων του πληθυσμού. Για αυτή την ενότητα ας συμβολίσουμε με ℓ_i τις δειγματικές ιδιοτιμές και με λ_i τις αντίστοιχες ιδιοτιμές του πληθυσμού. Κάτω από την υπόθεση της πολυμεταβλητής κανονικότητας θα ισχύει πως

$$Var(\ell_i) \approx \frac{2\lambda_i^2}{n}, \qquad Var(\ln \ell_i) = \frac{2}{n}$$

και $\sqrt{n}(\ell_i - \lambda_i) \underset{appr}{\sim} N(0, 2\lambda_i^2)$,

όπου $\underset{appr}{\sim}$ δηλώνει πως ακολουθεί ασυμπτωτικά την κατανομή αυτή.

Χρησιμοποιώντας αυτά τα αποτελέσματα μπορεί κάποιος να φτιάξει προσεγγιστικά διαστήματα εμπιστοσύνης για τις ιδιοτιμές του πληθυσμού. Ένα 95% προσεγγιστικό διάστημα εμπιστοσύνης είναι το

$$\frac{l_i}{1+1.96\sqrt{2/n}} < \lambda_i < \frac{l_i}{1-1.96\sqrt{2/n}}$$

το οποίο όμως συνήθως είναι μεγαλύτερο από 95% για λόγους που δεν θα ασχοληθούμε σε αυτές τις σημειώσεις. Μιλήσαμε προηγουμένως για τη χρήση αυτού του διαστήματος ως κριτηρίου για την επιλογή του αριθμού των συνιστωσών.

Γενικά η συνάρτηση πυκνότητας πιθανότητας των ιδιοτιμών είναι ιδιαίτερα περίπλοκη και επομένως δύσχρηστη. Επίσης μπορεί αν δειχθεί πως η συνδιακύμανση των δειγματικών ιδιοτιμών τείνει στο 0 όταν αυξάνει το μέγεθος του δείγματος, κάτι που υπονοεί πως για μεγάλα δείγματα οι ιδιοτιμές είναι ασυσχέτιστες.

Μερικές φορές είναι χρήσιμο να κάνουμε ελέγχους υποθέσεων για τα ιδιοδιανύσματα. Αυτό είναι χρήσιμο για να δούμε αν τα βάρη που δίνουμε σε κάθε μεταβλητή και επομένως η ερμηνεία στις συνιστώσες έχουν νόημα. Για παράδειγμα αν ο πίνακας συσχετίσεων περιέχει όλο θετικές συσχετίσεις η πρώτη κύρια συνιστώσα είναι ένα σταθμικός μέσος όρος. Έχουν οι μεταβλητές διαφορετικές σταθμίσεις ή είναι ένας απλός μέσος όρος; Γενικά για να ελέγξουμε αν ένα ιδιοδιάνυσμα είναι ίσο με ένα συγκεκριμένο ιδιοδιάνυσμα c, δηλαδή για να ελέγξουμε τη μηδενική υπόθεση

$H_0 : a_i = c$ έναντι της

$H_1 : a_i \neq c$

χρησιμοποιούμε την ελεγχοσυνάρτηση

$$A = (n-1)\left\{\ell_i c'\mathbf{S}^{-1}c + \frac{1}{\ell_i}c'\mathbf{S}c - 2\right\} \sim \chi^2(p-1),$$

όπου $\mathbf{S}$ ο δειγματικός πίνακας διακύμανσης (συσχετίσεων) και ℓ_i οι δειγματικές ιδιοτιμές. Η ελεγχοσυνάρτηση ακολουθεί κατανομή χ^2 με $p-1$ βαθμούς ελευθερίας.

Αν τα δεδομένα προέρχονται από πολυμεταβλητή κανονική κατανομή, οι κύριες συνιστώσες έχουν μια ενδιαφέρουσα γεωμετρική ερμηνεία. Συγκεκριμένα ο μετασχηματισμός των δεδομένων σε κύριες συνιστώσες αντιστοιχεί σε μετακίνηση των αξόνων κατά τη διεύθυνση της μεγαλύτερης διακύμανσης. Έτσι οι κύριες συνιστώσες αντιστοιχούν σε αυτούς τους καινούριους άξονες.

7.7 Μερικά Χρήσιμα Αποτελέσματα

Θα παρουσιάσουμε εν συντομία μερικά ενδιαφέροντα αποτελέσματα σχετικά με την ανάλυση σε κύριες συνιστώσες και κάποιες ειδικές της περιπτώσεις.

- Αν μια μεταβλητή είναι ασυσχέτιστη με τις υπόλοιπες καλό είναι να την αφαιρέσουμε από την ανάλυση, αφού αν παραμείνει κάποια από τις κύριες συνιστώσες θα ταυτιστεί μαζί της. Όταν δουλεύουμε με δεδομένα αυτό σημαίνει πως δεν έχει στατιστικά σημαντικές συσχετίσεις με τις υπόλοιπες και συνεπώς δεν έχει νόημα να την συμπεριλάβουμε στην ανάλυση.

- Αν δύο ιδιοτιμές προκύψουν ίδιες τότε αυτές αντιστοιχούν σε δύο όμοιες κύριες συνιστώσες κάτι που οδηγεί σε πλεονασμό. Φυσικά στην πράξη κάτι τέτοιο είναι σπάνιο. Αν λοιπόν συμβεί πρέπει να δούμε τα δεδομένα μας μήπως υπάρχει κάποιο πρόβλημα (π.χ. στήλες που επαναλαμβάνονται). Πρέπει να τονιστεί πως για δεδομένα από δείγμα έχει αποδειχτεί πως όλες οι ιδιοτιμές είναι διαφορετικές εκτός από συγκεκριμένες προβληματικές περιπτώσεις.

- Αν έχουμε μηδενικές ιδιοτιμές αυτό σημαίνει πως ο πίνακας που στηρίξαμε την ανάλυση δεν είναι πλήρους βαθμού και άρα κάποιες μεταβλητές είναι γραμμικά εξαρτημένες και πρέπει να τις διώξουμε. Στην πράξη δεν θα συναντήσουμε μηδενικές ιδιοτιμές αλλά πολύ μικρές, κοντά στο μηδέν, ιδιοτιμές. Αυτό υπονοεί ότι κάποιες μεταβλητές είναι σχεδόν γραμμικά εξαρτημένες. Αν αναλογιστεί κανείς πως τέτοιες ιδιοτιμές αντιστοιχούν σε συνιστώσες με σχεδόν μηδενική διακύμανση μπορούμε να τις αγνοήσουμε. Δηλαδή στην πράξη αφού δύο μεταβλητές θα παρέχουν την ίδια πληροφορία, όλη η πληροφορία θα πάει σε κάποια από τις πρώτες κύριες συνιστώσες και ότι μείνει θα πάει σε μια συνιστώσα με αμελητέα διακύμανση.

- Σε δύο διαφορετικά σετ δεδομένων μπορεί να πάρουμε τα ίδια ιδιοδιανύσματα ενώ οι ιδιοτιμές να αλλάξουν. Στην πράξη αυτό σημαίνει πως παίρνουμε τις ίδιες συνιστώσες αλλά σε κάθε περίπτωση η συνιστώσα εξηγεί άλλο ποσοστό της διακύμανσης. Συνεπώς δεν πρέπει να περιοριζόμαστε στα ιδιοδιανύσματα αλλά να κοιτάμε και τις ιδιοτιμές.

- Στη γενική περίπτωση που ο πίνακας συσχετίσεων έχει μόνο θετικά στοιχεία (όλες οι συσχετίσεις είναι θετικές) τότε η πρώτη κύρια συνιστώσα μπορεί να εκληφθεί σαν ένας σταθμικός μέσος όρος των μεταβλητών με σταθμίσεις τους αντίστοιχους συντελεστές. Επομένως σε τέτοιες περιπτώσεις μπορούμε να κατασκευάσουμε χρήσιμους δείκτες όπου οι σταθμίσεις έχουν επιλεγεί με έναν συγκεκριμένο τρόπο και όχι εμπειρικά.

- Η βασική ιδέα στην ανάλυση σε κύριες συνιστώσες είναι να γράψουμε τις συνιστώσες ως γραμμικό συνδυασμό των αρχικών μεταβλητών. Είναι εύκολο να δει κανείς πως ομοίως λύνοντας ως προς τις αρχικές μεταβλητές παίρνουμε $\mathbf{X=A'Y}$ επειδή ο πίνακας **A** είναι ορθογώνιος δηλαδή $\mathbf{A'=A^{-1}}$. Επομένως αν έχουμε τα σκορ των συνιστωσών μπορούμε εύκολα να βρούμε τα αρχικά δεδομένα.

7.8 Χρήση των Κυρίων Συνιστωσών

Όπως είπαμε και στην αρχή η μέθοδος των κυρίων συνιστωσών μπορεί να χρησιμοποιηθεί για διάφορους σκοπούς. Μερικοί από αυτούς είναι οι ακόλουθοι:

- Στη γραμμική παλινδρόμηση όταν οι ανεξάρτητες μεταβλητές είναι συσχετισμένες έχουμε το πρόβλημα της πολυσυγγραμικότητας, όπου πια οι εκτιμήτριες ελαχίστων τετραγώνων παύουν να είναι συνεπείς, οι διακυμάνσεις τους γίνονται πολύ μεγάλες και τα πρόσημα των συντελεστών δεν έχουν κάποια φυσική ερμηνεία. Αν αντί λοιπόν για τις αρχικές συσχετισμένες μεταβλητές χρησιμοποιήσουμε τις κύριες συνιστώσες (όχι απαραίτητα όλες) οι οποίες είναι ασυσχέτιστες το πρόβλημα της πολυσυγγραμικότητας έχει αποφευχθεί. Φυσικά δεν υπάρχει πια κάποια ερμηνεία των συντελεστών αλλά προβλέψεις μπορούν να γίνουν απλά μετασχηματίζοντας τα καινούρια δεδομένα σε κύριες συνιστώσες. Έχετε υπόψη σας ότι σε οικονομετρικές εφαρμογές οι κύριες συνιστώσες έχουν και ερμηνεία, ότι δηλαδή ποσοτικοποιούν κάποιες αφηρημένες έννοιες.

- Γενικά είναι δύσκολο κανείς να αναπαραστήσει γραφικά πολυδιάστατα δεδομένα. Αν λοιπόν αντί για τα αρχικά δεδομένα αναπαραστήσει γραφικά τις πρώτες κύριες συνιστώσες που ερμηνεύουν μεγάλο κομμάτι της μεταβλητότητας των δεδομένων επιτυγχάνει μια αξιόλογη οπτική παρουσίαση των δεδομένων

- Κοιτάζοντας τα σκορ των παρατηρήσεων στις κύριες συνιστώσες είναι μερικές φορές εύκολο να αποκτήσει κανείς μια ιδέα πως ομαδοποιούνται οι παρατηρήσεις. Αυτό έχει σχέση και με την ευκολότερη γραφική αναπαράσταση των δεδομένων που αναφέρθηκε αμέσως πριν.

- Data mining. Ένας καινούριος επιστημονικός τομέας που συνδυάζει την πληροφορική επιστήμη με τη στατιστική είναι το λεγόμενο data mining (εξόρυξη γνώσης). Η ιδέα είναι πως θα μπορέσουμε να εξάγουμε γνώση από τεράστιες βάσεις δεδομένων (όπως είναι πια οι βάσεις δεδομένων μεγάλων εταιρειών και οργανισμών). Από στατιστικής πλευράς ενδιαφερόμαστε στον να συμπυκνώσουμε την πληροφορία σε όσο γίνεται λιγότερες διαστάσεις και αυτό ακριβώς προσφέρει η ανάλυση σε κύριες συνιστώσες.

- Έλεγχος ποιότητας. Αν κάποιος παρατηρεί μια πληθώρα χαρακτηριστικών ενός προϊόντος με τη χρήση διαγραμμάτων ποιοτικού ελέγχου, είναι σχετικά δύσκολο να βρει πότε το προϊόν έχει βγει εκτός ελέγχου παρακολουθώντας τα πολλά επιμέρους χαρακτηριστικά. Αν όμως συμπυκνώσει την πληροφορία σε κάποιες κύριες συνιστώσες αυτόματα η δουλεία αυτή γίνεται πιο εύκολη.

7.9 Παραλλαγές της Μεθόδου

Πριν ολοκληρώσουμε την περιγραφή της μεθόδου και περάσουμε σε ένα λεπτομερές παράδειγμα θα αναφέρουμε εν συντομία μερικά ακόμα αποτελέσματα και ιδέες της ανάλυσης σε κύριες συνιστώσες για τον ενδιαφερόμενο αναγνώστη.

Κατ' αρχάς όλη η μέθοδος (που είναι στην ουσία ένας μαθηματικός μετασχηματισμός των δεδομένων) βασίστηκε στη φασματική ανάλυση ενός τετραγωνικού πίνακα όπως ο πίνακας συσχετίσεων ή ο πίνακας διακυμάνσεων. Επομένως μπορεί κανείς να ξεκινήσει από έναν τέτοιον πίνακα και να κάνει ανάλυση σε κύριες συνιστώσες άσχετα με τη μορφή που έχουν τα δεδομένα του. Ο πίνακας πρέπει να είναι συμμετρικός και τετραγωνικός. Για παράδειγμα ένας πίνακας συνάφειας όπου έχει για γραμμές και στήλες κάποιες δίτιμες μεταβλητές μπορεί να χρησιμοποιηθεί για την ανάλυση.

Φανταστείτε πως σε μια τάξη οι μαθητές γράφουν ένα διαγώνισμα με 10 ερωτήσεις πολλαπλών επιλογών και τα δεδομένα είναι 0 και 1 ανάλογα με το αν ο μαθητής απάντησε σωστά ή όχι. Χρησιμοποιώντας ένα κατάλληλο μέτρο συσχέτισης για τέτοιου είδους δεδομένα (π.χ. συντελεστής συνάφειας) μπορεί να προχωρήσει κανείς σε ανάλυση σε κύριες συνιστώσες για να διαπιστώσει τις σχέσεις ανάμεσα στις ερωτήσεις και να βρει βέλτιστα βάρη για τη βαθμολογία.

Έχουν αναπτυχθεί επίσης μέθοδοι ανθεκτικές στην ύπαρξη παράξενων τιμών στα δεδομένα (outliers). Οι μέθοδοι αυτοί δεν στηρίζονται στον πίνακα διακύμανσης ή συσχέτισης αλλά σε αντίστοιχους πίνακες που έχουν καλές ιδιότητες να μην επηρεάζονται από ακραίες τιμές. Σε παρόμοια θέματα έχει ερευνηθεί το θέμα της επίδρασης κάποιων μεμονωμένων τιμών στις κύριες συνιστώσες με παρόμοιο τρόπο όπως εξετάζουμε την ύπαρξη παρατηρήσεων με μεγάλη επίδραση στη γραμμική παλινδρόμηση.

Επίσης πρόσφατα αναπτύχθηκαν μεθοδολογίες που καταλήγουν σε κύριες συνιστώσες με ακέραιους συντελεστές. Τέτοιες κύριες συνιστώσες έχουν το πλεονέκτημα ότι η ερμηνεία τους είναι πιο εύκολη. Για να επιτευχθεί αυτό θυσιάζεται η μηδενική συσχέτιση μεταξύ των συνιστωσών αλλά αυτό στην πράξη δεν έχει ιδιαίτερη σημασία καθώς οι συνιστώσες που προκύπτουν έχουν απλά μη μηδενική συσχέτιση συνήθως κάτω από 0.05 σε απόλυτη τιμή.

Μια γενίκευση της ιδέας των κυρίων συνιστωσών είναι η ανάλυση σε κανονικές συσχετίσεις, όπου εκεί προσπαθούμε να βρούμε και να ερμηνεύσουμε τις αλληλεξαρτήσεις ανάμεσα σε δύο σύνολα μεταβλητών και όχι ανάμεσα σε απλές τυχαίες μεταβλητές όπως κάναμε στην ανάλυση σε κύριες συνιστώσες.

Τέλος καλό είναι να έχει υπόψη του κάποιος πως αν ο πληθυσμός δεν είναι πολυμεταβλητός κανονικός τότε η στατιστική συμπερασματολογία που αναφέρθηκε δεν είναι σωστή και πρέπει να χρησιμοποιείται με προσοχή.

7.10 Case Study: Αποτελέσματα Επτάθλου (Ολυμπιακοί αγώνες, Λος Άντζελες 1984)

Ας εξετάσουμε λοιπόν ένα παράδειγμα όσο γίνεται πληρέστερα. Τα δεδομένα αφορούν τις επιδόσεις 26 αθλητριών του επτάθλου στα 7 αγωνίσματα κατά τη διάρκεια της Ολυμπιάδας του Los Angeles το 1984. Τις επιδόσεις μπορείτε να τις δείτε στον πίνακα 7.1 που ακολουθεί. Το έπταθλο είναι ένα από τα πιο απαιτητικά αθλήματα του κλασικού αθλητισμού καθώς απαιτεί από τις αθλήτριες μια ποικιλία προσόντων και δυνατοτήτων. Τα αθλήματα περιλαμβάνουν δρόμους (100 μέτρα με εμπόδια, 200 μέτρα και 800 μέτρα), ρίψεις (σφαιροβολία και ακοντισμό) καθώς και άλματα (άλμα εις μήκος και άλμα εις ύψος). Η σειρά των αγωνισμάτων είναι αυτή που βλέπετε στον πίνακα 7.1.

Αθλήτρια	100 μ με εμπόδια (sec)	Άλμα εις ύψος (m)	Σφαιροβολία (m)	200 μέτρα (sec)	Άλμα εις μήκος (m)	Ακοντισμός (m)	800 μέτρα (sec)
1	13.87	1.70	13.11	25.44	6.23	45.42	134.31
2	14.03	1.79	13.05	24.39	6.22	45.18	130.90
3	14.79	1.55	10.71	25.66	5.76	0.00	141.59
4	13.48	1.82	13.23	23.93	6.01	48.10	129.49
5	13.25	1.88	13.77	23.34	6.82	41.90	125.08
6	14.31	1.76	12.96	25.01	6.01	43.30	138.84
7	13.25	1.94	14.23	24.27	6.02	51.12	134.35
8	13.97	1.85	14.35	24.54	6.10	37.58	128.62
9	16.62	1.64	12.24	25.44	5.88	44.40	144.30
10	14.10	1.82	15.33	24.86	6.13	45.14	128.83
11	13.48	1.58	13.85	23.95	6.10	52.12	151.21
12	13.23	1.70	14.68	23.31	6.11	44.48	127.90
13	13.59	1.79	14.35	24.60	6.38	40.78	134.16
14	12.85	1.91	14.13	23.12	7.10	44.98	131.75
15	13.64	1.73	12.83	25.29	6.35	47.42	139.61
16	13.48	1.70	14.49	24.40	6.12	44.12	130.96
17	13.75	1.88	13.48	25.24	5.99	41.28	135.57
18	13.94	1.67	12.40	24.43	5.74	41.08	137.53
19	12.86	1.82	14.34	23.70	6.49	41.30	131.22
20	13.75	1.76	14.49	25.20	6.03	44.42	134.95
21	13.96	1.70	12.97	25.09	5.90	49.02	135.18
22	13.73	1.82	12.07	24.48	6.08	35.42	142.19
23	14.06	1.79	12.69	26.13	5.65	52.58	142.42
24	14.04	1.73	14.96	25.28	6.11	49.00	138.40
25	13.57	1.82	13.91	24.18	6.20	43.46	134.96
26	13.64	1.82	14.26	23.83	5.99	45.12	151.84

Πίνακας 7.1. Τα δεδομένα: Οι επιδόσεις των 26 αθλητριών που έλαβαν μέρος στο έπταθλο στην Ολυμπιάδα του Λος Άντζελες το 1984

Ο πίνακας 7.2 περιέχει τις συσχετίσεις ανάμεσα στις μεταβλητές μας. Μόνο τα στοιχεία της κάτω διαγωνίου δίνονται. Μπορούμε να παρατηρήσουμε πως οι συσχετίσεις είναι μέτριες, δηλαδή υπάρχουν κάποιες μεταβλητές με ισχυρές συσχετίσεις (π.χ. τα 100 μέτρα με εμπόδια και τα 200 μέτρα) ενώ για κάποιες άλλες μεταβλητές οι συσχετίσεις είναι σχετικά χαμηλές (π.χ. ο ακοντισμός έχει υψηλή συσχέτιση μόνο με τη σφαιροβολία). Αν υπολογίσουμε το στατιστικό φ αυτό έχει τιμή 0.42, κάτι που επαληθεύει ότι οι συσχετίσεις χωρίς να είναι έντονες είναι τουλάχιστον ενδιαφέρουσες.

	100 μ με εμπόδια (sec)	**Άλμα εις ύψος (m)**	**Σφαιρο-βολία (m)**	**200 μέτρα (sec)**	**Άλμα εις μήκος (m)**	**Ακοντι-σμός (m)**	**800 μέτρα (sec)**
100 μ με εμπόδια (sec)	1						
Άλμα εις ύψος (m)	-0.493	1					
Σφαιροβολία (m)	-0.500	0.441	1				
200 μέτρα (sec)	0.629	-0.374	-0.433	1			
Άλμα εις μήκος (m)	-0.527	0.461	0.391	-0.621	1		
Ακοντισμός (m)	-0.252	0.331	0.531	-0.170	0.135	1	
800 μέτρα (sec)	0.368	-0.442	-0.384	0.328	-0.445	-0.036	1

Πίνακας 7.2. Πίνακας Συσχετίσεων για τα δεδομένα του επτάθλου

Το επόμενο δίλημμα είναι ποιόν από τους δύο πίνακες (διακύμανσης και συσχέτισης) θα χρησιμοποιήσουμε. Θα προχωρήσουμε στην ανάλυση και με τους δύο πίνακες για να συγκρίνουμε τα αποτελέσματά τους. Αν κάποιος όμως κοιτάξει τον πίνακα διακυμάνσεων (Πίνακας 7.3) παρατηρεί αφενός ότι οι μονάδες είναι διαφορετικές (κάποιες μεταβλητές μετριούνται σε μέτρα ενώ κάποιες σε δευτερόλεπτα) και αφετέρου το μέγεθος των διακυμάνσεων διαφέρει πάρα πολύ. Ο ακοντισμός έχει διακύμανση 93 περίπου ενώ στο άλμα σε ύψος η διακύμανση είναι μόλις 0.009. Αυτό αποτελεί μια πολύ ισχυρή ένδειξη ότι η χρήση του πίνακα διακυμάνσεων δεν είναι καλή ιδέα.

Ας ξεκινήσουμε λοιπόν την ανάλυση με τη χρήση του πίνακα διακυμάνσεων. Στον πίνακα 7.4 μπορεί κάποιος να δει τις ιδιοτιμές του πίνακα διακύμανσης που είναι 94.160, 45.308, 0.959, 0.438, 0.200, 0.050 και 0.005. Παρατηρείστε ότι οι 2 πρώτες είναι πολύ μεγάλες σε σχέση με τις υπόλοιπες και επομένως διαλέγοντας κάποιος δύο συνιστώσες καταφέρνει να εξηγεί το 99% της συνολικής διακύμανσης. Αν διαλέξουμε μόνο μια συνιστώσα εξηγούμε το 66,7% . Στον πίνακα 7.4 μπορούμε να δούμε το ποσοστό που κάθε φορά επιτυγχάνουμε με συγκεκριμένο αριθμό συνιστωσών στη στήλη 'Αθροιστικό ποσοστό της διακύμανσης'

	100 μ με εμπόδια (sec)	**Άλμα εις ύψος (m)**	**Σφαιροβολία (m)**	**200 μέτρα (sec)**	**Άλμα εις μήκος (m)**	**Ακοντισμός (m)**	**800 μέτρα (sec)**
100 μ με εμπόδια (sec)	0.5085						
Άλμα εις ύψος (m)	-0.0337	0.0091					
Σφαιροβολία (m)	-0.3739	0.0443	1.0980				
200 μέτρα (sec)	0.3493	-0.0278	-0.3532	0.6073			
Άλμα εις μήκος (m)	-0.1166	0.0137	0.1271	-0.1501	0.0963		
Ακοντισμός (m)	-1.7400	0.3065	5.3859	-1.2833	0.4050	93.6593	
800 μέτρα (sec)	1.7624	-0.2845	-2.7019	1.7199	-0.9275	-2.3498	45.1422

Πίνακας 7.3. Πίνακας Διακύμανσης για τα δεδομένα του επτάθλου

Παρατηρείστε πως στον αρχικό πίνακα διακυμάνσεων οι διακυμάνσεις του ακοντισμού και των 800 μέτρων είναι πολύ μεγαλύτερες από τις υπόλοιπες και επομένως υπάρχει η υποψία ότι οι δύο συνιστώσες σχεδόν ταυτίζονται με αυτές τις μεταβλητές.

	Πίνακας Συσχετίσεων			**Πίνακας Διακύμανσης**		
Α/Α	**Ιδιοτιμή**	**Ποσοστό της διακύμανσης**	**Αθροιστικό Ποσοστό της διακύμανσης**	**Ιδιοτιμή**	**Ποσοστό της διακύμανσης**	**Αθροιστικό Ποσοστό της διακύμανσης**
1	3.4392	0.491	0.491	94.160	0.667	0.667
2	1.1408	0.163	0.654	45.308	0.321	0.988
3	0.7811	0.112	0.766	0.959	0.007	0.995
4	0.5461	0.078	0.844	0.438	0.003	0.998
5	0.4580	0.065	0.909	0.200	0.001	1.000
6	0.3248	0.046	0.956	0.050	0.000	1.000
7	0.3100	0.044	1.000	0.005	0.000	1.000

Πίνακας 7.4. Οι ιδιοτιμές και τα ποσοστά της διακύμανσης που εξηγούνται με την επιλογή συγκεκριμένου αριθμού συνιστωσών

Στον πίνακα 7.5 που ακολουθεί μπορεί κανείς να δει τις κύριες συνιστώσες που προέκυψαν. Τα στοιχεία του πίνακα είναι οι συντελεστές. Επομένως η πρώτη κύρια συνιστώσα είναι η

Υ_1 = 0.019802 (100μ) -0.003443 (ύψος) -0.059340 (σφαιροβολία) + 0.014941 (200 μ.) - 0.004937 (μήκος) -0.996535 (ακοντισμός) + 0.052393 (800μ)

Παρατηρείστε πως η κύρια συνιστώσα δεν έχει μονάδες. Ακόμα και αν έχουμε χρησιμοποιήσει τον πίνακα διακυμάνσεων όπου υπήρχαν μονάδες το αποτέλεσμα πρέπει να το

βλέπουμε σαν αριθμό χωρίς μονάδες (δεν έχει νόημα να προσθέσουμε δευτερόλεπτα με μέτρα!)

Κύρια Συνιστώσα	**100 μ με εμπόδια (sec)**	**Άλμα εις ύψος (m)**	**Σφαιροβολία (m)**	**200 μέτρα (sec)**	**Άλμα εις μήκος (m)**	**Ακοντισμός (m)**	**800 μέτρα (sec)**
1	0.019802	-0.003443	-0.059340	0.014941	-0.004937	-0.996535	0.052393
2	0.037744	-0.005976	-0.054564	0.037455	-0.020285	0.057006	0.995238
3	-0.491883	0.024709	0.616033	-0.588679	0.148971	-0.051861	0.080738
4	-0.319610	0.011340	-0.780255	-0.524139	0.114575	0.031103	-0.010309
5	-0.805739	0.026062	-0.071453	0.583657	-0.065764	-0.002578	0.003639
6	-0.061431	-0.066489	0.008981	-0.190797	-0.977351	-0.000086	-0.010312
7	-0.033084	-0.997052	0.004458	0.007154	0.068589	0.002106	-0.003479

Πίνακας 7.5. Οι κύριες συνιστώσες που προκύπτουν από την ανάλυση με τον πίνακα διακυμάνσεων

Επομένως για κάθε αθλήτρια μπορούμε να βρούμε την τιμή της στη συγκεκριμένη συνιστώσα απλά χρησιμοποιώντας την τιμή της σε κάθε μεταβλητή (αγώνισμα). Στον πίνακα έχουμε ξεχωρίσει με σκίαση τους συντελεστές που είναι μεγαλύτεροι σε απόλυτη τιμή από 0.30. Αυτό είναι μια συνηθισμένη τεχνική για να μπορούμε να δούμε εύκολα ποιες μεταβλητές έχουν μεγάλους συντελεστές σε κάθε συνιστώσα και επομένως αν μπορούμε να ερμηνεύσουμε τη συνιστώσα αυτή. Αυτό συμβαίνει γιατί μικροί συντελεστές μπορεί απλά να οφείλονται στην τυχαιότητα του δείγματος και επομένως να μην αναδεικνύουν κάποια πραγματική διάσταση του πληθυσμού. Σε πολλά στατιστικά πακέτα παρέχεται η δυνατότητα να εμφανίζονται στην οθόνη μόνο οι συντελεστές που είναι μεγαλύτεροι από κάποια τιμή.

Παρατηρείστε λοιπόν πως οι δύο πρώτες συνιστώσες σχεδόν ταυτίζονται με τον ακοντισμό και τα 800 μέτρα αντίστοιχα (θυμηθείτε για μια ακόμα φορά πως τα πρόσημα δεν παίζουν κάποιο ρόλο). Αυτές ήταν οι δύο μεταβλητές με τη μεγαλύτερη διακύμανση. Το ίδιο συμβαίνει και με τις δύο τελευταίες συνιστώσες οι οποίες σχεδόν ταυτίζονται με το άλμα εις μήκος και το άλμα εις ύψος αντίστοιχα. Αυτές είναι και οι δυο μεταβλητές με τις μικρότερες διακυμάνσεις. Αυτό δηλαδή που βλέπουμε είναι πως αν στηρίξουμε την ανάλυση στον πίνακα διακυμάνσεων τότε τα αποτελέσματα δεν έχουν κάποιο ενδιαφέρον και αυτό οφείλεται στις μεγάλες διαφορές των μεταβλητών ως προς τις διακυμάνσεις τους.

Τέλος θυμηθείτε πως το άθροισμα τετραγώνων των συντελεστών κάθε συνιστώσας είναι ίσο με 1, δηλαδή για την πρώτη συνιστώσα έχουμε πως

$$(0.019802)^2 + (-0.003443)^2 + (-0.059340)^2 + (0.014941)^2 +$$
$$+(-0.004937)^2 + (-0.996535)^2 + (0.052393)^2 = 1$$

Ας επαναλάβουμε την ανάλυση με τον πίνακα συσχετίσεων. Παρατηρώντας τις ιδιοτιμές στον πίνακα 7.4 διαπιστώνουμε πως τώρα οι 2 πρώτες συνιστώσες εξηγούν μόλις το 65% της συνολικής διακύμανσης, ενώ οι 3 ανεβαίνουν στο 76%. Δηλαδή τα αποτελέσματα δεν

είναι τόσο θεαματικά όσο πριν αλλά ελπίζουμε πως οι συνιστώσες που προκύπτουν είναι ερμηνεύσιμες. Ο πίνακας 7.6α περιέχει τις συνιστώσες που προέκυψαν. Μπορεί κάποιος να παρατηρήσει πως τώρα τα πράγματα είναι καλύτερα. Οι μεταβλητές δείχνουν να μπερδεύονται μεταξύ τους κάτι που σημαίνει πως κάθε συνιστώσα εξαρτάται από πολλές μεταβλητές. Για παράδειγμα παρατηρείστε ότι η πρώτη συνιστώσα έχει θετικό πρόσημο στους δρόμους και αρνητικό στα υπόλοιπα αγωνίσματα. Κρατήστε αυτή την παρατήρηση προς το παρόν.

Δεν θα προσπαθήσουμε να δώσουμε ακόμα κάποια ερμηνεία στις συνιστώσες καθώς αυτό εμπεριέχει έναν υποκειμενικό χαρακτήρα. Για παράδειγμα κάποιος μπορεί να πει πως η πρώτη συνιστώσα είναι ένα κοντράστ ανάμεσα στους δρόμους και τα λοιπά αγωνίσματα. Συνήθως οι άνθρωποι προσπαθούν να δουν τα αποτελέσματα με βάση τις προσωπικές τους προκαταλήψεις δηλαδή να ανακαλύψουν στα αποτελέσματα αυτό που ήθελαν ή τους συμφέρει. Δηλαδή κάποιος μπορεί να δώσει την παραπάνω ερμηνεία στη συνιστώσα και να την εξηγήσει ότι αυτοί που είναι καλοί στους δρόμους είναι δεν είναι καλοί στις ρίψεις που απαιτούν δύναμη και σωματικό όγκο (αν και η σύγχρονη αθλητική επιστήμη διαφωνεί με αυτό). Όπως θα δούμε σε λίγο, αν και μια τέτοια ερμηνεία φαίνεται λογική σε πολλούς, τα πράγματα δεν είναι έτσι

Συνεπώς για το παράδειγμα αυτό ο πίνακας συσχετίσεων μοιάζει να είναι η καλύτερη επιλογή.

Κύρια Συνιστώσα	**100 μ με εμπόδια (sec)**	**Άλμα εις ύψος (m)**	**Σφαιροβολία (m)**	**200 μέτρα (sec)**	**Άλμα εις μήκος (m)**	**Ακοντισμός (m)**	**800 μέτρα (sec)**
1	0.429654	-0.391738	-0.401327	0.407340	-0.408511	-0.240791	0.331796
2	-0.082156	-0.066687	-0.377340	-0.230331	0.316381	-0.770490	-0.315616
3	-0.292121	-0.381523	-0.073663	-0.511351	0.159719	0.036265	0.689474
4	-0.100602	0.759488	-0.491387	0.113055	0.068449	0.019332	0.392117
5	-0.655432	0.063855	0.105759	0.030354	-0.693881	-0.267612	-0.033737
6	0.193855	0.226798	0.661771	0.162150	0.128028	-0.522687	0.396430
7	-0.496006	-0.253373	-0.023201	0.692493	0.452729	0.043106	0.053611

Πίνακας 7.6α. Οι κύριες συνιστώσες που προκύπτουν από την ανάλυση με τον πίνακα συσχετίσεων

Είπαμε πριν πως η ερμηνεία της πρώτης συνιστώσας που δώσαμε (κοντράστ ανάμεσα στους δρόμους και τα λοιπά αγωνίσματα) και η όποια εξήγηση της φάσκει στο εξής απλό σημείο. Στους δρόμους οι καλές επιδόσεις είναι οι μικρές επιδόσεις (μικρότερος χρόνος τερματισμού λοιπόν) ενώ στις ρίψεις και τα άλματα οι μεγάλες επιδόσεις. Συνεπώς αυτή η διαφορά που βλέπουμε στα πρόσημα δεν σημαίνει κοντράστ αλλά μπορεί κανείς να πει πως είναι ένας σταθμικός μέσος όρος των επιδόσεων των αθλητριών. Για να γίνει αυτό πιο σαφές θα μπορούσε κανείς να αλλάξει τις τιμές των μεταβλητών 100μ με εμπόδια, 200μ και 800μ έτσι ώστε οι μεγάλες τιμές να δείχνουν μεγαλύτερη απόδοση. Αυτό μπορεί να γίνει είτε απλά αλλάζοντας πρόσημο, είτε παίρνοντας μια άλλη συνάρτηση των δεδομένων όπως 1/x για παράδειγμα. Κάτω λοιπόν από αυτή την πιο λογική προσέγγιση η πρώτη κύρια συνιστώσα είναι ένα μέτρο της ικανότητας του αθλητή και μιλώντας γενικά μας επιτρέπει να

ποσοτικοποιήσουμε την αθλητική ικανότητα του αθλητή, ποσότητα μη μετρήσιμη με αντικειμενικά μέτρα (βέβαια και τώρα θα μπορούσε κανείς να διαμαρτυρηθεί πως δεν έχει μετρηθεί αντικειμενικά, αφού η ερμηνεία που δίνουμε στη συνιστώσα είναι μάλλον υποκειμενική).

Προχωράμε λοιπόν αφού αλλάξαμε τα πρόσημα στους 3 δρόμους. Ως προς τον πίνακα συσχετίσεων αλλάζουν μόνο τα πρόσημα των συσχετίσεων αυτών των μεταβλητών με τις υπόλοιπες, ενώ οι ιδιοτιμές παραμένουν οι ίδιες. Στα ιδιοδιανύσματα αλλάζουν απλά τα πρόσημα που αφορούν τις συγκεκριμένες μεταβλητές και επομένως ο πίνακας με τους συντελεστές των συνιστωσών είναι αυτό που βλέπεται στον πίνακα 7.6β

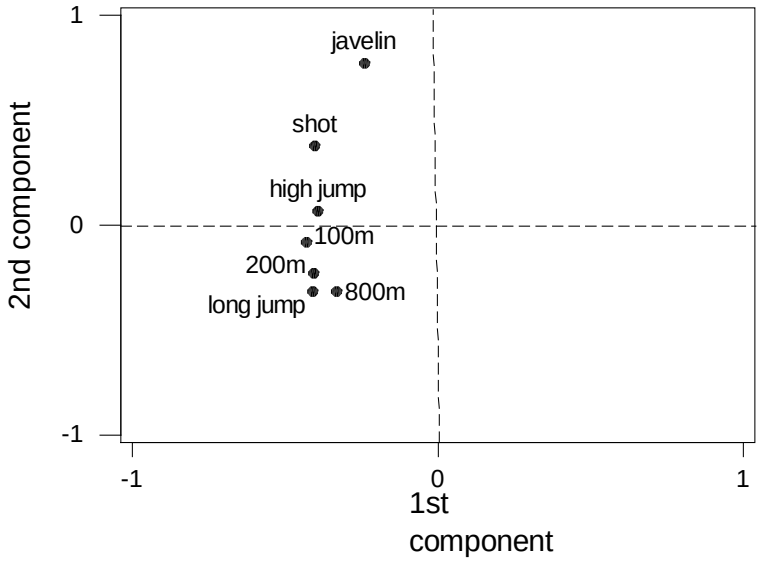
Ο πίνακας 7.6β προσφέρεται καλύτερα για προσπάθεια ερμηνείας των συνιστωσών. Υπενθυμίζουμε ότι τα πρόσημα αυτά καθαυτά δεν έχουν σημασία. Έτσι για την πρώτη κύρια συνιστώσα μπορεί κάποιος να της δώσει την ερμηνεία μιας γενικής αθλητικής ικανότητας αφού είναι ένας σταθμικός μέσος όρος των αθλημάτων. Είναι ευνόητο πως μια τέτοια ερμηνεία είναι πολύ ενδιαφέρουσα καθώς επιτρέπει στους αναλυτές να δουν πόσο σημαντικό είναι κάθε αγώνισμα στη διαμόρφωση του τελικού αποτελέσματος. Παρατηρείστε ότι τους μικρότερους συντελεστές (σε απόλυτη τιμή) τους έχουν ο ακοντισμός και τα 800 μέτρα, κάτι που και εμπειρικά είναι γνωστό στους φίλους του κλασικού αθλητισμού. Τα 800 μέτρα είναι το τελευταίο αγώνισμα όπου συνήθως έχουν κριθεί οι περισσότερες θέσεις, ενώ στον ακοντισμό, επειδή είναι ένα αγώνισμα με πολλά τεχνικά χαρακτηριστικά συνήθως οι αθλήτριες δεν τα πάνε καλά. Μια παρατήρηση που ισχύει γενικά είναι πως όταν όλα τα στοιχεία ενός πίνακα συσχέτισης είναι θετικά τότε η πρώτη κύρια συνιστώσα προκύπτει ως σταθμικός μέσος των μεταβλητών.

Κύρια Συνιστώσα	**100 μ με εμπόδια (sec)**	**Άλμα εις ύψος (m)**	**Σφαιρο-βολία (m)**	**200 μέτρα (sec)**	**Άλμα εις μήκος (m)**	**Ακοντι-σμός (m)**	**800 μέτρα (sec)**
1	-0.429654	-0.391738	-0.401327	-0.407340	-0.408511	-0.240791	-0.331796
2	-0.082156	0.066686	0.377340	-0.230331	-0.316381	0.770490	-0.315615
3	-0.292121	0.381523	0.073663	-0.511351	-0.159720	-0.036265	0.689474
4	0.100602	0.759488	-0.491387	-0.113056	0.068450	0.019332	-0.392117
5	0.655432	0.063855	0.105760	-0.030354	-0.693881	-0.267612	0.033736
6	-0.193857	0.226798	0.661771	-0.162148	0.128028	-0.522687	-0.396429
7	-0.496004	0.253373	0.023199	0.692493	-0.452730	-0.043105	0.053611

Πίνακας 7.6β. Οι κύριες συνιστώσες που προκύπτουν από την ανάλυση με τον πίνακα συσχετίσεων μετά το μετασχηματισμό των δεδομένων

Συνήθως το να πάρει κανείς απλά τους αριθμούς που αφορούν τους συντελεστές δεν είναι ο κύριος ή τουλάχιστον όχι ο μοναδικός σκοπός. Πολλές φορές είναι πιο χρήσιμο να απεικονίσει κανείς αυτούς τους συντελεστές και πολύ περισσότερο τους συντελεστές των πρώτων συνιστωσών που έχουν και μεγαλύτερη ερμηνευτικότητα. Στο γράφημα 7.2 βλέπουμε αυτό ακριβώς. Τέτοια γραφήματα μας βοηθούν να δούμε πως ομαδοποιούνται οι μεταβλητές. Για παράδειγμα σε σχέση με τις δύο πρώτες κύριες συνιστώσες βλέπουμε πως οι δρόμοι (100μ, 200μ, 800μ) είναι πολύ κοντά μεταξύ τους, μαζί τους είναι και το άλμα εις μήκος, άρα αυτό

αποτελεί μια ένδειξη πως τα αθλήματα αυτά έχουν κάποια κοινά στοιχεία. Κάτι τέτοιο θα ήταν πολύ δύσκολο να προκύψει από τον πίνακα των συντελεστών και θα έπρεπε να έχει κανείς αρκετή εμπειρία για να το δει. Παρατηρείστε ότι αυτές οι μεταβλητές έχουν αρνητικά πρόσημα στη 2η κύρια συνιστώσα, για την οποία μπορούμε να δώσουμε μια ερμηνεία ως ένα κοντράστ των δρόμων με τις ρίψεις.



Γράφημα 7.2. Απεικόνιση των συντελεστών των μεταβλητών στις δύο πρώτες κύριες συνιστώσες.

Όπως πολλές φορές έχουμε πει σκοπός της ανάλυσης σε κύριες συνιστώσες είναι η δημιουργία νέων μεταβλητών από τις ήδη υπάρχουσες, οι νέες μεταβλητές ονομάζονται κύριες συνιστώσες. Πως όμως θα κατασκευάσουμε τις συνιστώσες αυτές; Τα πράγματα είναι απλά. Είδαμε πριν πως η πρώτη κύρια συνιστώσα είναι γραμμικός συνδυασμός των αρχικών μεταβλητών, αρκεί λοιπόν για κάθε παρατήρηση (κάθε αθλήτρια στο παράδειγμα μας) να χρησιμοποιήσουμε τις τιμές που έχουμε για κάθε μεταβλητή και να αντικαταστήσουμε στο τύπο της κύριας συνιστώσας. Για παράδειγμα έχουμε βρει (πίνακας 7.6β) πως η πρώτη κύρια συνιστώσα είναι η

$\Upsilon_1 =$ -0.429 (100μ) -0.392 (ύψος) -0.401 (σφαιροβολία) - 0.407 (200 μ.) -
-0.408 (μήκος) - 0.240 (ακοντισμός) - 0.332 (800μ)

και επίσης ξέρουμε πως για την πρώτη αθλήτρια οι επιδόσεις της (μετά το μετασχηματισμό που είπαμε) και οι τυποποιημένες τους τιμές είναι

	100 μ με εμπόδια (sec)	**Άλμα εις ύψος (m)**	**Σφαιρο-βολία (m)**	**200 μέτρα (sec)**	**Άλμα εις μήκος (m)**	**Ακοντι-σμός (m)**	**800 μέτρα (sec)**
Τιμή	-13.87	1.70	13.11	-25.44	6.23	45.42	-134.31
Τυποποιη-μένη τιμή	-0.074	-0.710	-0.441	-6.102	0.304	0.247	0.252

Επομένως η τιμή της αθλήτριας στην πρώτη κύρια συνιστώσα είναι

$$\Upsilon_1 = -0.429\ (-0.074)\ -0.392\ (-0.710)\ -0.401\ (-0.441) - 0.407\ (-1.102)\ -0.408\ (0.304)\ -$$
$$-0.240\ (0.247) -\ 0.332\ (0.252) = 0.66845$$

Χρησιμοποιούμε τις τυποποιημένες τιμές καθώς η ανάλυση στηρίχτηκε στον πίνακα συσχέτισης που στην ουσία είναι ο πίνακας διακύμανσης για τα τυποποιημένα δεδομένα.

Με όμοιο τρόπο μπορούμε να βρούμε τις τιμές για όλες τις συνιστώσες για όλες τις αθλήτριες. Ο πίνακας 7.7 που ακολουθεί έχει τις τιμές για όλες τις αθλήτριες για τις δύο πρώτες κύριες συνιστώσες. Επίσης στον πίνακα για λόγους σύγκρισης υπάρχουν η πραγματική βαθμολογία από τους αγώνες (στο έπταθλο η επίδοση σε κάθε αγώνισμα παίρνει συγκεκριμένο αριθμό βαθμών, η αθλήτρια με το μεγαλύτερο σύνολο ανακηρύσσεται νικήτρια), καθώς επίσης και η θέση που πήρε κάθε αθλήτρια μαζί με τη θέση που θα έπαιρνε αν για την αξιολόγηση είχαμε χρησιμοποιήσει την πρώτη κύρια συνιστώσα. Επειδή έχουμε χρησιμοποιήσει τον πίνακα συσχέτισης, η μέση τιμή της κάθε συνιστώσας είναι 0 και μπορεί εύκολα να επαληθευτεί από τα δεδομένα πως και η διακύμανση της είναι ίση με την ιδιοτιμή. Επίσης οι συνιστώσες είναι ασυσχέτιστες, αν δηλαδή κανείς υπολογίσει το συντελεστή συσχέτισης ανάμεσα στις καινούριες μεταβλητές που δημιούργησε θα τον βρει ίσο με 0.

Αθλήτρια	**Τελική επίδοση στο αγώνισμα (βαθμοί)**	**Κατάταξη με βάση την επίδοση**	**Σκορ της αθλήτριας στην 1η συνιστώσα**	**Σκορ της αθλήτριας στην 2η συνιστώσα**	**Κατάταξη αθλήτριας με την πρώτη κύρια συνιστώσα**
1	6030	18	0.66845	0.06047	18
2	6251	12	-0.27815	-0.35951	12
3	4530	26	4.97808	-3.5323	26
4	6434	6	-0.90751	-0.0929	7
5	6845	2	-2.93584	-1.58375	2
6	5897	21	1.0869	0.24001	22
7	6649	3	-1.59017	0.88324	4
8	6256	11	-0.74466	-0.40209	10
9	5278	25	3.88302	0.76698	25
10	6388	7	-0.96898	0.62067	6
11	5994	19	0.7007	1.2179	19
12	6464	5	-1.56705	-0.3311	5
13	6300	9	-0.87144	-0.24021	8
14	7044	1	-3.668	-1.27114	1
15	6095	17	0.49016	0.1956	17
16	6263	10	-0.62701	0.05633	11
17	6141	15	0.09514	0.22018	15
18	5749	24	1.48899	-0.20147	23
19	6619	4	-2.20334	-0.78184	3
20	6142	14	0.01656	0.66866	14
21	5993	20	0.98095	0.58074	21
22	5869	22	0.82542	-0.80352	20
23	5847	23	1.92205	1.73972	24
24	6152	13	0.12688	1.31927	16
25	6333	8	-0.84747	-0.06985	9
26	6123	16	-0.05369	1.0999	13

Πίνακας 7.7. Αποτελέσματα επτάθλου πραγματικά και με τη χρήση κυρίων συνιστωσών.

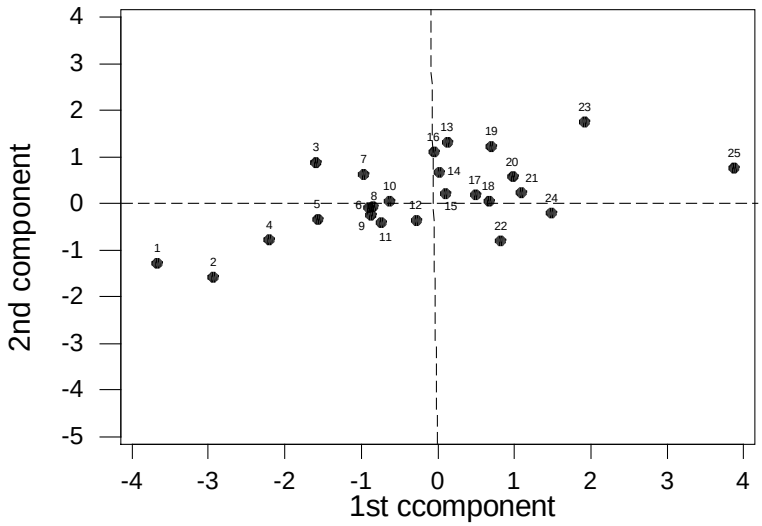
Από τον πίνακα 7.7 μπορεί κάποιος να παρατηρήσει πως υπάρχει μεγάλη συμφωνία στην τελική κατάταξη με τη χρήση κυρίων συνιστωσών. Θα πρέπει να σημειώσουμε πως εξαιτίας των αρνητικών πρόσημων των συντελεστών της πρώτης κύριας συνιστώσας, καλά σκορ είναι τα μικρά σκορ. Δηλαδή το καλύτερο σκορ το έχει η αθλήτρια 14, η οποία και στην πραγματικότητα κέρδισε (Ίζάκι Ίζοινερ-Κερσυ από τις ΗΠΑ). Στο γράφημα 7.3 που ακολουθεί έχουμε απεικονίσει τις αθλήτριες και τις τιμές τους στις δύο πρώτες κύριες συνιστώσες. Δηλαδή η συντεταγμένη κάθε σημείου είναι οι τιμές στις δύο συνιστώσες. Οι ετικέτες που έχουμε χρησιμοποιήσει είναι η τελική τους κατάταξη με βάση τα αποτελέσματα της Ολυμπιάδας. Αυτό που μπορεί κάποιος να δει από το γράφημα 7.3 είναι

- Αν και πως ομαδοποιούνται οι αθλήτριες (πιθανότατα θα μπορούσε κάποιος να χρησιμοποιήσει κάποια άλλα χαρακτηριστικά, όχι απαραίτητα μεταβλητές που χρησιμοποιήσαμε για τη μέθοδο της ανάλυσης σε κύριες συνιστώσες, όπως για παράδειγμα ηλικία, φυλή κλπ) ώστε να δει αν οι συνιστώσες έχουν διακριτική ικανότητα ανάμεσα σε υποπληθυσμούς. Για παράδειγμα, γενικά οι αθλητές της μαύρης φυλής είναι καλύτεροι σους δρόμους και μια ερμηνεία για τη δεύτερη συνιστώσα ήταν ένα κοντράστ δρόμων και ρίψεων. Ομαδοποιούνται οι αθλήτριες με βάση τέτοια κριτήρια;

- Παρατηρείστε πόσο μικρότερη είναι η μεταβλητότητα στη 2η συνιστώσα, πόσο πιο μαζεμένες είναι οι τιμές δηλαδή. Αυτό είναι αναμενόμενο, αφού η διακύμανση της 1ης συνιστώσας είναι 3.43 έναντι μόλις 1.14 της δεύτερης.

- Δείτε επίσης πως είναι σχεδόν σε σειρά από δεξιά προς τα αριστερά οι αθλήτριες με την τελική τους κατάταξη

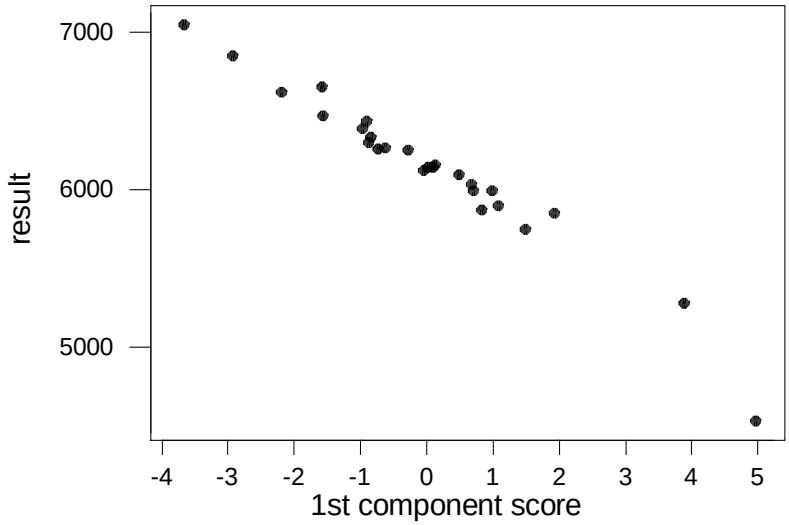
- Συνήθως το ενδιαφέρον περιορίζεται στο να απεικονίζουμε συνιστώσες με μεγάλη διακύμανση (δηλαδή τις πρώτες) καθώς μετά η πληροφορία που παίρνουμε είναι μικρή



Γράφημα 7.3. Τα σκορ στις δύο πρώτες κύριες συνιστώσες για όλες τις αθλήτριες

Στο γράφημα 7.4 μπορεί κανείς να δει τη σχέση της επίδοσης της αθλήτριας και του σκορ

στην πρώτη συνιστώσα. Φαίνεται καθαρά η ύπαρξη μιας έντονης γραμμικής σχέσης (η αρνητική κλίση οφείλεται στο γεγονός πως τα *skor* της πρώτης συνιστώσας είναι καλύτερα όταν είναι μικρά).



Γράφημα 7.4. Οι πραγματικές επιδόσεις και οι τιμές της πρώτης συνιστώσας.

Αυτό δηλαδή που προέκυψε από την ανάλυση σε κύριες συνιστώσες είναι πως κατασκευάσαμε μια συνιστώσα (την πρώτη) η οποία μπορεί να περιγράψει αρκετά καλά την πραγματική κατάσταση να βρει δηλαδή ποια αθλήτρια μοιάζει να είναι καλύτερη. Το πλεονέκτημα είναι πως αυτό έγινε με έναν επιστημονικό τρόπο, δηλαδή προέκυψαν σταθμίσεις για κάθε άθλημα. Στο έπταθλο η βαθμολογία είναι προκαθορισμένη και συνήθως αλλάζει κάθε 10 περίπου χρόνια, ενσωματώνοντας στον τρόπο βαθμολόγησης την πρόοδο σε συγκεκριμένα αθλήματα. Η ανάλυση σε κύριες συνιστώσες μπορεί να αποτελέσει εργαλείο για τη δημιουργία τέτοιων δεικτών τόσο σε αθλητικά όρια (π.χ. συνεισφορά παίκτη στην καλαθοσφαίριση) ή σε άλλες εκφάνσεις της ζωής όπως για παράδειγμα δείκτες αξιολόγησης εταιρειών ή πανεπιστημίων.

Φυσικά κανείς θα πρέπει να έχει σοβαρά υπόψη του πως τα όποια συμπεράσματα μας βασίστηκαν σε 26 παρατηρήσεις και επομένως τα αποτελέσματα πρέπει να τα δεχτούμε με προσοχή. Ίσως κάποιοι το παρατήρησαν αλλά αν όχι δείτε τα δεδομένα του πίνακα 7.1. Η 3η αθλήτρια έχει ακυρωθεί στον ακοντισμό. Αυτό σημαίνει πως έχουμε κάποια missing δεδομένα, καθώς υπό φυσιολογικές συνθήκες η αθλήτρια θα είχε κάποια επίδοση στον ακοντισμό. Αν λοιπόν δεν χρησιμοποιήσουμε αυτή τη μηδενική τιμή και είτε αφαιρέσουμε όλη την παρατήρηση, είτε απλά θεωρήσουμε αυτή την τιμή ως missing, τα αποτελέσματα θα είναι διαφορετικά.

Επίσης θυμηθείτε πως επειδή έχουμε απλά ένα μικρό δείγμα, η ερμηνεία των συντελεστών της πρώτης συνιστώσας εμπεριέχει μέσα της στατιστικό λάθος, δηλαδή το γεγονός ότι οι συντελεστές είναι διαφορετικοί μπορεί απλά να οφείλεται στις κυμάνσεις της τυχαίας δειγματοληψίας και στην πραγματικότητα να είναι όλοι οι συντελεστές ίσοι.

Εναλλακτικά θα μπορούσε κάποιος να αναλύσει τα δεδομένα αυτά όχι χρησιμοποιώντας τις επιδόσεις σε κάθε άθλημα αλλά τη βαθμολογία σε κάθε άθλημα. Με αυτόν τον τρόπο όλες οι μεταβλητές θα είχαν τις ίδιες μονάδες μέτρησης και οι διαφορετικές διακυμάνσεις θα αντιπροσώπευαν το διαφορετικό βάρος κάθε μεταβλητής, σε αυτή την περίπτωση η ανάλυση θα μπορούσε να στηριχτεί στον πίνακα διακύμανσης.

Επιλογή αριθμού συνιστωσών

Μέχρι τώρα δεν αντιμετωπίσαμε καθόλου το πρόβλημα επιλογής του αριθμού των συνιστωσών. Μερικές φορές η επιλογή του αριθμού είναι πολύ χρήσιμη και δυστυχώς δεν υπάρχει κάποιο εργαλείο κοινώς αποδεκτό για να γίνει αυτό. Για τα δεδομένα του επτάθλου εφαρμόσαμε διάφορα κριτήρια που περιγράψαμε προηγουμένως (όχι όλα όμως). Τα στατιστικά πακέτα δυστυχώς προσφέρουν πολύ λίγα από αυτά και μάλλον όχι τα καλύτερα αλλά τα πιο εύκολα.

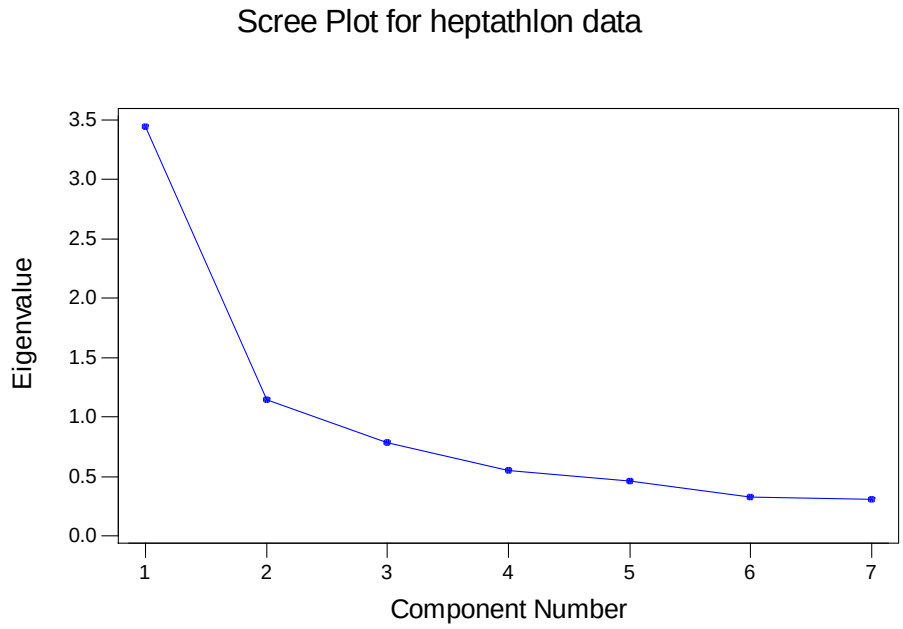
Στον πίνακα 7.8 βλέπουμε τις ιδιοτιμές και τις αναμενόμενες τιμές με τη μέθοδο του σπασμένου ραβδιού. Σύμφωνα με το κριτήριο του Kaiser έχουμε 2 ιδιοτιμές πάνω από 1 επομένως διαλέγουμε 2 συνιστώσες. Βέβαια το γεγονός ότι η τιμή 1.14 μπορεί να οφείλεται σε τυχαία σφάλματα δεν το εξετάζει η μέθοδος.

Αν θέλουμε να σταματήσουμε σε ποσοστό συνολικής διακύμανσης που εξηγείται πάνω από 80% πρέπει να διαλέξουμε 4 συνιστώσες ενώ αν ρίξουμε το ποσοστό αυτό στο 70% χρειαζόμαστε 3 συνιστώσες. Με τη μέθοδο του σπασμένου ραβδιού διαλέγουμε μια μόνο συνιστώσα αφού $0.49 > 0.37$ αλλά για τη δεύτερη συνιστώσα το ποσοστό που εξηγεί είναι 0.16 μικρότερο από το 0.22 που η μέθοδος υποθέτει. Χρησιμοποιώντας δηλαδή 4 κριτήρια έχουμε ήδη καταλήξει σε διαφορετικό αριθμό συνιστωσών! Ας προχωρήσουμε όμως και στα άλλα κριτήρια.

Eigenvalues	Proportion	Cumulative Proportion	Broken Stick
3.4392	0.491	0.491	0.370408
1.1408	0.163	0.654	0.227551
0.7811	0.112	0.766	0.156122
0.5461	0.078	0.844	0.108503
0.458	0.065	0.909	0.072789
0.3248	0.046	0.956	0.044218
0.3100	0.044	1	0.020408

Πίνακας 7.8. Οι ιδιοτιμές του πίνακα συσχετίσεων μαζί με τα ποσοστά της διακύμανσης που ερμηνεύουν και τις αναμενόμενες τιμές για τη μέθοδο του σπασμένου ραβδιού

Στο γράφημα 7.5 μπορούμε να δούμε το scree plot για τα δεδομένα του επτάθλου. Από τη δεύτερη ιδιοτιμή και μετά παρατηρούμε πως οι ιδιοτιμές είναι σε μια περίπου ευθεία γραμμή και επομένως πρέπει να κρατήσουμε μια μόνο συνιστώσα. Φυσικά αυτό είναι μάλλον μια υποκειμενική εξήγηση καθώς θα μπορούσε να υποστηρίξει κάποιος πως η αλλαγή κλίσης ξεκινά μετά τη δεύτερη ιδιοτιμή και άρα πρέπει να κρατήσουμε δύο συνιστώσες. Αυτό αποτελεί και το μειονέκτημα της μεθόδου, ότι δηλαδή δεν είναι πάντα ξεκάθαρος ο αριθμός των συνιστωσών που θα πρέπει να επιλέξουμε.



Γράφημα 7.5. Το scree plot για τα δεδομένα του επτάθλου

Στον πίνακα 7.9 έχουμε κατασκευάσει 95% διαστήματα εμπιστοσύνης για τις ιδιοτιμές με τη χρήση τόσο της κανονικής προσέγγισης όσο και με τη μέθοδο bootstrap. Ως συνήθως είναι πολύ δύσκολο να ελέγξουμε αν ο πληθυσμός από όπου προήλθαν τα δεδομένα μας είναι κανονικός. Από τον πίνακα παρατηρείστε πως τα διαστήματα εμπιστοσύνης βασισμένα στην κανονική προσέγγιση είναι πολύ μεγαλύτερα από αυτά της μεθόδου bootstrap. Αυτό είναι γνωστό και στη βιβλιογραφία, ότι δηλαδή αυτή η προσέγγιση δίνει διαστήματα εμπιστοσύνης κάπως μεγαλύτερα από τα πραγματικά. Με βάση λοιπόν αυτά τα κριτήρια διαλέγουμε τόσες συνιστώσες όσες και οι ιδιοτιμές για τις οποίες με πιθανότητα 95% η τιμή στον πληθυσμό είναι μεγαλύτερη της μονάδας, δηλαδή ιδιοτιμές για τις οποίες το 95% διάστημα εμπιστοσύνης είναι μεγαλύτερο του 1 και δεν το περιέχει. Αυτό συμβαίνει και για τις δύο μεθόδους μόνο για μια ιδιοτιμή και επομένως διαλέγουμε να διατηρήσουμε μόνο μια συνιστώσα. Η τελευταία στήλη του πίνακα μας δείχνει την τυπική απόκλιση κάθε δειγματικής ιδιοτιμής όπως αυτή εκτιμήθηκε με τη μέθοδο bootstrap.

Eigenvalue	95% CI based on normal approximation		Bootstrap 95% CI		Τυπική απόκλιση
3.439	2.228	7.536	2.269	4.610	0.597
1.141	0.739	2.500	0.699	1.582	0.225
0.781	0.506	1.712	0.401	1.161	0.194
0.546	0.354	1.196	0.264	0.828	0.144
0.458	0.297	1.004	0.174	0.742	0.145
0.325	0.210	0.712	0.183	0.466	0.072
0.310	0.201	0.679	0.131	0.489	0.091

Πίνακας 7.9. Διαστήματα εμπιστοσύνης για τις ιδιοτιμές βασισμένα σε προσέγγιση με την κανονική κατανομή και στη μέθοδο bootstrap

Στον πίνακα 7.10 μπορούμε να δούμε τα αποτελέσματα της μεθόδου που βασίζεται στο ποσοστό της διακύμανσης κάθε μεταβλητής που ερμηνεύεται αν κρατήσουμε συγκεκριμένο αριθμό συνιστωσών. Για παράδειγμα, αν κρατήσουμε μόνο μια συνιστώσα ερμηνεύουμε το 63% της διακύμανσης των 100 μέτρων με εμπόδια ενώ μόλις το 20% του ακοντισμού. Όπως είπαμε το κριτήριο συνήθως βασίζεται στο γεγονός πως για όλες τις μεταβλητές ερμηνεύουμε κάποιο ποσοστό μεγαλύτερο από το 70% ή 80%. Και επομένως διαλέγουμε, με βάση αυτό το κριτήριο, 4 και 5 συνιστώσες αντίστοιχα. Το κριτήριο αυτό συνήθως οδηγεί σε μεγάλο αριθμό επιλεγέντων συνιστωσών και αυτό οφείλεται στο γεγονός πως αν υπάρχει κάποια μεταβλητή για την οποία αποτυγχάνουμε να ερμηνεύσουμε τη διακύμανση αναγκαζόμαστε να προσθέτουμε συνιστώσες. Αρκεί δηλαδή μια μεταβλητή με χαλαρή συσχέτιση με τις υπόλοιπες για να οδηγήσει σε μεγάλο αριθμό συνιστωσών.

	Αριθμός συνιστωσών που διατηρούμε						
	1	2	3	4	5	6	7
100 μ με εμπόδια (sec)	0.63	0.64	0.71	0.71	0.91	0.92	1.00
Άλμα εις ύψος (m)	0.53	0.53	0.65	0.96	0.96	0.98	1.00
Σφαιροβολία (m)	0.55	0.72	0.72	0.85	0.86	1.00	1.00
200 μέτρα (sec)	0.57	0.63	0.84	0.84	0.84	0.85	1.00
Άλμα εις μήκος (m)	0.57	0.69	0.71	0.71	0.93	0.94	1.00
Ακοντισμός (m)	0.20	0.88	0.88	0.88	0.91	1.00	1.00
800 μέτρα (sec)	0.38	0.49	0.86	0.95	0.95	1.00	1.00

Πίνακας 7.10. Το ποσοστό της διακύμανσης κάθε μεταβλητής που ερμηνεύεται αν κρατήσουμε k συνιστώσες.

Στον πίνακα 7.11 υπάρχουν τα αποτελέσματα του ελέγχου του Bartlett για την επιλογή συνιστωσών. Ο έλεγχος δουλεύει ως εξής. Αν κάποιες συνιστώσες δεν έχουν στατιστική σημαντικότητα (δηλαδή δεν εξηγούν σημαντικό κομμάτι ή αλλιώς δεν οφείλονται σε κάποια υπάρχουσα δομή των δεδομένων αλλά απλά οφείλονται στην τυχαία δειγματοληψία), τότε οι ιδιοτιμές τους θα είναι θεωρητικά ίσες στον πληθυσμό. Επομένως ο έλεγχος ελέγχει τη μηδενική υπόθεση ότι οι τελευταίες ιδιοτιμές είναι ίσες έναντι της εναλλακτικής ότι δεν είναι. Ξεκινάμε λοιπόν από τον έλεγχο ότι οι 2 τελευταίες ιδιοτιμές είναι ίσες και αν δεν απορρίψουμε συνεχίζουμε ελέγχοντας για τις 3 τελευταίες και ούτω καθεξής μέχρι να απορρίψουμε για πρώτη φορά τη μηδενική υπόθεση. Για το παράδειγμα μας, δεν μπορούμε να

απορρίψουμε την υπόθεση πως οι 6 τελευταίες συνιστώσες είναι ίσες, αλλά μόνο την υπόθεση πως όλες οι συνιστώσες είναι ίσες (σε επίπεδο στατιστικής σημαντικότητας 5%). Επομένως οι 6 τελευταίες συνιστώσες δεν έχουν ενδιαφέρον και κρατάμε μόνο την πρώτη. Παρατηρείστε πως επειδή ο έλεγχος γίνεται ακολουθιακά, υπάρχει πρόβλημα με το πραγματικό επίπεδο στατιστικής σημαντικότητας όμοιο με αυτό που εμφανίζεται στην ανάλυση διακύμανσης με τις πολλαπλές συγκρίσεις. Επίσης ο έλεγχος ισχύει για ιδιοτιμές από πίνακα συσχέτισης.

k	Test statistic	df	p-value
7	64.04026	27	0.01
6	16.65579	20	0.68
5	7.506863	14	0.92
4	2.81215	9	0.95
3	1.176038	5	0.95
2	0.01374	2	0.99

Πίνακας 7.11. Ο έλεγχος του Bartlett για την επιλογή αριθμού συνιστωσών

Είδαμε μέχρι τώρα αρκετά κριτήρια σχετικά με την επιλογή του αριθμού των συνιστωσών. Η ερώτηση που εύλογα προκύπτει είναι ποια από τις μεθόδους προτιμάμε; Η απάντηση στο ερώτημα δεν υπάρχει με σιγουριά. Όλα τα κριτήρια που έχουν προταθεί (καθώς και κάποια άλλα που δεν αναφέραμε ή κάποια που δεν εφαρμόσαμε στο παράδειγμα) έχουν σχεδιαστεί να δουλεύουν όταν υπάρχει πραγματικά κάτι στα δεδομένα, δηλαδή όταν υπάρχει μια πραγματική δομή στον πληθυσμό. Σε αυτή την περίπτωση όλες οι μέθοδοι δίνουν σχετικά όμοια αποτελέσματα. Αυτό δυστυχώς συνήθως δεν συμβαίνει με πραγματικά δεδομένα όπως αυτά που συζητάμε. Αν επομένως πρέπει να προτείνουμε κάποιο κριτήριο, αυτό θα ήταν με τη χρήση της μεθόδου boostrap. Όμως αυτή η μέθοδος έχει ένα μεγάλο μειονέκτημα, απαιτεί αρκετό χρόνο και πολύ υπολογιστικό κόπο για να εφαρμοστεί, δεν προσφέρεται από τα υπάρχοντα στατιστικά πακέτα και επομένως στην πράξη δεν μπορεί κάποιος να την εφαρμόσει εύκολα. Όπως είπαμε και προηγουμένως τα στατιστικά πακέτα προσφέρουν πολύ λίγα κριτήρια. Το scree plot και το κριτήριο του Kaiser τα οποία είναι διαθέσιμα σε όλα τα πακέτα συνήθως υπερεκτιμούν τον αριθμό των συνιστωσών, το ίδιο συμβαίνει και με τα κριτήρια που στηρίζονται στο ποσοστό της διακύμανσης που εξηγείται. Ο έλεγχος του Bartlett έχει ενδιαφέρον αν και μόνο αν μπορούμε να εξασφαλίσουμε την κανονικότητα του πληθυσμού, κάτι που πολλές φορές δεν είναι λογικό αλλά και δεν μπορεί να ελεγχθεί.

Μέθοδος	**Αριθμός Συνιστωσών**
Ιδιοτιμές >1	2
Ποσοστό συνολικής διακύμανσης > 90%	5
Ποσοστό συνολικής διακύμανσης >80%	4
Μέθοδος σπασμένου ραβδιού	1
Bootstrap διαστήματα εμπιστοσύνης	1
Διαστήματα εμπιστοσύνης με κανονική προσέγγιση	1
Έλεγχοι υποθέσεων του Bartlett	1
Scree plot	1
Διακύμανση κάθε μεταβλητής >80%	5
Διακύμανση κάθε μεταβλητής >80%	4

Πίνακας 7.12. Επιλογή αριθμού συνιστωσών με διάφορες μεθόδους

Τελειώνοντας λοιπόν την παρουσίαση της μεθόδου των κυρίων συνιστωσών θα πρέπει να τονίσουμε ότι πρόκειται για έναν χρήσιμο μετασχηματισμό των δεδομένων μας που δεν βασίζεται σε κάποιο στατιστικό μοντέλο αλλά απλά στις ιδιότητες του πίνακα διακύμανσης (και συσχέτισης). Η μέθοδος μπορεί να μας ερμηνεύσει τη διακύμανση των μεταβλητών και πως αυτές σχετίζονται μεταξύ τους. Για πολλούς συγγραφείς η μέθοδος των κυριών συνιστωσών είναι απλά μια ειδική περίπτωση της παραγοντικής ανάλυσης που θα δούμε σε λίγο, σε μερικά μάλιστα στατιστικά πακέτα δεν υπάρχει σαν αυτόνομη επιλογή. Θα πρέπει να τονιστεί πως η ανάλυση σε κύριες συνιστώσες διαφέρει πολύ από την παραγοντική ανάλυση και η τελευταία απλά την χρησιμοποιεί.

7.11 Bootstrap στην ανάλυση κυρίων συνιστωσών

Όπως είπαμε και πριν στατιστική συμπερασματολογία στην ανάλυση σε κύριες συνιστώσες είναι ιδιαίτερα δύσκολη. Η δυσκολία οφείλεται σε δύο γεγονότα: α) χρειάζεται να κάνουμε υποθέσεις για τις οποίες δύσκολα μπορούμε να δούμε αν ισχύουν και β) λόγω της πολυμεταβλητής φύσης των δεδομένων ακόμα και αν υπάρχουν ασυμπτωτικά αποτελέσματα είναι δύσκολο να τα εφαρμόσουμε με επιτυχία στην πράξη. Λύση σε αυτά τα προβλήματα μπορεί να δώσουν οι μέθοδοι Bootstrap.

Η μεθοδολογία bootstrap η οποία είναι πια πολύ διαδεδομένη στη στατιστική βασίζεται στην ιδέα πως ακόμα και αν δεν γνωρίζουμε την κατανομή μιας στατιστικής συνάρτησης, αν είμαστε σε θέση να προσομοιώσουμε από αυτήν μπορούμε να εκτιμήσουμε

την άγνωστη αυτή κατανομή παίρνοντας ένας δείγμα με προσομοίωση από αυτήν. Αυτό το δείγμα επειδή προέρχεται από την κατανομή που θέλουμε μπορεί να μας αποκαλύψει πολλά από τα χαρακτηριστικά της.

Το πρόβλημα όμως είναι από ποια κατανομή θα προσομοιώσουμε. Δηλαδή αυτό που είπαμε στην προηγούμενη παράγραφο στηρίζεται στο γεγονός πως ξέρουμε την κατανομή του πληθυσμού και επομένως επαναλαμβάνουμε το πείραμα πολλές φορές. Στα περισσότερα όμως προβλήματα δεν γνωρίζουμε την κατανομή του πληθυσμού και επομένως χρειαζόμαστε κάποια άλλη τεχνική για να ξεπεράσουμε το πρόβλημα. Αυτή η τεχνική είναι η μέθοδος bootstrap. Αν σκεφτούμε πως η εμπειρική κατανομή είναι μια καλή προσέγγιση της άγνωστης κατανομής του πληθυσμού τότε μπορούμε να χρησιμοποιήσουμε την εμπειρική κατανομή (δηλαδή την κατανομή που δίνει πιθανότητα *1/n* σε κάθε παρατήρηση) για να προσομοιώσουμε το πείραμα / φαινόμενο που εξετάζουμε.

Αυτό ακριβώς θα χρησιμοποιήσουμε και στην ανάλυση σε κύριες συνιστώσες. Για παράδειγμα είπαμε πριν ότι μας ενδιαφέρει να βρούμε τη μεταβλητότητα των ιδιοτιμών ώστε να αποφασίσουμε πόσες συνιστώσες θα κρατήσουμε. Επομένως αυτό μπορούμε να το κάνουμε παίρνοντας Bootstrap δείγματα και κατασκευάζοντας διαστήματα εμπιστοσύνης για κάθε ιδιοτιμή. Θα δούμε αμέσως μια εφαρμογή. Περισσότερες λεπτομέρειες για τη μέθοδο bootstrap ο ενδιαφερόμενος αναγνώστης μπορεί να βρει στο βιβλίο των Efron και Tibshirani (1993).

Θα δούμε λοιπόν τη χρήση της μεθόδου bootstrap σε αυτό το πρόβλημα καθώς και άλλα σχετικά με την ανάλυση σε κύριες συνιστώσες εξετάζοντας ένα σετ από πραγματικά δεδομένα. Τα δεδομένα που θα αναλύσουμε αφορούν το αγώνισμα του επτάθλου στους Ολυμπιακούς αγώνες του Σίδνευ το 2000. Συγκεκριμένα μόνο οι 26 αθλήτριες που βαθμολογήθηκαν και στα 7 αγωνίσματα θα χρησιμοποιηθούν για την ανάλυση. Σκοπός της ανάλυσης είναι να δούμε αν, και κατά πόσο, τα 7 αγωνίσματα μπορούν να αντικατασταθούν από συνιστώσες που μετράνε κάποιες μη παρατηρήσιμες ποσότητες, όπως η δύναμη ή η ταχύτητα. Εμείς θα παρουσιάσουμε το παράδειγμα επικεντρώνοντας περισσότερο στο τι μπορεί να μας προσφέρει η μέθοδος bootstrap. Τα δεδομένα υπάρχουν στον πίνακα 7.13 Για τα αγωνίσματα των δρόμων (100μ με εμπόδια, 200 μέτρα και 800 μέτρα) τα δεδομένα είναι σε δευτερόλεπτα ενώ για τα υπόλοιπα (ρίψεις και άλματα) σε μέτρα.

Κατά αρχάς θα πρέπει να τονίσουμε πως θα δουλέψουμε με τον πίνακα συσχετίσεων, λόγω των διαφορετικών μονάδων μέτρησης σε κάθε μεταβλητή (αγώνισμα). Επίσης επειδή για τους δρόμους καλές επιδόσεις είναι οι μικρές τιμές (μικροί χρόνοι) θα αλλάξουμε τα πρόσημα των παρατηρήσεων ώστε σε όλα τα αγωνίσματα οι καλές επιδόσεις να είναι οι μεγάλες τιμές. Αυτό διευκολύνει πολύ την όποια ερμηνεία των συνιστωσών.

Οι συναρτήσεις για τις οποίες ενδιαφερόμαστε είναι οι εξής:

- Ποια είναι η κατανομή των ιδιοτιμών του πίνακα συσχετίσεων; Πόσες ιδιοτιμές είναι μεγαλύτερες της μονάδας αν κανείς λάβει υπόψη την τυχαιότητα λόγω του δείγματος (οι

επταθλήτριες είναι ένα δείγμα από τον πληθυσμό των επταθλητριών, δεν θα ασχοληθούμε με το θέμα αν και κατά πόσο είναι τυχαίο δείγμα ή άλλα τέτοια προβλήματα).

- Η πρώτη συνιστώσα που προκύπτει ως ένας σταθμικός μέσος των αγωνισμάτων δίνει διαφορετικά βάρη σε κάθε αγώνισμα ή όχι.

Αθλήτρια	Χώρα	100μ εμπόδια	Άλμα εις ύψος	Σφαιροβολία	200 μ.	άλμα εις μήκος	Ακοντισμός	800 μ.
Azzizi Yasmina	ALG	13.64	1.6	14.17	24.59	5.88	46.28	141.82
Bacher Gertrud	ITA	13.82	1.75	12.75	24.96	5.84	41.14	129.08
Biswas Soma	IND	14.11	1.63	11.69	24.73	5.64	39.59	142.17
Braun Sabine	GER	13.49	1.81	14.33	24.74	6.22	48.56	139.14
Ganapathy G. Pramila	IND	14.22	1.69	11.14	24.69	5.96	36.02	140.86
Garcva Magalys	CUB	13.46	1.66	13.29	24.58	5.92	50.31	139.64
Hautala Tiia	FIN	13.62	1.78	13.31	25.00	6.12	45.4	134.9
Jamieson Jane	AUS	14.09	1.81	13.59	25.27	6.09	45.32	136.57
Kabanova Sofiya	UZB	14.89	1.72	11.56	27.27	5.22	36.61	140.11
Kazanina Svetlana	KZK	14.71	1.75	12.97	25.04	5.84	43.53	130.45
Koritskaya Diana	RUS	13.88	1.72	13.53	24.08	5.56	40.67	129.77
Kovalenko- Lyudmila	UKR	15.12	1.72	13.57	26.36	5.57	42.5	133.52
Lewis Denise	GBR	13.23	1.75	15.55	24.34	6.48	50.19	136.83
Mark Marsha	TRI	13.72	1.66	11.44	25.35	5.9	48.99	152.36
Nathan Le Shundra	USA	13.74	1.78	14.22	24.84	6.06	43.48	136.67
Naumenko Irina	KZK	14.26	1.84	11.26	25.19	5.88	32.53	138.49
Prokhorova Yelena	RUS	13.63	1.81	13.21	23.72	6.59	45.05	130.32
Rajamδki Susanna	FIN	13.6	1.66	13.87	24.03	6.36	37	138.47
Roshchupkina Natalya	RUS	13.7	1.84	14.03	23.53	5.47	43.87	132.24
Sazanovich Natalya	BLR	13.45	1.84	14.79	24.12	6.5	43.97	136.41
Skujyte Austra	LIT	14.37	1.78	15.09	25.35	5.97	45.43	140.25
Specht-Ertl Karin	GER	13.43	1.78	13.55	24.64	6.22	42.7	136.25
Teppe Nathalie	FRA	14.02	1.72	13.44	26.39	5.94	46.98	138.56
Teteryuk Larisa	UKR	14.53	1.72	13.56	25.76	5.89	44.57	139.94
Tigau Viorica	ROM	13.39	1.72	11.53	24.8	6.01	43.38	139.65
Wlodarczyk Urszula	POL	13.33	1.78	14.45	24.29	6.31	46.16	132.15

Πίνακας 7.13 Οι επιδόσεις των 26 επταθλητριών στην ολυμπιάδα του Σίδνευ, 2000

Σκοπός της εφαρμογής είναι να αναδείξει τη χρησιμότητα της μεθόδου bootstrap και όχι τα υπέρ και τα κατά της ανάλυσης σε κύριες συνιστώσες. Σε κάθε περίπτωση όμως θα δούμε πως λειτουργεί η μέθοδος bootstrap σε ένα πολυμεταβλητό παράδειγμα όπου κλασικές μέθοδοι βασισμένες σε υποθέσεις και ασυμπτωτικά αποτελέσματα είναι μάλλον δύσκολο να χρησιμοποιηθούν.

Για να πάρουμε ένα bootstrap δείγμα από τα δεδομένα μας αρκεί να πάρουμε με επανάθεση παρατηρήσεις, δηλαδή αθλήτριες. Ως παρατήρηση εννοούμε όλο το διάνυσμα με τα 7 αγωνίσματα. Συνεπώς για το παράδειγμα μας το bootstrap δείγμα μπορεί να περιέχει

περισσότερες από μια φορές κάποια αθλήτρια. Στη συνέχεια για κάθε bootstrap δείγμα που δημιουργήσαμε θα προχωρήσουμε σε ανάλυση σε κύριες συνιστώσες με τη χρήση του πίνακα συσχετίσεων τις ποσότητες που μας ενδιαφέρουν: τις ιδιοτιμές, τα ιδιοδιανύσματα και την ορίζουσα του πίνακα συσχετίσεων.

7.11.1 Η κατανομή των ιδιοτιμών.

Ας ξεκινήσουμε από το πρόβλημα που σχετίζεται με τον αριθμό των κυρίων συνιστωσών που πρέπει να χρησιμοποιήσουμε. Όπως είπαμε ένα από τα κριτήρια που συνήθως χρησιμοποιούμε, αυτό του Kaiser, προτρέπει να κρατήσουμε τόσες συνιστώσες όσες και ιδιοτιμές μεγαλύτερες από το 1 έχουμε. Στην περίπτωση μας αυτές είναι 3, όπως μπορούμε να δούμε στον πίνακα 7.14. Η τρίτη ιδιοτιμή είναι 1.05 και είναι αρκετά λογικό να έχει αυτή την τιμή απλά για λόγους τυχαιότητας και όχι γιατί απαραίτητα αντιστοιχεί σε κάποια συνιστώσα με σημαντική συνεισφορά.

Α/Α	Ιδιοτιμή	Ποσοστό της συνολικής διακύμανσης που εξηγεί η συνιστώσα	Αθροιστικό Ποσοστό της συνολικής διακύμανσης που εξηγούν οι συνιστώσες
1	2.96009	42.29%	42.287%
2	1.51991	21.71%	64.000%
3	1.05053	15.01%	79.008%
4	0.64641	9.23%	88.242%
5	0.38602	5.51%	93.757%
6	0.27779	3.97%	97.725%
7	0.15924	2.27%	100%

Πίνακας 7.14. Ιδιοτιμές του πίνακα συσχετίσεων

Θέλουμε επομένως να εκτιμήσουμε τη διακύμανση κάθε ιδιοτιμής με σκοπό να δούμε αν αυτή είναι πράγματι μεγαλύτερη από τη μονάδα ή αν απλά έτυχε για λόγους τυχαίας δειγματοληψίας. Η μέθοδος bootstrap μπορεί να βοηθήσει σε αυτή την κατεύθυνση. Παρατηρείστε πως η κατανομή μιας ιδιοτιμής είναι ιδιαίτερα δύσκολο να μελετηθεί θεωρητικά αν και υπάρχουν θεωρητικά αποτελέσματα τα οποία στηρίζονται είτε σε κανονικότητα του πληθυσμού είτε σε ασυμπτωτικά θεωρήματα. Προφανώς η μέθοδος bootstrap δεν χρειάζεται τέτοιες υποθέσεις.

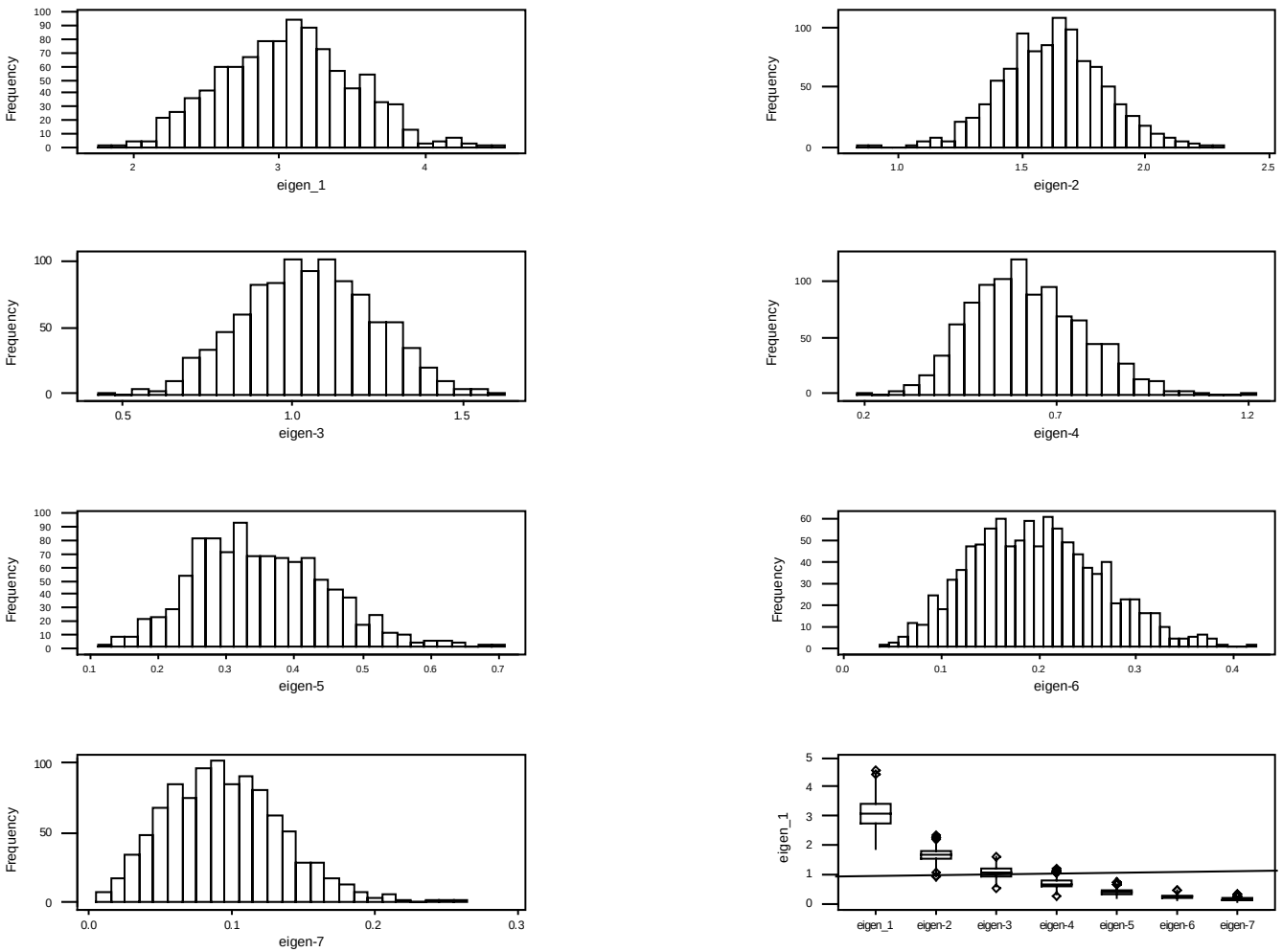
Στον πίνακα 7.15 μπορεί κανείς να δει τη μέση τιμή για κάθε ιδιοτιμή καθώς και την εκτίμηση της τυπικής της απόκλισης βασισμένες σε 1000 bootstrap δείγματα. Επίσης έχουμε κατασκευάσει ένα 95% διάστημα εμπιστοσύνης για κάθε ιδιοτιμή, βασισμένο στη μέθοδο των ποσοστιαίων σημείων (το πώς χρησιμοποιούμε τη μέθοδο bootstrap για να κατασκευάσουμε διαστήματα εμπιστοσύνης δεν θα συζητηθεί εδώ, ο αναγνώστης μπορεί να απευθυνθεί σε

κλασικά βιβλία για τα τη μέθοδο bootstrap). Αν κοιτάξει κανείς στο γράφημα 7.6 μπορεί να δει κατά τα ιστογράμματα των ιδιοτιμών καθώς και ένα διάγραμμα πλαισίου και απολήξεων για όλες τις ιδιοτιμές. Είναι πολύ ενδιαφέρον να παρατηρήσει κανείς πως το διάστημα εμπιστοσύνης για την τρίτη ιδιοτιμή περιέχει ξεκάθαρα την τιμή 1 και συνεπώς η τρίτη ιδιοτιμή δεν μπορεί να ισχυριστεί κανείς πως είναι στατιστικά σημαντικά μεγαλύτερη από το 1, δηλαδή δεν υπάρχει κάποιος λόγος να κρατήσουμε την τρίτη συνιστώσα. Επίσης παρατηρείστε πως η κατανομή των ιδιοτιμών δεν μοιάζει με την κανονική για όλες τις ιδιοτιμές. Συγκεκριμένα οι ουρές της κατανομής είναι συνήθως πιο παχιές από αυτές της κανονικής κατανομής, ιδιαίτερα για τις ιδιοτιμές μεγάλης τάξης. Κάνοντας έλεγχο κανονικότητας Anderson Darling απορρίπτουμε την υπόθεση της κανονικότητας για όλες τις ιδιοτιμές εκτός από τη δεύτερη.

Επομένως η μέθοδος bootstrap μπορεί να χρησιμοποιηθεί για να μελετηθεί η κατανομή των ιδιοτιμών και άρα να επιλεγεί ο αριθμός των συνιστωσών που θα κρατηθούν για περαιτέρω ανάλυση.

Α/Α	Ιδιοτιμή	Μέση τιμή ιδιοτιμής βασισμένη σε 1000 επαναλήψεις	Τυπική απόκλιση βασισμένη σε 1000 επαναλήψεις	95% Διάστημα εμπιστοσύνης (percentile μέθοδος)	
1	2.96009	3.057	0.450	2.229	3.896
2	1.51991	1.627	0.206	1.227	2.043
3	1.05053	1.049	0.191	0.685	1.416
4	0.64641	0.624	0.145	0.375	0.932
5	0.38602	0.350	0.097	0.183	0.552
6	0.27779	0.198	0.067	0.080	0.333
7	0.15924	0.095	0.041	0.025	0.184

Πίνακας 7.15 Εκτιμηθείσες τυπικές αποκλίσεις και 95% διαστήματα εμπιστοσύνης για τις ιδιοτιμές



Γράφημα 7.6 Ιστογράμματα και διάγραμμα Boxplot για τις ιδιοτιμές.

7.11.2 Ερμηνεία των ιδιοδιανυσμάτων – συνιστωσών

Μια άλλη ενδιαφέρουσα χρησιμότητα της μεθόδου bootstrap στην ανάλυση σε κύριες συνιστώσες είναι πως μπορούμε να κάνουμε συμπερασματολογία σχετικά με τις ίδιες τις συνιστώσες και τους συντελεστές τους. Θυμηθείτε πως υπάρχουν ασυμπτωτικά αποτελέσματα τα οποία όμως αφορούν πολυμεταβλητούς κανονικούς πληθυσμούς και μεγάλο μέγεθος δείγματος.

Από τα δεδομένα μας προκύπτει πως η πρώτη κύρια συνιστώσα είναι η

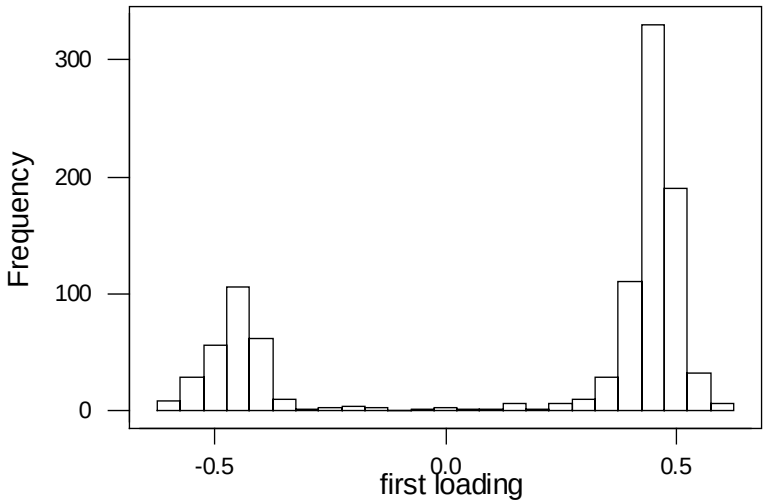
$Y_1 =$ 0.466 x (100 μ) + 0.246 x (ύψος) +0.420 x (σφαιροβολία)
+ 0.433 x (200 μ) +0.463 x (άλμα εις μήκος)
+0.323 x (ακοντισμός) +0.202 x (800μ)

Μια ερμηνεία αυτής της συνιστώσας θα μπορούσε να είναι ότι είναι ένας σταθμικός μέσος των αγωνισμάτων (θυμηθείτε πως έχουμε αλλάξει πρόσημα στους δρόμους ώστε μεγάλες τιμές δείχνουν μεγάλες επιδόσεις). Και το ερώτημα που προκύπτει είναι

- Μπορούμε να κάνουμε συμπερασματολογία για κάθε συντελεστή ξεχωριστά; Ένας μηδενικός συντελεστής σημαίνει πως η μεταβλητή δεν είναι σημαντική για τη συνιστώσα (κατ' αναλογία με το ότι συμβαίνει και στη γραμμική παλινδρόμηση).
- Μπορούμε να πούμε πως όλες οι μεταβλητές έχουν την ίδια στάθμιση και πως απλά έτυχε για λόγους τυχαιότητας να βρούμε διαφορετικούς συντελεστές;

Θα εξετάσουμε και τα δύο ερωτήματα. Κατά αρχάς ας θυμηθούμε πως το άθροισμα τετραγώνων των συντελεστών αθροίζει αναγκαστικά στη μονάδα, επομένως αν κάποιος αλλάξει τα πρόσημα μπορεί να πάρει πάλι μια λύση. Αν δούμε το ιστόγραμμα που ακολουθεί (γράφημα 7.7) και αφορά το συντελεστή των 100 μέτρων με εμπόδια παρατηρείστε πως έχει ξεκάθαρα μια δίκορφη μορφή. Οι δύο κορυφές εξηγούνται λόγω της δυνατότητας να αλλάξει κανείς αυθαίρετα πρόσημο στη συνιστώσα. Παρατηρείστε μια κάποια συμμετρία ως προς το 0. Τα υπάρχοντα στατιστικά πακέτα, αποφασίζουν αυθαίρετα αν θα κρατήσουν τη λύση με το θετικό ή το αρνητικό πρόσημο για τον πρώτο συντελεστή. Κάποιος που θα χρησιμοποιήσει τη μέθοδο bootstrap στην ανάλυση σε κύριες συνιστώσες πρέπει να έχει το νου του σε αυτό το πρόβλημα.

Για να αποφύγουμε αυτό το πρόβλημα αποφασίσαμε πως για κάθε λύση το πρόσημο του πρώτου συντελεστή θα είναι θετικό. Με αυτό τον περιορισμό πήραμε τα ιστογράμματα και τα διαγράμματα Boxplot που βλέπετε στο γράφημα 7.8



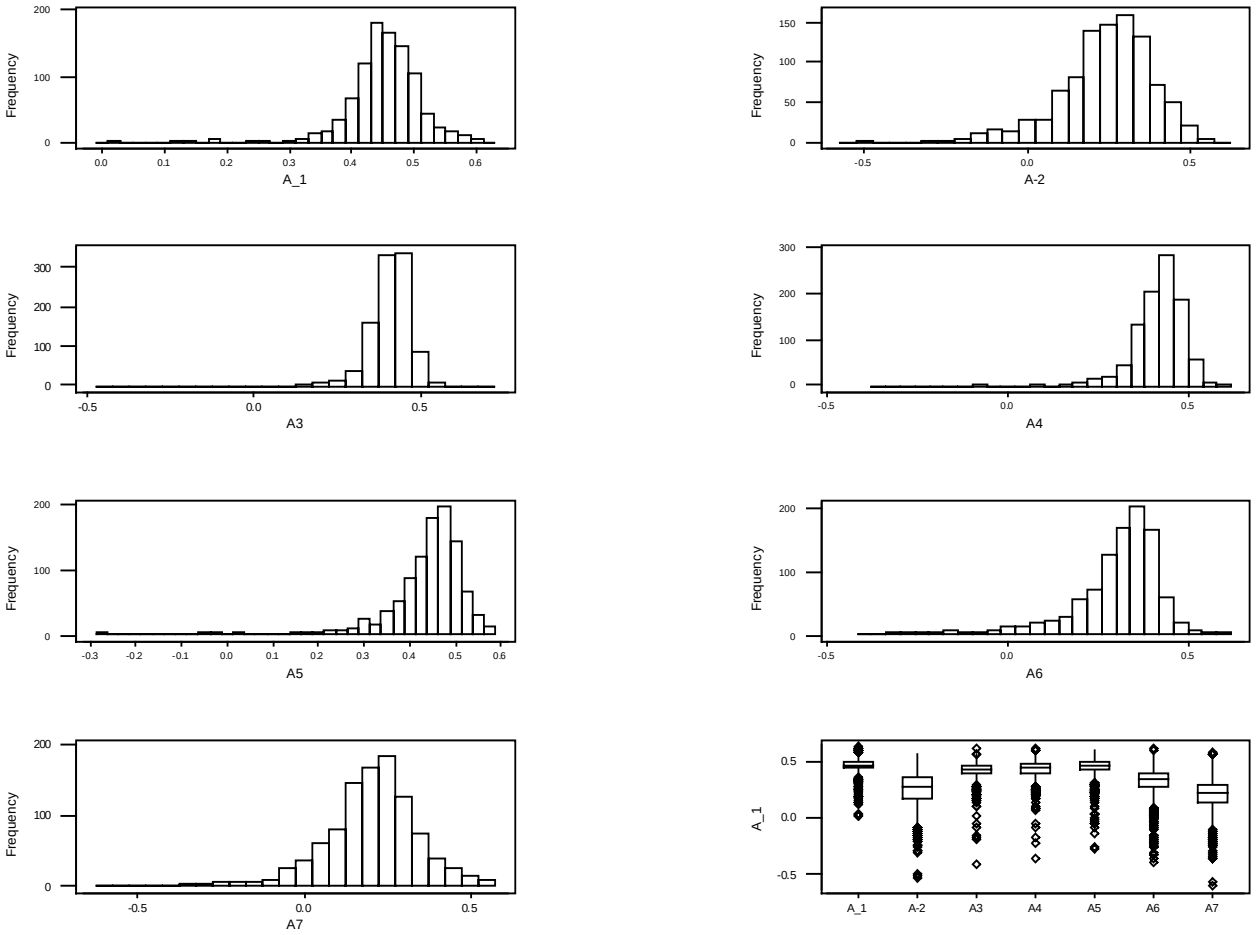
Γράφημα 7.7 Ιστόγραμμα για το συντελεστή των 100 μέτρων στην πρώτη συνιστώσα. Δεν έχει τεθεί περιορισμός σχετικά με το πρόσημο.

Παρατηρείστε τη μεγάλη αριστερή ουρά των περισσοτέρων συντελεστών. Ο συντελεστής του ύψους, του ακοντισμού και των 800 μέτρων είναι συστηματικά μικρότερος από τα άλλα αγωνίσματα κάτι που μάλλον δείχνει πως αυτά τα αγωνίσματα έχουν μικρότερο βάρος.

Ενδιαφέρον παρουσιάζει και ο πίνακας 7.16 με τις μέσες τιμές και τις τυπικές αποκλίσεις των συντελεστών καθώς και 95% διαστήματα εμπιστοσύνης για τον καθένα.

Α/Α	Τιμή συντελεστή	Μέση τιμή συντελεστή βασισμένη σε 1000 επαναλήψεις	Τυπική απόκλιση βασισμένη σε 1000 επαναλήψεις	95% Διάστημα εμπιστοσύνης (percentile μέθοδος)	
1	0.466	0.448	0.067	0.267	0.561
2	0.246	0.238	0.150	-0.137	0.478
3	0.420	0.404	0.079	0.230	0.509
4	0.433	0.415	0.081	0.228	0.527
5	0.463	0.437	0.092	0.206	0.546
6	0.323	0.300	0.131	-0.084	0.471
7	0.202	0.200	0.143	-0.133	0.464

Πίνακας 7.16 Οι τιμές των συντελεστών και εκτιμήσεις των τυπικών τους σφαλμάτων με τη μέθοδο Bootstrap. Μπορείτε επίσης να δείτε και ένα 95% διάστημα εμπιστοσύνης για τους συντελεστές



Γράφημα 7.8. Ιστογράμματα και διάγραμμα Boxplot για τους συντελεστές της πρώτης συνιστώσας

Μπορεί κανείς να παρατηρήσει πως το 95% διάστημα εμπιστοσύνης του συντελεστή για το ύψους, τον ακοντισμό και τα 800 μέτρα περιέχει την τιμή 0 και επομένως αυτοί οι συντελεστές δεν διαφέρουν στατιστικά σημαντική από το 0, δηλαδή οι μεταβλητές δεν είναι συσχετισμένες με τη συνιστώσα.

Ένα άλλο ενδιαφέρον θέμα είναι κατά πόσο οι συντελεστές είναι όλοι ίδιοι. Αν σκεφτούμε πως το άθροισμα τετραγώνων είναι 1 τότε κάθε συντελεστής θα έπρεπε να είναι 0.377. Επομένως θέλω να ελέγξω αν το διάνυσμα των συντελεστών είναι το (0.377, 0.377, 0.377, 0.377, 0.377, 0.377, 0.377).

Με βάση αυτά που είπαμε στο κεφάλαιο 5, έχω 100 παρατηρήσεις (100 διανύσματα) και θέλω να ελέγξω τη μηδενική υπόθεση πως ο μέσος τους είναι το διάνυσμα (0.377, 0.377, 0.377, 0.377, 0.377, 0.377, 0.377). Μπορώ επομένως να χρησιμοποιήσω τον έλεγχο του Hotelling. Προκειμένου όμως να αποφύγω την υπόθεση της πολυμεταβλητής κανονικότητας που απαιτεί ο έλεγχος θα καταφύγω σε ελέγχους bootstrap>

Κάτω από την υπόθεση της πολυμεταβλητής κανονικότητας για το διάνυσμα των συντελεστών θα μπορούσε κανείς να κατασκευάσει ένα από κοινού 95% διάστημα εμπιστοσύνης το οποίο θα περιείχε το 95% των παρατηρήσεων. Συγκεκριμένα αν με $\overline{X}$ συμβολίσουμε το διάνυσμα με τις μέσες τιμές από το δείγμα και με **S** το δειγματικό πίνακα διακύμανσης συνδιακύμανσης, τότε το ελλειψοειδές που περικλείεται μέσα στα σημεία που ικανοποιούν τη σχέση

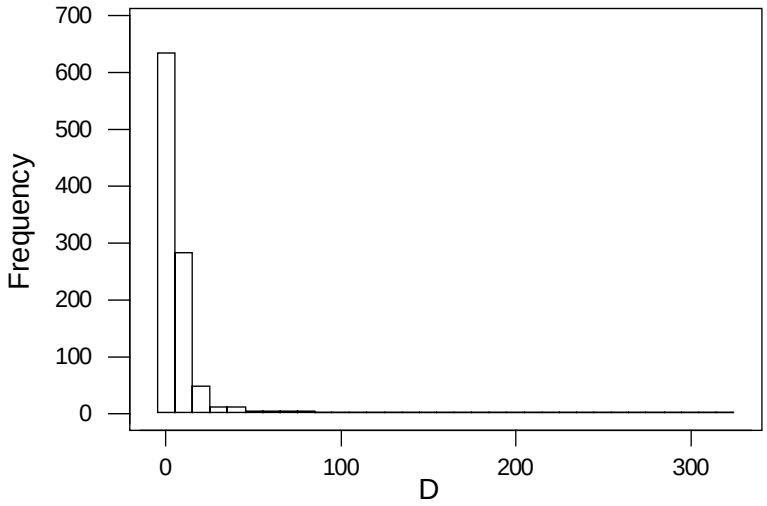
$$(x_i - \overline{x})' S^{-1} (x_i - \overline{x}) \le C$$

είναι ένα διάστημα εμπιστοσύνης. Η τιμή C είναι μια κατάλληλα επιλεγμένη κριτική τιμή από μια F κατανομή (για την ακρίβεια από είναι συνάρτηση μιας F κριτικής τιμής). Στην περίπτωση μας, όπου δεν μπορούμε και δεν θέλουμε να υποθέσουμε κανονικότητα για τα δεδομένα μας, θα πρέπει να κατασκευάσουμε από τις bootstrap τιμές την κατανομή της συνάρτησης $D_i = (a_i - \overline{a})' S^{-1} (a_i - \overline{a})$, όπου a_i είναι το διάνυσμα των συντελεστών από το i bootstrap δείγμα, $\overline{a}$ είναι το διανυσμάτων μέσων από τις 1000 επαναλήψεις, δηλαδή $\overline{a}$ = (0.448, 0.238, 0.404, 0.415, 0.437, 0.300, 0.200) και **S** είναι ο πίνακας διακύμανσης – συνδιακύμανσης των συντελεστών, τον οποίο έχουμε εκτιμήσεις από τις 1000 bootstrap τιμές που πήραμε. Δηλαδή δεν θα χρησιμοποιήσουμε την κριτική τιμή της F κατανομής.

Θα πρέπει να σημειωθεί πως υπάρχουν σημαντικές συσχετίσεις ανάμεσα στους συντελεστές. Ο πίνακας συσχετίσεων , όπως εκτιμήθηκε με τη μέθοδο Bootstrap είναι ο

$$\mathbf{R}=\begin{bmatrix} 1.00000 & & & & & & \\ -0.59102 & 1 & & & & & \\ -0.02199 & 0.20992 & 1 & & & & \\ 0.29812 & -0.08829 & -0.20557 & 1 & & & \\ 0.36047 & -0.17295 & -0.09661 & 0.12633 & 1 & & \\ 0.40529 & -0.35193 & 0.26834 & -0.39413 & 0.15731 & 1 & \\ -0.54235 & 0.64511 & 0.18617 & 0.15738 & -0.34500 & -0.43607 & 1 \end{bmatrix}$$

και μπορεί κανείς να δει καθαρά ότι κάποιες συσχετίσεις είναι αρκετά μεγάλες.



Γράφημα 7.9 Ιστόγραμμα συχνοτήτων για τη συνάρτηση *D*.

Στο γράφημα 7.9 μπορεί κανείς να δει την κατανομή της συνάρτησης Di. Παρατηρείστε πως υπάρχει μια πολύ μεγάλη τιμή κοντά στο 300, δηλαδή για ένα bootstrap δείγμα οι εκτιμηθέντες συντελεστές απείχαν πάρα πολύ από ότι συνέβηκε στα υπόλοιπα bootstrap δείγματα. Η μέση τιμή είναι 6.99 αλλά λόγω της ύπαρξης πολλών ακραίων τιμών δεν είναι αξιόπιστο μέτρο θέσης. Η διάμεσος είναι 3.65 και ο 5% περικομμένος μέσος 4.95. Η τυπική απόκλιση της είναι 15.137. Η κατανομή είναι ιδιαίτερα ασύμμετρη και έχει μεγάλη δεξιά ουρά. Αν κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης με τη μέθοδο των ποσοστιαίων σημείων, αυτό είναι το (0.816, 35.157). Βέβαια θα χρειαζόντουσαν περισσότερες επαναλήψεις για να είναι το διάστημα εμπιστοσύνης πιο αξιόπιστο, θυμηθείτε την ύπαρξη κάποιων πολύ ακραίων τιμών. Παρόλα αυτά το διάνυσμα (0.377, 0.377, 0.377, 0.377, 0.377, 0.377, 0.377) που μας ενδιαφέρει έχει τιμή *D=5.142* η οποία σε καμιά περίπτωση δεν είναι ακραία. Μάλιστα από τις 1000 τιμές που έχουμε η τιμή αυτή είναι μεγαλύτερη μόλις από τις 355 και επομένως δεν μπορούμε να ισχυριστούμε πως το διάνυσμα αυτό έχει πολύ μικρή πιθανότητα να παρατηρηθεί. Δηλαδή η υπόθεση πως η πρώτη συνιστώσα είναι ένας σταθμικός μέσος των αγωνισμάτων δείχνει να ευσταθεί.

8 ΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ

8.1 Εισαγωγή

Η παραγοντική ανάλυση είναι μια στατιστική μέθοδος που έχει σκοπό να βρει την ύπαρξη παραγόντων κοινών ανάμεσα σε μια ομάδα μεταβλητών. Έτσι, εκφράζοντας αυτούς τους παράγοντες (οι οποίοι δεν είναι μια υπαρκτή ποσότητα αλλά την 'κατασκευάζουμε' για τις ανάγκες μας) μπορούμε

- Να μειώσουμε τις διαστάσεις του προβλήματος. Αντί να δουλεύουμε με τις αρχικές μεταβλητές να δουλέψουμε με λιγότερες αφού οι παράγοντες είναι έτσι κατασκευασμένοι ώστε να διατηρούν όσο γίνεται την πληροφορία που υπήρχε στις αρχικές μεταβλητές.
- Να δημιουργήσουμε νέες μεταβλητές, τους παράγοντες, στις οποίες μπορούμε με έναν υποκειμενικό τρόπο να αναγνωρίσουμε ως κάποιες μη μετρήσιμες μεταβλητές όπως π.χ. η ευφυΐα στην ψυχολογία ή η ελκυστικότητα ενός προϊόντος στο Μάρκετινγκ .
- Να εξηγήσουμε τις συσχετίσεις που υπάρχουν στα δεδομένα, για τις οποίες έχουμε υποθέσει ότι οφείλονται αποκλειστικά στην ύπαρξη κάποιων κοινών παραγόντων που δημιούργησαν τα δεδομένα.

Αυτό που πρέπει να έχει κανείς υπόψη του είναι πως η παραγοντική ανάλυση προσπαθεί περισσότερο να ερμηνεύσει τη δομή παρά τη μεταβλητότητα.

Οι διαφορές της με την Ανάλυση σε Κύριες Συνιστώσες είναι

- Στην παραγοντική ανάλυση υπάρχει ένα δομημένο μοντέλο και κάποιες υποθέσεις. Από αυτή την άποψη είναι μια στατιστική τεχνική κάτι που δεν ισχύει με την ανάλυση σε κύριες συνιστώσες η οποία είναι καθαρά ένας μαθηματικός μετασχηματισμός.
- Στην ανάλυση σε κύριες συνιστώσες το ενδιαφέρον στηρίζεται στο να εξηγηθεί η διακύμανση ενώ με την παραγοντική ανάλυση εξηγούμε την συνδιακύμανση των μεταβλητών.

Η παραγοντική ανάλυση έχει δεχτεί πολλές κριτικές από πολλούς επιστήμονες. Τα κυριότερα προβλήματα που συνδυάζονται με την παραγοντική ανάλυση είναι ότι

- Στηρίζεται σε ένα πλήθος υποθέσεων οι οποίες δεν είναι απαραίτητα ρεαλιστικές για πραγματικά προβλήματα και συνήθως ο ερευνητής δεν μπορεί να τις ελέγξει εύκολα.
- Δεν έχει μοναδική λύση. Όπως θα δούμε στη συνέχεια, μπορούμε να χρησιμοποιήσουμε διάφορες μεθόδους εκτίμησης, και ακόμα και για την ίδια μέθοδο εκτίμησης μπορούμε να πάρουμε ένα μεγάλο αριθμό ισοδύναμων εκτιμήσεων. Έτσι βασισμένοι στα ίδια δεδομένα διαφορετικοί επιστήμονες θα μπορούσαν να καταλήξουν σε διαφορετικά αποτελέσματα. Το πόσο διαφορετικά εξαρτάται και αυτό από διάφορα άλλα στοιχεία
- Οι παράγοντες οι οποίοι προκύπτουν μπορούν να δεχτούν διαφορετικές ερμηνείες οι οποίες μπορεί και να έρχονται σε αντιπαράθεση. Συνδυάζοντας το με την προηγούμενη παρατήρηση, μπορούμε από τα ίδια δεδομένα να καταλήξουμε σε εντελώς διαφορετικές ερμηνείες κάτι που επιστημονικά δεν είναι αποδεκτό.
- Ο αριθμός των παραγόντων που χρειάζεται να εξάγουμε ώστε τα αποτελέσματα να είναι χρήσιμα, δεν είναι προφανής κι εξαρτάται και από τη μέθοδο εκτίμησης που θα χρησιμοποιηθεί. Αυτό επιτρέπει στον επιστήμονα να δουλεύει σε μια μεροληπτική βάση έτσι ώστε να εμφανίζει τα αποτελέσματα όπως τον συμφέρουν.

Παρόλα αυτά η παραγοντική ανάλυση αποτελεί πολύτιμο εργαλείο σε πολλές επιστήμες και κυρίως στην Ψυχομετρία και την έρευνα αγοράς. Ο βασικός λόγος είναι πως αποτελεί μεθοδολογία για την ποσοτικοποίηση μη παρατηρήσιμων ποσοτήτων οι οποίες εμφανίζονται συχνά σε αυτές τις επιστήμες.

8.2 Το Ορθογώνιο Μοντέλο

Στο ορθογώνιο μοντέλο της παραγοντικής ανάλυσης, το οποίο είναι και το πιο διαδεδομένο, υποθέτουμε πως οι όποιες συσχετίσεις μεταξύ των μεταβλητών οφείλονται αποκλειστικά στην ύπαρξη κάποιων κοινών παραγόντων τους οποίους δεν ξέρουμε και θέλουμε να εκτιμήσουμε.

Έτσι υποθέτουμε πως οι p μεταβλητές μας μπορούν να γραφτούν ως γραμμικός συνδυασμός των k παραγόντων, δηλαδή

$$\mathbf{X}-\boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$$

όπου

$\mathbf{X}$ είναι το διάνυσμα των αρχικών μεταβλητών μεγέθους p x 1 (υποθέτω ότι έχω p μεταβλητές),

$\boldsymbol{\mu}$ είναι το διάνυσμα των μέσων μεγέθους p x 1,

L είναι ένας πίνακας p x k όπου το L_{ij} είναι η επιβάρυνση (loading) του παράγοντα F_j στη μεταβλητή X_i ,

F είναι ένας k x 1 πίνακας με τους παράγοντες και

ε είναι το σφάλμα ή μοναδικός παράγοντας . Το σφάλμα ε_i είναι ο μοναδικός παράγοντας της i μεταβλητής και είναι το μέρος της μεταβλητής το οποίο δεν μπορεί να εξηγηθεί από τους παράγοντες.

Μπορούμε να υποθέσουμε πως όλες οι μεταβλητές έχουν μέσο 0 οπότε το διάνυσμα **μ** δεν χρειάζεται στο παραπάνω μοντέλο (αυτό μπορεί να επιτευχθεί εύκολα αφαιρώντας από κάθε μεταβλητή τη μέση της τιμή). Επίσης είναι προφανές ότι $k<p$, δηλαδή ο αριθμός των παραγόντων πρέπει να είναι μικρότερος του αριθμού των μεταβλητών γιατί αλλιώς θα ήταν χωρίς νόημα να γίνει παραγοντική ανάλυση. Σύμφωνα με τα παραπάνω υποθέτουμε ότι κάθε μεταβλητή μπορούμε να την γράψουμε στη μορφή

$$X_1 = L_{11}F_1 + L_{12}F_2 + ... + L_{1k}F_k + \varepsilon_1$$
$$X_2 = L_{21}F_1 + L_{22}F_2 + ... + L_{2k}F_k + \varepsilon_2$$
$$...$$
$$X_p = L_{p1}F_1 + L_{p2}F_2 + ... + L_{pk}F_k + \varepsilon_p$$

Να σημειωθεί ότι

- Το παραπάνω μοντέλο αν και μοιάζει με ένα γραμμικό μοντέλο έχει μερικές διαφορές. Κατά αρχάς τα X_i δεν είναι παρατηρήσεις αλλά μεταβλητές. Αφετέρου το δεξί μέλος της εξίσωσης δεν είναι παρατηρήσιμο και έτσι πρέπει να εκτιμηθεί.

- Οι παράγοντες F_i μπορούν να γραφτούν και αυτοί σαν γραμμικός συνδυασμός των μεταβλητών. Αυτό είναι χρήσιμο να γίνεται όταν θέλουμε να δημιουργήσουμε νέες μεταβλητές. Θα πρέπει όμως να γίνει σαφές ότι οι συντελεστές αυτοί διαφέρουν από τις επιβαρύνσεις και δεν πρέπει να γίνεται σύγχυση. Οι συντελεστές κάθε παράγοντα όταν εκφράζουμε τις μεταβλητές ως γραμμικό συνδυασμό των παραγόντων καλούνται επιβαρύνσεις ενώ αντίστοιχα οι συντελεστές κάθε μεταβλητής όταν εκφράζουμε κάθε παράγοντα ως γραμμικό συνδυασμό των μεταβλητών καλούνται συντελεστές των σκορ (factor scores coefficients)

- Παρατηρείστε πως οι παράγοντες έχουν την ίδια διακύμανση. Αυτό αποτελεί βασική διαφορά από την ανάλυση σε κύριες συνιστώσες όπου θέλαμε οι κύριες συνιστώσες να είναι σε φθίνουσα τάξη διακύμανσης. Συνεπώς οι παράγοντες που προκύπτουν δεν είναι απαραίτητα σε κάποια σειρά (αν και αυτό, όπως θα δούμε στη συνέχεια, εξαρτάται και από τη μέθοδο εκτίμησης).

- Μια θεμελιώδης διαφορά με την ανάλυση σε κύριες συνιστώσες είναι πως εδώ το μοντέλο προσπαθεί να εκφράσει τις μεταβλητές ως γραμμικό συνδυασμό των παραγόντων ενώ στην ανάλυση σε κύριες συνιστώσες νοιαζόμασταν περισσότερο να εκφράσουμε τις κύριες συνιστώσες ως γραμμικό συνδυασμό των αρχικών μεταβλητών.

8.3 Υποθέσεις του Ορθογώνιου Μοντέλου

Ένα πολύ βασικό κομμάτι του παραγοντικού μοντέλου είναι οι υποθέσεις που πρέπει να γίνουν. Αυτές είναι:

1. $E(\mathbf{F})=\mathbf{0}$
2. $Cov(\mathbf{F}) = \mathbf{I}$
3. $E(\boldsymbol{\varepsilon})=\mathbf{0}$
4. $Cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$ όπου $\boldsymbol{\Psi}$ είναι ένας διαγώνιος πίνακας της μορφής

$$\Psi = \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & \psi_p \end{bmatrix}$$

5. $Cov(\boldsymbol{\varepsilon}_i, \mathbf{F}_j)=0$, για κάθε $i \neq j$.

Δηλαδή υποθέτουμε πως οι μοναδικοί παράγοντες και οι κοινοί παράγοντες είναι ασυσχέτιστοι (υπόθεση 5). Επίσης από τις παραπάνω υποθέσεις έχουμε πως τόσο οι παράγοντες όσο και οι μοναδικοί παράγοντες είναι ασυσχέτιστοι μεταξύ τους (υποθέσεις 2 και 4) κι έχουν μηδενικές μέσες τιμές (υποθέσεις 1 και 3). Σημειώστε επίσης πως υποθέτουμε ότι τα δεδομένα προέρχονται από πολυμεταβλητούς κανονικούς πληθυσμούς. Αυτή η υπόθεση χρησιμοποιείται ως βάση για ελέγχους καλής προσαρμογής του μοντέλου καθώς και για την εκτίμηση με τη μέθοδο μεγίστης πιθανοφάνειας. Συνεπώς μπορεί να αγνοηθεί στην περίπτωση που δουλεύουμε με άλλες μεθόδους εκτίμησης.

Η υπόθεση 2 σημαίνει ότι οι παράγοντες είναι ορθογώνιοι μεταξύ τους. Για αυτό το λόγο ονομάζουμε το μοντέλο ως ορθογώνιο. Αυτό δεν είναι καθόλου ρεαλιστικό σε πραγματικές εφαρμογές. Αν επιτρέψουμε κάποια μορφή συσχέτισης τότε μπορούμε να ορίσουμε ένα γενικότερο μοντέλο παραγοντικής ανάλυσης το οποίο δεν είναι ορθογώνιο. Παρατηρείστε επίσης πως οι διακυμάνσεις των παραγόντων είναι ίσες με τη μονάδα, άρα όλοι οι παράγοντες έχουν την ίδια διακύμανση.

Από τις παραπάνω υποθέσεις μπορεί να δειχθεί ότι

$$\begin{aligned} \boldsymbol{\Sigma} &= Cov(\mathbf{X}) = Cov(\mathbf{LF} + \boldsymbol{\varepsilon}) = \\ &= \mathbf{L}Cov(\mathbf{F})\mathbf{L}' + Cov(\boldsymbol{\varepsilon}) = \mathbf{LL}' + \Psi \end{aligned}$$

καθώς από τις υποθέσεις του μοντέλου η συνδιακύμανση μεταξύ $\mathbf{F}$ και $\boldsymbol{\varepsilon}$ είναι μηδέν. Συνεπώς βλέπουμε πως ο πίνακας διακύμανσης μπορεί να διασπαστεί σε δυο μέρη, το πρώτο είναι το κομμάτι που ερμηνεύουν οι κοινοί παράγοντες και ονομάζεται *εταιρικότητα*

(communality) και το δεύτερο το κομμάτι που οφείλεται στους μοναδικούς παράγοντες, και άρα το μοντέλο δεν μπορεί να ερμηνεύσει και ονομάζεται *ιδιαιτερότητα (specificity).*

Επίσης παρατηρείστε ότι η επιβάρυνση είναι η συσχέτιση κάθε μεταβλητής με τον αντίστοιχο παράγοντα.

Στην παραγοντική ανάλυση σκοπός μας είναι να εκτιμήσουμε τους πίνακες **L** και **Ψ,** να αναπαραστήσουμε δηλαδή τον πίνακα διακύμανσης του πληθυσμού. Για να το επιτύχουμε αυτό έχουν αναπτυχθεί διάφορες μέθοδοι εκτίμησης τις οποίες θα εξετάσουμε αργότερα.

Τα βήματα για να κάνω παραγοντική ανάλυση πρέπει να είναι τα εξής:

- Έλεγχος για το αν υπάρχουν συσχετίσεις ικανοποιητικές να κάνω παραγοντική ανάλυση
- Εύρεση του αριθμού των παραγόντων και εκτίμηση των παραμέτρων του μοντέλου
- Περιστροφή του μοντέλου με σκοπό να αυξήσω την ερμηνευτική του ικανότητα
- Εκτίμηση των *scor* των παραγόντων για περαιτέρω στατιστική χρήση

Στη συνέχεια θα αναφερθούμε αναλυτικά σε κάθε ένα από αυτά τα βήματα.

8.4 Έλεγχος Συσχετίσεων

Όπως όλες οι στατιστικές μέθοδοι, έτσι και στην παραγοντική ανάλυση πρέπει να ξεκινώ εξετάζοντας περιγραφικά τα δεδομένα. Για την παραγοντική ανάλυση είναι σημαντικό να υπάρχουν συσχετίσεις ανάμεσα στις μεταβλητές καθώς αυτές τις συσχετίσεις θα προσπαθήσω να εξηγήσω. Επομένως πρέπει να ξεκινήσω από τις συσχετίσεις.

Όπως είπαμε αν τα δεδομένα είναι σχετικά ασυσχέτιστα δεν έχει νόημα να συνεχίσω αφού αυτό σημαίνει ότι δε θα βρω κοινούς παράγοντες που να μου επιτρέψουν να δουλέψω με αυτούς. Τι σημαίνει όμως μεγάλες συσχετίσεις; Σε καμιά περίπτωση δεν σημαίνει στατιστικά σημαντικές συσχετίσεις, δηλαδή συσχετίσεις διάφορες του μηδέν. Είναι γνωστό στη στατιστική ότι όσο αυξάνει το μέγεθος του δείγματος τότε συσχετίσεις κοντά στο μηδέν τείνουν να είναι στατιστικά σημαντικά διάφορες του μηδέν αν και πολύ μικρές σε απόλυτη τιμή. Συνεπώς αυτό που μας ενδιαφέρει είναι να υπάρχουν μεγάλες συσχετίσεις τουλάχιστον σε μεγάλο ποσοστό του πίνακα συσχετίσεων. Τιμές μεγαλύτερες του 0.40 σε απόλυτη τιμή είναι ευπρόσδεκτες. Σε αντίθετη περίπτωση δεν έχει έννοια να συνεχίσουμε. Αν υπάρχουν κάποια ή κάποιες μεταβλητές που είναι ασυσχέτιστες με τις υπόλοιπες καλό είναι να τις αγνοήσουμε καθώς, επειδή δεν σχετίζονται με τις άλλες, θα προκύψουν από μόνες τους ως ένας ξεχωριστός παράγοντας.

Για να ελέγξουμε τη στατιστική σημαντικότητα ενός δειγματικού συντελεστή συσχέτισης χρειαζόμαστε κάποιον έλεγχο. Έτσι για να ελέγξουμε την

H_0: $\rho=0$ έναντι της εναλλακτικής

H_1: $\rho\neq 0$,

υπολογίζουμε την ελεγχοσυνάρτηση

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

η οποία ακολουθεί την t κατανομή με n-2 βαθμούς ελευθερίας. Συνεπώς απορρίπτουμε τη μηδενική υπόθεση όταν $|t| \geq t_{1-a/2,n-2}$ όπου $t_{a,n-2}$ είναι το a% ποσοστιαίο σημείο από την κατανομή t με $n-2$ βαθμούς ελευθερίας. Παρατηρείστε ότι:

- Ο παρονομαστής της ελεγχοσυνάρτησης είναι το τυπικό σφάλμα της δειγματικής εκτιμήτριας της συσχέτισης.
- Ο παραπάνω έλεγχος δεν μπορεί να χρησιμοποιηθεί για να ελέγξουμε υποθέσεις διάφορες από τη μορφή $\rho=0$, κι αυτό γιατί η κατανομή της ελεγχοσυνάρτησης παύει να είναι η κατανομή t.
- Απόρριψη της μηδενικής υπόθεσης σημαίνει ότι υπάρχει μη μηδενική συσχέτιση. Για σχετικά μέτρια μεγέθη δείγματος ακόμα και δειγματικές συσχετίσεις της τάξης του 0.10 σε απόλυτη τιμή τείνουν να είναι στατιστικά σημαντικές αν και μια τέτοια συσχέτιση είναι εξαιρετικά χαμηλή. Συνεπώς μη μηδενική συσχέτιση δεν σημαίνει και ότι υπάρχει κάποια έντονη σχέση μεταξύ των μεταβλητών.
- Ο συντελεστής συσχέτισης r του Pearson εξετάζει μόνο γραμμικής μορφής συσχέτιση και συνεπώς δεν μπορεί να διαγνώσει άλλες μορφές συσχέτισης.

Για να ελέγξουμε υποθέσεις της μορφής

H_0: $\rho=\rho_0$ έναντι της εναλλακτικής

H_1: $\rho\neq\rho_0$,

υπολογίζουμε την ελεγχοσυνάρτηση

$$Z = \frac{z(r) - z(\rho_0)}{\sqrt{\frac{1}{n-3}}}$$

όπου $z(a) = 0.5\ln\left(\frac{1+a}{1-a}\right)$, γνωστός και ως μετασχηματισμός του Fisher. Η ελεγχοσυνάρτηση ακολουθεί προσεγγιστικά την τυποποιημένη κανονική κατανομή και άρα απορρίπτουμε όταν η τιμή είναι μεγαλύτερη σε απόλυτη τιμή από το αντίστοιχο ποσοστιαίο

σημείο της κανονικής κατανομής. Για τον έλεγχο χρειάζεται η τιμή κάτω από τη μηδενική υπόθεση να μην είναι κοντά στο ± 1.

Για να ελέγξουμε αν υπάρχουν συσχετίσεις στα δεδομένα μας μπορούμε να χρησιμοποιήσουμε τον έλεγχο σφαιρικότητας του Bartlett (Bartlett's test of sphericity). Αν τα δεδομένα ήταν ασυσχέτιστα τότε το νέφος των σημείων θα ήταν μια υπερσφαίρα στο πολυεπίπεδο. Ο έλεγχος ελέγχει την υπόθεση

H_0: $\Sigma = \sigma^2 I_p$ έναντι της εναλλακτικής

H_1: $\Sigma \neq \sigma^2 I_p$.

Ο πιο πάνω έλεγχος είναι γενικότερος καθώς εξετάζει αν ο πίνακας διακύμανσης είναι διαγώνιος. Ένας έλεγχος βασισμένος στο λόγο πιθανοφανειών υπολογίζει τη στατιστική συνάρτηση

$$L = -\left[n - \frac{1}{6p}(2p^2 + p + 2) \right] \left[\ln|\mathbf{S}| - \ln\left(\prod_{i=1}^{p} s_i^{\,2} \right) \right]$$

και απορρίπτουμε τη μηδενική υπόθεση συγκρίνοντας την τιμή αυτή με το ποσοστιαίο σημείο της χ^2 κατανομής με $p(p-1)/2$ βαθμούς ελευθερίας. Στον παραπάνω τύπο $\boldsymbol{S}$ είναι ο δειγματικός πίνακας διακύμανσης συνδιακύμανσης, s_i^2 είναι η δειγματική διακύμανση της i μεταβλητής και ο πρώτος όρος του γινομένου είναι η διόρθωση που προτάθηκε από τον Bartlett έτσι ώστε η κατανομή της ελεγχοσυνάρτησης να προσεγγίζεται καλά από μια χ^2 κατανομή.

Στην περίπτωση που θέλουμε να ελέγξουμε για έναν πίνακα συσχέτισης, η μηδενική υπόθεση παίρνει τη μορφή

H_0: $R = I_p$ έναντι της εναλλακτικής

H_1: $R \neq I_p$,

δηλαδή ελέγχουμε την υπόθεση πως ο πίνακας συσχετίσεων του πληθυσμού είναι ο μοναδιαίος. Η ελεγχοσυνάρτηση που χρησιμοποιούμε είναι η

$$L = -\left[n - \frac{1}{6(2p+5)} \right] \ln|\mathbf{R}|$$

η οποία και ακολουθεί και πάλι χ^2 κατανομή με $p(p-1)/2$ βαθμούς ελευθερίας. Ο έλεγχος αυτός είναι στην πραγματικότητα ένας έλεγχος λόγου πιθανοφανειών κάτω από τις δύο υποθέσεις. Παρατηρείστε πως ο πολλαπλασιαστής μπροστά από το λογάριθμο της ορίζουσας δεν είναι ο ίδιος με αυτόν που είχαμε πριν στον έλεγχο για έναν πίνακα

διακύμανσης. Αυτό οφείλεται στο γεγονός πως η προσέγγιση από την κατανομή χ^2 είναι καλύτερη με τη χρήση του καινούριου πολλαπλασιαστή.

Μερικός συντελεστής συσχέτισης.

Ο απλός συντελεστής συσχέτισης υπολογίζει τη συσχέτιση μεταξύ δυο μεταβλητών αγνοώντας τις υπόλοιπες. Έτσι μπορεί να εμφανίζει συσχετισμένες κάποιες μεταβλητές απλά και μόνο επειδή κάποιες άλλες έχουν μεγάλη συσχέτιση με αυτές και όταν ακυρώσουμε την επίδρασή τους οι αρχικές μεταβλητές να μην εμφανίζουν πια καμιά συσχέτιση (όπως στη γραμμική παλινδρόμηση όπου προσθέτοντας κάποια μεταβλητή σε ένα μοντέλο μπορεί μια μεταβλητή που ήταν πριν σημαντική να πάψει να είναι). Για αυτό είναι χρήσιμος ένας συντελεστής συσχέτισης ο οποίος θα υπολογίζει τη συσχέτιση αφού αφαιρέσει την επίδραση των υπόλοιπων μεταβλητών. Αυτός είναι ο μερικός συντελεστής συσχέτισης. Ο τρόπος υπολογισμού του είναι αρκετά πολύπλοκος και συνήθως γίνεται με τη χρήση υπολογιστή. Για να προχωρήσουμε σε παραγοντική ανάλυση μας ενδιαφέρει οι μερικοί συντελεστές συσχέτισης να είναι μικροί.

Αν οι μεταβλητές μοιράζονται κοινούς παράγοντες θα περίμενε κανείς ότι ο μερικός συντελεστής συσχέτισης ανάμεσα σε δύο μεταβλητές, όταν ακυρωθεί η επίδραση όλων των υπολοίπων μεταβλητών, θα είναι μικρή, αφού η ακύρωση της επίδρασης των υπολοίπων μεταβλητών ακυρώνει σε μεγάλο βαθμό την επίδραση των κοινών παραγόντων. Επομένως οι μερικοί συντελεστές συσχέτισης είναι εκτιμήσεις των συσχετίσεων μεταξύ των μοναδικών παραγόντων και θα πρέπει να είναι κοντά στο 0 όταν οι υποθέσεις του παραγοντικού μοντέλου ισχύουν. Θυμηθείτε ότι από τις υποθέσεις του μοντέλου οι μοναδικοί παράγοντες είναι ασυσχέτιστοι.

Ένα μέτρο για να συγκρίνουμε το σχετικό μέγεθος των συντελεστών συσχέτισης σχετικά με τους μερικούς συντελεστές συσχέτισης είναι το Kaiser-Meyer-Olkin στατιστικό που υπολογίζεται ως

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} a_{ij}^2},$$

όπου r_{ij} και a_{ij} είναι οι δειγματικοί συντελεστές συσχέτισης και μερικής συσχέτισης αντίστοιχα. Αν η τιμή του KMO είναι μεγάλη τότε τα δεδομένα μας είναι κατάλληλα για παραγοντική ανάλυση. Τιμές κάτω από 0.5 είναι πολύ κακές τιμές. Στην πράξη τιμές γύρω στο 0.8 θεωρούνται αρκετά καλές για να προχωρήσουμε. Μικρότερες τιμές αποτελούν ένδειξη ότι η παραγοντική ανάλυση δεν θα μας δώσει ικανοποιητικά αποτελέσματα.

Τέλος ένα άλλο μέτρο που μας επιτρέπει να εξετάσουμε μια-μια τις μεταβλητές και το κατά πόσο είναι κατάλληλες για να χρησιμοποιηθούν στην ανάλυση είναι το μέτρο της

δειγματικής καταλληλότητας (measure of sampling adequacy) το οποίο υπολογίζεται για την *i* μεταβλητή ως

$$MSA_i = \frac{\sum_j r_{ij}^2}{\sum_j r_{ij}^2 + \sum_j a_{ij}^2}$$

Τιμές κοντά στο 1 είναι ενδείξεις ότι η μεταβλητή είναι πολύ καλή για να χρησιμοποιηθεί στην ανάλυση.

8.5 Αριθμός Παραγόντων και Εκτίμηση των Παραγόντων

Ένα από τα βασικά ερωτήματα στην Παραγοντική Ανάλυση είναι ο καθορισμός του αριθμού των παραγόντων που θα χρησιμοποιήσουμε. Όπως είπαμε και προηγουμένως ο αριθμός αυτός δεν είναι γνωστός και υπάρχουν διάφορες μέθοδοι για να εκτιμηθεί. Πολλά στατιστικά πακέτα επιτρέπουν στον ερευνητή να καθορίσει εκ των προτέρων τον αριθμό αυτό αλλά γενικά αυτό γίνεται κυρίως για λόγους ευκολίας.

Για να βρεθεί ο αριθμός λοιπόν των παραγόντων ο ερευνητής μπορεί να χρησιμοποιήσει παρόμοιες τεχνικές με αυτές που είδαμε στην ανάλυση σε κύριες συνιστώσες. Δηλαδή, τις τιμές των ιδιοτιμών του πίνακα διακύμανσης συνδιακύμανσης, τιμές που εξηγούν κάποιο ποσοστό της διακύμανσης ή το scree plot (το γράφημα των ιδιοτιμών ως προς τον αύξοντα αριθμό τους).

Παρατηρείστε ότι ο αριθμός των παραγόντων χρειάζεται να καθοριστεί πριν γίνει η εκτίμηση τους. Επομένως κάποιος θα μπορούσε να δουλέψει με διαδοχικά αυξανόμενο αριθμό παραγόντων και να κρατήσει το μοντέλο με βάση κάποιο κριτήριο καλής προσαρμοστικότητας. Τέτοια κριτήρια είναι:

- Από τον πίνακα των επιβαρύνσεων μπορεί κάποιος να εκτιμήσει τον πίνακα **Σ**. Οι αποκλίσεις του πραγματικού πίνακα με τον εκτιμημένο (συνήθως ονομάζεται reproduced matrix) θα πρέπει να είναι μικρές. Δυστυχώς δεν υπάρχει ένα κριτήριο του πόσο μικρές.

- Έλεγχος λόγου πιθανοφανειών αν οι εκτιμήσεις έχουν γίνει με τη μέθοδο μεγίστης πιθανοφάνειας. Τέτοιοι έλεγχοι στηρίζονται σε υποθέσεις για την κατανομή του πληθυσμού.

Σημειώστε επίσης πως:

- Η ερμηνεία των παραγόντων μπορεί να εξαρτάται και από τον αριθμό τους, δηλαδή προσθέτοντας παράγοντες αυτοί να παύουν να έχουν την ίδια ερμηνεία (αν και αυτό είναι μια ένδειξη ακαταλληλότητας του μοντέλου).

- Για μερικές μεθόδους εκτίμησης υπάρχει περιορισμός στον αριθμό των παραγόντων που μπορούν να εκτιμηθούν.

Οι δύο βασικές μέθοδοι εκτίμησης που χρησιμοποιούνται στην πράξη είναι η μέθοδος των κυρίων συνιστωσών και η μέθοδος μεγίστης πιθανοφάνειας. Συγκριτικά έχουμε:

- Όταν εκτιμούμε το μοντέλο με τη μέθοδο των κυρίων συνιστωσών, προσθέτοντας παράγοντες δεν αλλάζουν οι επιβαρύνσεις των παραγόντων που είχαμε πάρει πριν. Αυτό δεν ισχύει με τη μέθοδο μεγίστης πιθανοφάνειας όπου προσθέτοντας παράγοντες αλλάζουν οι επιβαρύνσεις των προηγούμενων παραγόντων και άρα η ερμηνεία τους.
- Με τη μέθοδο μεγίστης πιθανοφάνειας μπορούμε να κάνουμε ελέγχους καλής προσαρμογής του μοντέλου βασισμένοι στον κλασικό έλεγχο λόγου πιθανοφανειών
- Η μέθοδος των κυρίων συνιστωσών εξαρτάται από τις μονάδες μέτρησης κι έτσι αν αλλάξουν μπορεί να αλλάξει ριζικά η λύση που έχουμε πάρει. Αυτό δεν ισχύει με τη μέθοδο μεγίστης πιθανοφάνειας που είναι ανεξάρτητη των μονάδων μέτρησης. Έτσι ενώ στη μέθοδο κυρίων συνιστωσών πρέπει να διαλέξω ανάμεσα στον πίνακα διακύμανσης και τον πίνακα συσχέτισης στη μέθοδο μεγίστης πιθανοφάνειας δεν έχω τέτοιο πρόβλημα.
- Η μέθοδος των κυρίων συνιστωσών δεν βάζει περιορισμούς στον αριθμό των παραγόντων που μπορούμε να εκτιμήσουμε.
- Όταν η μέθοδος μεγίστης πιθανοφάνειας δεν δουλεύει αυτό είναι μια ένδειξη ότι υπάρχει πρόβλημα με το μοντέλο. Αντίθετα η μέθοδος κυρίων συνιστωσών επειδή είναι στην ουσία ένας μαθηματικός μετασχηματισμός των δεδομένων δουλεύει πάντα χωρίς όμως να μας δίνει κάποια ένδειξη αν καλώς δουλεύει ή όχι.
- Με τη μέθοδο μεγίστης πιθανοφάνειας τα *scor* των παραγόντων δεν μπορούν να υπολογιστούν ακριβώς όπως συμβαίνει με τη μέθοδο κυρίων συνιστωσών.

8.5.1 Εκτίμηση με τη μέθοδο Κυρίων Συνιστωσών

Η εκτίμηση με τη μέθοδο των κυρίων συνιστωσών βασίζεται στη φασματική ανάλυση του πίνακα διακύμανσης (συσχέτισης). Όταν λέμε πως θέλουμε να εκτιμήσουμε τις παραμέτρους του παραγοντικού μοντέλου εννοούμε πως θέλουμε να εκτιμήσουμε τα στοιχεία του πίνακα επιβαρύνσεων $\mathbf{L}$ και τα στοιχεία της διαγωνίου του πίνακα $\mathbf{\Psi}$.

Παρατηρείστε πως το πλήθος των στοιχειών του πίνακα **L** έχει να κάνει με το πλήθος των παραγόντων που έχουμε υποθέσει πως υπάρχουν. Επομένως σκοπός μας είναι να βρούμε πίνακες $\hat{\mathbf{L}}, \hat{\mathbf{\Psi}}$ για τους όποιους ο πίνακας $\hat{\mathbf{L}}\hat{\mathbf{L}}'+\hat{\mathbf{\Psi}}$ να είναι όσο γίνεται πιο κοντά στον πίνακα δειγματικής διακύμανσης (συσχέτισης).

Από τη φασματική ανάλυση ενός πίνακα διακύμανσης γνωρίζουμε πως μπορούμε να τον γράψουμε στη μορφή $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}'$, όπου $\mathbf{A} = \mathbf{\Pi}\mathbf{\Lambda}^{1/2}$, $\mathbf{\Lambda}$ είναι ο διαγώνιος πίνακας που περιέχει στη διαγώνιο τις ιδιοτιμές και $\mathbf{\Pi}$ είναι ο πίνακας με στήλες τα ιδιοδιανύσματα του πίνακα $\mathbf{\Sigma}$ Επομένως αν χρησιμοποιήσουμε ως $\hat{\mathbf{L}} = \mathbf{\Pi}\mathbf{\Lambda}^{1/2}$ τότε μπορούμε να αναπαραστήσουμε πλήρως τον πίνακα **Σ.** Στην πράξη δουλεύουμε με το δειγματικό πίνακα διακύμανσης **S.**

Αν το πλήθος των παραγόντων *k* είναι ίδιο με το πλήθος των μεταβλητών *p*, επιτυγχάνουμε την πλήρη αναπαράσταση του δειγματικού πίνακα διακύμανσης (συσχέτισης) και επομένως οι εκτιμήσεις των ιδιαιτεροτήτων ψ_i είναι 0, δηλαδή οι παράγοντες εξηγούν όλη τη διακύμανση.

Αν $k < p$ τότε ο πίνακας $\hat{\mathbf{L}}\hat{\mathbf{L}}'$ δεν μπορεί να αναπαραστήσει πλήρως τον αρχικό πίνακα διακύμανσης. Έτσι σε αυτή την περίπτωση μπορούμε να εκτιμήσουμε και τις ιδιαιτερότητες ως

$$\hat{\psi}_i = s_i^2 - \sum_{j=1}^{p} L_{ij}^2 ,$$

όπου L_{ij} είναι το *ij* στοιχείο του πίνακα $\hat{\mathbf{L}}\hat{\mathbf{L}}'$, δηλαδή η επιβάρυνση του *j* παράγοντα στην *i* μεταβλητή, *j=1,. . . , k* και *i=1, . . ., p.* Ο δεύτερος όρος στο δεξί μέλος της ισότητας είναι η εταιρικότητα της μεταβλητής.

Μια εναλλακτική παρουσίαση της εκτίμησης των ιδιαιτεροτήτων είναι πως αυτές εκτιμώνται ως τα διαγώνια στοιχεία του πίνακα $\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}'$ (προσοχή, μόνο τα διαγώνια). Επομένως μετά την εκτίμηση του μοντέλου, αν έχουμε χρησιμοποιήσει λιγότερους παράγοντες από το πλήθος των μεταβλητών μας, θα υπάρχουν κατάλοιπα ανάμεσα στον πραγματικό πίνακα διακύμανσης από όπου ξεκινήσαμε και τον εκτιμηθέντα από το μοντέλο.

Μερικές χρήσιμες παρατηρήσεις σχετικά με τη μέθοδο κυρίων συνιστωσών είναι οι εξής:

- Αν χρησιμοποιήσουμε πολλούς παράγοντες μπορούμε να αναπαραστήσουμε πλήρως τον αρχικό πίνακα. Σε αυτήν όμως την περίπτωση δεν έχουμε κερδίσει κάτι σημαντικό αφού χρησιμοποιήσαμε πολλούς παράγοντες και στην ουσία απλά μετασχηματίσαμε τα δεδομένα μας
- Δεν υπάρχει περιορισμός ως προς τον αριθμό των παραγόντων που μπορώ να εκτιμήσω με τη μέθοδο κυρίων συνιστωσών.
- Είναι διαφορετικό πράγμα να χρησιμοποιώ την μέθοδο ανάλυσης σε κύριες συνιστώσες και το να χρησιμοποιώ τη μέθοδο κυρίων συνιστωσών για να εκτιμήσω

το παραγοντικό μοντέλο. Μπορεί τα ονόματα να είναι όμοια αλλά υπάρχουν σημαντικές διαφορές. Η μια είναι αυτοτελής μέθοδος ανάλυσης και η άλλη απλά ένα εργαλείο εκτίμησης του παραγοντικού μοντέλου.

8.5.2 Εκτίμηση με τη μέθοδο μεγίστης πιθανοφάνειας

Για να χρησιμοποιήσουμε τη μέθοδο μεγίστης πιθανοφάνειας χρειάζεται να κάνουμε κάποιες υποθέσεις σχετικά με τον πληθυσμό από όπου προήλθαν τα δεδομένα μας.

Συγκεκριμένα υποθέτουμε πως τα σφάλματα (μοναδικοί όροι) ακολουθούν πολυμεταβλητή κανονική κατανομή με διάνυσμα μέσων το μηδενικό διάνυσμα και πίνακα διακύμανσης το διαγώνιο πίνακα **Ψ,** δηλαδή $\boldsymbol{\varepsilon} \sim N_p(\mathbf{0}, \boldsymbol{\Psi})$. Παρατηρείστε ότι η παραπανίσια υπόθεση είναι η μορφή της κατανομής. Επομένως το διάνυσμα των τυχαίων μεταβλητών **X** δοθέντος του διανύσματος των παραγόντων **F** ακολουθεί την πολυδιάστατη κανονική κατανομή, δηλαδή $\mathbf{X} \mid \mathbf{F} \sim N_p(\mathbf{LF}, \boldsymbol{\Psi})$ και άρα αν υποθέσουμε πως και οι παράγοντες προέρχονται από πολυδιάστατη κανονική κατανομή, δηλαδή $\mathbf{F} \sim N_k(\mathbf{0}, \mathbf{I})$ προκύπτει πως $\mathbf{X} \sim N_k(\mathbf{LF}, \mathbf{LL'}+\boldsymbol{\Psi})$.

Δηλαδή οι παραπανίσιες υποθέσεις είχαν να κάνουν με την κανονικότητα των σφαλμάτων και των παραγόντων. Άρα τώρα έχουμε ένα παραμετρικό μοντέλο και τα δεδομένα μας προέρχονται από πολυμεταβλητή κανονική κατανομή. Αυτό αφενός σημαίνει πως έχουμε να ελέγξουμε μια υπόθεση η οποία μάλιστα δεν είναι εύκολο να ελεγχθεί (πολυμεταβλητή κανονικότητα) αλλά αφετέρου μπορούμε να κάνουμε στατιστική συμπερασματολογία. Επίσης η υπόθεση της κανονικότητας ισοδυναμεί με το ότι οι μεταβλητές μας είναι συνεχείς.

Αν λοιπόν έχουμε ένα δείγμα από πολυμεταβλητή κανονική κατανομή μπορεί αν δειχτεί ότι η πιθανοφάνεια είναι ως συνάρτηση του πίνακα διακύμανσης **Σ** του πληθυσμού

$$\ell(X, \boldsymbol{\Sigma}) = -\frac{n}{2}\left[p\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + tr(\boldsymbol{\Sigma}^{-1}\mathbf{S})\right],$$

όπου *n* είναι το μέγεθος του δείγματος, *p* ο αριθμός των μεταβλητών και **S** ο δειγματικός πίνακας διακυμάνσεων. Από την παραπάνω πιθανοφάνεια έχουμε εξαφανίσει το διάνυσμα των μέσων **μ** αφού αυτό δεν επηρεάζει το μοντέλο μας (ή ισοδύναμα έχουμε κεντροποιήσει όλες τις μεταβλητές να έχουν μέση τιμή 0). Για να εκτιμήσουμε το μοντέλο με τη μέθοδο μεγίστης πιθανοφάνειας πρέπει να μεγιστοποιήσουμε τη συνάρτηση

$$\ell(\mathbf{X}, \mathbf{L}, \boldsymbol{\Psi}) = -\frac{n}{2}\left[p\ln(2\pi) + \ln|\mathbf{LL'}+\boldsymbol{\Psi}| + tr((\mathbf{LL'}+\boldsymbol{\Psi})^{-1}\mathbf{S})\right]$$

ως προς **L** και **Ψ.** Αν το μοντέλο έχει *k* παράγοντες τότε ο πίνακας **L** έχει $p \times k$ στοιχεία ενώ ο πίνακας **Ψ** επειδή είναι διαγώνιος έχει *p* στοιχεία. Συνολικά έχουμε $(p+1)k$ παραμέτρους ενώ ο πίνακας **Σ** από όπου ξεκινάμε έχει $\frac{p}{2}(p+1)$ διαφορετικά στοιχεία

(θυμηθείτε πως είναι συμμετρικός). Για να έχει λύση λοιπόν θα πρέπει να βάλουμε περιορισμό στο *k* τον αριθμό των παραγόντων που μπορούμε να εκτιμήσουμε. Επομένως με τη μέθοδο μέγιστης πιθανοφάνειας υπάρχει περιορισμός στον αριθμό των παραγόντων (κάτι που δεν υπήρχε στη μέθοδο κυρίων συνιστωσών). Μπορεί κανείς να δει πως ο μέγιστος αριθμός *k* των παραγόντων που μπορούμε να εκτιμήσουμε είναι $\left[p/2\right]$ όπου [α] είναι το ακέραιο μέρος του *a*. Ο πίνακας 8.1 που ακολουθεί δείχνει πόσους παράγοντες μπορούμε να εκτιμήσουμε με τη μέθοδο μεγίστης πιθανοφάνειας για κάθε πλήθος μεταβλητών

p	3	4	5	7	10	12	15	20	30
Μέγιστο *k*	1	2	2	3	5	6	7	10	15

Πίνακας 8.1 Ο μέγιστος αριθμός παραγόντων που μπορούμε να εκτιμήσουμε με τη μέθοδο μεγίστης πιθανοφάνειας για κάθε πλήθος μεταβλητών *p*.

Επίσης για να μπορούμε να ταυτοποιήσουμε χρειαζόμαστε έναν ακόμα περιορισμό. Αυτός που συνήθως χρησιμοποιείται (και που χρησιμοποιούν τα περισσότερα στατιστικά πακέτα) είναι πως ο πίνακας $\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}$ είναι διαγώνιος και τα στοιχεία του είναι σε φθίνουσα σειρά.

Για να μεγιστοποιήσουνε αυτή την πιθανοφάνεια με τον περιορισμό που δώσαμε χρειάζονται αριθμητικές μέθοδοι και για αυτό πρέπει και να ορίσουμε κάποιο κριτήριο τερματισμού αυτών των μεθόδων. Επίσης δεν έχει σημασία με ποιον πίνακα (διακύμανσης ή συσχετίσεων) θα δουλέψουμε αφού η λύση είναι αδιάφορη των μονάδων μέτρησης.

Ένα από τα πλεονεκτήματα της μεθόδου μεγίστης πιθανοφάνειας είναι πως μας επιτρέπει να κάνουμε έλεγχο καλής προσαρμογής του ορθογώνιου μοντέλου που προσαρμόσαμε. Συγκεκριμένα ελέγχουμε τη μηδενική υπόθεση

H_0: $\mathbf{\Sigma} = \mathbf{LL}'+\mathbf{\Psi}$ έναντι της εναλλακτικής

H_1: δεν υπάρχει περιορισμός στον πίνακα $\mathbf{\Sigma}$

Η ελεγχοσυνάρτηση είναι η

$$LR = n(tr\mathbf{D} - \ln|\mathbf{D}| - p)$$

όπου $\mathbf{D} = (\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}})^{-1}\mathbf{S}$. Η μηδενική υπόθεση είναι ότι το μοντέλο προσαρμόζει καλά τα δεδομένα. Η τιμή του LR συγκρίνεται με την τιμή της X^2 κατανομής με $s=[(p-k)^2 - (p+k)]/2$ βαθμούς ελευθερίας και αν είναι μεγαλύτερη απορρίπτουμε την καλή προσαρμοστικότητα του μοντέλου. Αν το μοντέλο είναι καλό ο πίνακας $\mathbf{D}$ είναι περίπου ο μοναδιαίος πίνακας και άρα το ίχνος του είναι *p* και ο λογάριθμος της ορίζουσας 0. Επομένως η ελεγχοσυνάρτηση παίρνει την τιμή 0.

8.5.3 Κριτήρια Επιλογής Μοντέλου

Εκτός από κριτήρια καλής προσαρμογής, που είδαμε προηγουμένως, η μέθοδος μεγίστης πιθανοφάνειας μας επιτρέπει να κάνουμε και επιλογή μοντέλου, δηλαδή πόσοι παράγοντες μας δίνουν το καλύτερο αποτέλεσμα. Θυμηθείτε πως στην περίπτωση της μεθόδου κυρίων συνιστωσών ο μόνος τρόπος να το κάνουμε αυτό ήταν να χρησιμοποιήσουμε τα κατάλοιπα από τη διαφορά του πραγματικού πίνακα διακύμανσης με τον εκτιμώμενο πίνακα. Στη μέθοδο μεγίστης πιθανοφάνειας μπορούμε να χρησιμοποιήσουμε πληροφοριακά κριτήρια (information criteria) όπως χρησιμοποιούμε και σε άλλες στατιστικές μεθόδους (π.χ. γραμμική παλινδρόμηση). Έτσι για κάθε μοντέλο με *r* παράγοντες υπολογίζουμε είτε το Akaike Information Criterion που είναι

$$AIC(r) = -2\left[\frac{n}{2}\sum_{i=r+1}^{p}\ln\lambda_i\right] + \left[2p(r+1) - r(r-1)\right]$$

ή το κριτήριο του Schwartz που είναι

$$SIC(r) = -\left[\frac{n}{2}\sum_{i=r+1}^{p}\ln\lambda_i\right] + \left[\frac{p(r+1)}{2} - \frac{r(r-1)}{4}\right]\ln n\,,$$

όπου λ_i είναι οι ιδιοτιμές του πίνακα **S** σε φθίνουσα σειρά.

Επιλέγουμε για κάθε κριτήριο το μοντέλο με τη μικρότερη τιμή. Παρατηρείστε πως η λογική και των δύο κριτηρίων είναι να επιβάλουν κάποια ποινή για κάθε μοντέλο με περισσότερες παραμέτρους γιατί αλλιώς η πιθανοφάνεια είναι λογικό να αυξάνει προσθέτοντας παραμέτρους. Επομένως αυτή η ποινή αποζημιώνει για τις παραπανίσιες παραμέτρους. Το κριτήριο του Schwartz λαμβάνει υπόψη του στην ποινή αυτή τόσο των αριθμό των παραπανίσιων παραμέτρων αλλά και το μέγεθος του δείγματος κάτι το οποίο δεν συμβαίνει στην περίπτωση του AIC.

8.5.4 Άλλες μέθοδοι Εκτίμησης

Εκτός από τις δύο μεθόδους εκτίμησης που περιγράψαμε υπάρχουν αρκετές άλλες μέθοδοι εκτίμησης στη βιβλιογραφία. Η παραγοντική ανάλυση, αν και θεωρητικά ήταν γνωστή πολλά χρόνια, δεν ήταν εύκολο να χρησιμοποιηθεί λόγω της πολυπλοκότητας των υπολογισμών για την εκτίμηση των παραμέτρων. Έτσι διάφορες τεχνικές αναπτύχθηκαν με βασικό σκοπό να απλοποιηθεί η διαδικασία εκτίμησης. Μερικές από τις μεθόδους αυτές είναι

*Μέθοδος ελαχίστων τετραγώνων***:** Η μέθοδος αυτή προσπαθεί να ελαχιστοποιήσει το άθροισμα των τετραγωνικών διαφορών των πραγματικών συνδιακυμάνσεων με αυτές που το μοντέλο εκτιμά. Το πρόβλημα επομένως ανάγεται σε πρόβλημα ελαχίστων τετραγώνων το οποίο από

αρκετά χρόνια πριν ήταν σχετικά ευκολότερο να αντιμετωπιστεί. Στην πράξη η μέθοδος μπορεί να δώσει εκτιμήσεις σε προβλήματα που η μέθοδος μέγιστης πιθανοφάνειας αποτυγχάνει. Τα αποτελέσματα όμως αλλάζουν αν αλλάξει η κλίμακα. Όσο προσθέτουμε παράγοντες αλλάζει και η εκτίμηση των επιβαρύνσεων τους

Γενικευμένη μέθοδος ελαχίστων τετραγώνων. Η μέθοδος είναι παραλλαγή της προηγούμενης. Στη γραμμική παλινδρόμηση είναι γνωστό πως οι απλοί εκτιμητές ελαχίστων τετραγώνων δεν είναι συνεπείς όταν η διακύμανση δεν είναι σταθερή. Κάτι αντίστοιχο έχουμε και εδώ όπου η διακύμανση των τυχαίων όρων δεν είναι η ίδια για όλες τις μεταβλητές. Επομένως αυτή η μέθοδο χρησιμοποιεί ως βάρη τις αντίστροφες τιμές των μοναδικών διακυμάνσεων. Ισχύουν όλα τα πλεονεκτήματα και τα μειονεκτήματα που αναφέραμε για την απλή μέθοδο των ελαχίστων τετραγώνων

Μέθοδος των κυρίων αξόνων (Principal Axis Model): Η μέθοδος είναι παραλλαγή της μεθόδου των κυρίων συνιστωσών. Αντικαθιστά τις μονάδες στη διαγώνιο του πίνακα συσχέτισης με εκτιμήσεις της εταιρικότητας. Στην πραγματικότητα δίνει σε κάθε μεταβλητή διαφορετικό βάρος καθώς τα διαγώνια στοιχεία δεν είναι πια μονάδες. Η μέθοδος λειτουργεί επαναληπτικά και ξεκινά με αρχικές εκτιμήσεις για την εταιρικότητα κάθε μεταβλητής τον συντελεστή προσδιορισμού από τη γραμμική παλινδρόμηση που έχει τη μεταβλητή αυτή ως εξαρτημένη και τις υπόλοιπες ως ανεξάρτητες. Χρησιμοποιώντας τη λογική της μεθόδου των κυρίων συνιστωσών (υπολογίζοντας ιδιοτιμές και ιδιοδιανύσματα και εκτιμώντας την εταιρικότητα όπως είπαμε) αντικαθιστά τις αρχικές τιμές της εταιρικότητας και επαναλαμβάνει τη διαδικασία (βρίσκοντας πάλι τα ιδιοδιανύσματα κλπ) μέχρι να σταματήσουν να υπάρχουν αλλαγές ανάμεσα σε δύο επαναλήψεις. Στην περίπτωση που οι αρχικές εταιρικότητες είναι ίσες με 1 τα αποτελέσματα θα ταυτιστούν με αυτά της μεθόδου κυρίων συνιστωσών.

Παρόμοια λογική χρησιμοποιεί και η μέθοδος *Image Factoring* η οποία επιτρέπει η μοναδικοί παράγοντες να είναι συσχετισμένοι (και άρα ο πίνακας $\mathbf{\Psi}$ να μην είναι διαγώνιος). Στη βιβλιογραφία είναι γνωστό πως και οι δύο δίνουν μη συνεπή αποτελέσματα και για αυτό δεν χρησιμοποιούνται συχνά στην πράξη.

8.6 Περιστροφή

Με την περιστροφή των παραγόντων προσπαθώ να κάνω τους παράγοντες πιο ερμηνεύσιμους. Με την περιστροφή δεν αλλάζουν κάποια από τα χαρακτηριστικά του μοντέλου όπως η καλή του προσαρμοστικότητα και το ποσό της διακύμανσης συνδιακύμανσης που ερμηνεύει το μοντέλο παρά μόνο οι τιμές των επιβαρύνσεων. Γενικά αν **L** είναι ένας πίνακας που περιέχει τις επιβαρύνσεις και **G** ένας ορθογώνιος πίνακας (δηλαδή

ισχύει **G'G=**I) τότε ισχύει πως $\mathbf{LG(LG)' = LGG'L' = LL'}$ κι επομένως και ο πίνακας **LG** μπορεί να θεωρηθεί ως ένας πίνακας επιβαρύνσεων. Μαθηματικά ο πίνακας **G** ορίζει έναν ορθογώνιο μετασχηματισμό.

Κάνοντας λοιπόν την περιστροφή ελπίζουμε ότι οι επιβαρύνσεις κάποιων παραγόντων θα είναι μεγάλες σε απόλυτη κλίμακα μόνο για κάποιες από τις μεταβλητές κι έτσι βλέποντας ποιες μεταβλητές εξαρτώνται με ποιους παράγοντες να μπορέσουμε να δώσουμε μια ερμηνεία σε αυτούς. Οι βασικές μέθοδοι περιστροφής είναι

- Varimax: Προσπαθεί να ελαχιστοποιήσει τον αριθμό των μεταβλητών που έχουν μεγάλες επιβαρύνσεις για κάθε παράγοντα

- Quartimax: Προσπαθεί να ελαχιστοποιήσει τον αριθμό των παραγόντων που εξηγούν μια μεταβλητή

- Equimax: Συνδυασμός των varimax και quartimax

- Oblique: Μη ορθογώνια περιστροφή, οι άξονες που προκύπτουν δεν είναι πια ορθογώνιοι (και άρα οι παράγοντες δεν είναι ανεξάρτητοι). Η ερμηνεία των αποτελεσμάτων είναι πιο δύσκολη. Στην πράξη τον χρησιμοποιούμε όταν δεν θέλουμε οι παράγοντες που προκύπτουν να είναι ασυσχέτιστοι.

8.7 Υπολογισμός των Σκορ των Παραγόντων

Όπως είπαμε προηγουμένως ένας από του σκοπούς της παραγοντικής ανάλυσης είναι να μειώσει τον αριθμό των μεταβλητών. Για να επιτευχθεί αυτό μπορούμε να δημιουργήσουμε καινούριες μεταβλητές, τους παράγοντες, ως γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε ξεκινώντας από έστω 10 αρχικές μεταβλητές να μας μείνουν έστω 4 νέες, οι κοινοί παράγοντες. Κάθε παράγοντας μπορεί να γραφτεί στη μορφή

$$F_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$
$$F_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$
$$...$$
$$F_k = a_{k1}X_1 + a_{k2}X_2 + ... + a_{kp}X_p$$

Οι συντελεστές a_{ij} είναι το σκορ της μεταβλητής X_j στον παράγοντα F_i και δεν πρέπει να συγχέονται με τις επιβαρύνσεις. Όταν το μοντέλο έχει εκτιμηθεί με τη μέθοδο κυριών συνιστωσών οι παράγοντες είναι ακριβής, δηλαδή μπορούν να υπολογιστούν χωρίς σφάλμα. Αντίθετα για μοντέλα εκτιμημένα με τη μέθοδο μεγίστης πιθανοφάνειας προσεγγιστικές μέθοδοι χρησιμοποιούνται. Σημειώστε ότι εξ ορισμού οι νέες μεταβλητές θα έχουν μέση τιμή 0 και θα είναι ασυσχέτιστες, δεδομένου πως το μοντέλο είναι ορθογώνιο.

Με τη χρήση του παραπάνω μοντέλου μπορούμε να δημιουργήσουμε καινούριες μεταβλητές για περαιτέρω χρήση, όπως π.χ. για διακριτική ανάλυση, να δούμε πως κάποιοι υποπληθυσμοί διαφέρουν κλπ.

Έχοντας λοιπόν εκτιμήσει ένα παραγοντικό μοντέλο και έστω **L** και **Ψ** οι εκτιμήσεις μας για τις παραμέτρους αυτού, (πριν η μετά την περιστροφή) τότε μπορούμε να βρούμε τα factor scores δηλαδή τις τιμές των καινούριων μεταβλητών για κάθε μεταβλητή. Οι μέθοδοι που προσφέρονται είναι πολλές. Αυτές που τα περισσότερα στατιστικά πακέτα και ανάμεσα τους το SPSS προσφέρουν είναι οι εξής.

- Regression method. Το διάνυσμα F των καινούριων μεταβλητών για υπολογίζεται ως εξής

$$\mathbf{F} = \left(\mathbf{L}'\mathbf{L}\right)^{-1}\mathbf{L}'\mathbf{X}.$$

Η μέθοδος αυτή βασίζεται στη μέθοδο ελαχίστων τετραγώνων ανάμεσα στις πραγματικές τιμές και αυτές που το παραγοντικό μοντέλο προβλέπει.

- Bartlett method. Σε σχέση με την παραπάνω μέθοδο ο Bartlett πρότεινε αντί να χρησιμοποιήσει κανείς την απλή μέθοδο ελαχίστων τετραγώνων να χρησιμοποιήσει γενικευμένα ελάχιστα τετράγωνα καθώς η διακύμανση δεν είναι η ίδια για όλες τις παρατηρήσεις. Επομένως η μέθοδος εκτίμησης εκτιμά τους παράγοντες ως

$$\mathbf{F} = \left(\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1}\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{X}.$$

- Μέθοδος του Anderson. Η μέθοδος αυτή χρησιμοποιεί τον τύπο

$$\mathbf{F} = \left(\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)\left(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1/2}\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{X}$$

Συνοψίζοντας ο πίνακας 8.2 μας δίνει τον ορισμό του πίνακα **A** με τους συντελεστές των σκορ των παραγόντων για τις διάφορες μεθόδους.

	Factor Score Coefficient
Regression	$\left(\mathbf{L}'\mathbf{L}\right)^{-1}\mathbf{L}'$
Bartlett	$\left(\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1}\mathbf{L}'\mathbf{\Psi}^{-1}$
Anderson	$\left(\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)\left(\mathbf{I} + \mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1/2}\mathbf{L}'\mathbf{\Psi}^{-1}$

Πίνακας 8.2. Ο πίνακας με τους συντελεστές των factor scores, για διάφορες μεθόδους εκτίμησης

Και οι τρεις μέθοδοι δίνουν παράγοντες με μέση τιμή μηδέν (άλλωστε αυτή ήταν και η αρχική υπόθεση). Η μέθοδος του Anderson οδηγεί πάντα σε ασυσχέτιστους

παράγοντες ακόμα και αν εξαιτίας μη ορθογώνιας περιστροφής οι παράγοντες θα έπρεπε να είναι συσχετισμένοι. Η μέθοδος της παλινδρόμησης μπορεί να οδηγήσει σε πίνακα διακύμανσης των παραγόντων ο οποίος δεν είναι ο μοναδιαίος, δηλαδή τα διαγώνια στοιχεία να μην είναι 1 και να υπάρχουν συσχετίσεις.

8.8 Confirmatory Factor Analysis

Όσα περιγράψαμε μέχρι τώρα αφορούν τη χρήση της παραγοντικής ανάλυσης κυρίως ως περιγραφικό εργαλείο. Δηλαδή ψάχνουμε να βρούμε αν υπάρχουν κάποιοι παράγοντες που μπορούν να ερμηνεύσουν τις συσχετίσεις μεταξύ των μεταβλητών των δεδομένων μας και να δώσουμε σε αυτούς κάποια ερμηνεία (αν αυτό βέβαια είναι δυνατόν). Τα τελευταία χρόνια όμως έχει αναπτυχθεί μια διαφορετική προσέγγιση στο θέμα της παραγοντικής ανάλυσης με τον τίτλο Confirrmatory παραγοντική ανάλυση. Θυμηθείτε άλλωστε πως η παραγοντική ανάλυση προσπαθεί να ερμηνεύσει τη δομή και όχι τη μεταβλητότητα.

Σε αυτή την προσέγγιση σκοπός δεν είναι να προκύψει μια περιγραφή των δεδομένων αλλά να ελέγξουμε συγκεκριμένες υποθέσεις όπως αυτές προκύπτουν από συγκεκριμένες θεωρίες στα γνωστικά αντικείμενα των οποίων εφαρμόζονται. Τέτοιες προσεγγίσεις είναι δημοφιλείς στις κοινωνικές επιστήμες αλλά και το Marketing, όπου γίνονται διάφορες θεωρητικές υποθέσεις σχετικά με την ύπαρξη κάποιων κοινών παραγόντων αλλά και μεταξύ των σχέσεων που αυτοί οι παράγοντες έχουν. Στη συνέχεια ο ερευνητής προσπαθεί να ελέγξει στατιστικά τις υποθέσεις αυτές μέσω ενός μοντέλου παραγοντικής ανάλυσης. Επομένως στην περίπτωση της Confirmatory παραγοντικής ανάλυσης ο αριθμός των παραγόντων είναι επιλεγμένος καθώς και οι όποιες συσχετίσεις τους και απλά γίνεται έλεγχος αν αυτό το μοντέλο προσαρμόζει καλά τα δεδομένα μας.

Τα τελευταία χρόνια που η μέθοδος αυτή έχει αναπτυχθεί , έχουν προκύψει μια σειρά από ενδιαφέροντα αποτελέσματα (κατάλληλα στατιστικά μέτρα, αλγόριθμοι κλπ) καθώς και έχουν αναπτυχθεί στατιστικά πακέτα που να υποστηρίζουν αυτές τις προσεγγίσεις (π.χ. AMOS, LISREL κλπ).

8.9 Μη Ορθογώνια Παραγοντική Ανάλυση

Το ορθογώνιο παραγοντικό μοντέλο βασίστηκε στην υπόθεση πως οι παράγοντες είναι ορθογώνιοι μεταξύ τους. Πολλές φορές μια τέτοια υπόθεση δεν είναι καθόλου ρεαλιστική και πρέπει να επιτρέψουμε στους παράγοντες να συσχετίζονται μεταξύ τους. Σε αυτή την περίπτωση δηλαδή υποθέτουμε πως $Cov(\mathbf{F}) = \mathbf{\Omega}$ όπου $\mathbf{\Omega}$ είναι ένας οποιοσδήποτε πίνακας διακύμανσης. Σε αυτή την περίπτωση έχουμε πως

$$\mathbf{\Sigma} = Cov(\mathbf{X}) = Cov(\mathbf{LF} + \boldsymbol{\varepsilon}) =$$

$$\mathbf{L}Cov(\mathbf{F})\mathbf{L}'+Cov(\boldsymbol{\varepsilon}) = \mathbf{L\Omega L}'+\mathbf{\Psi}$$

και επομένως ως προς την εκτίμηση των παραμέτρων έχουμε να εκτιμήσουμε έναν ακόμα μεγαλύτερο αριθμό παραμέτρων, καθώς χρειαζόμαστε και τα στοιχεία του πίνακα **Ω**. Μπορεί βέβαια να παρατηρήσει κανείς πως ο πίνακας **Ω** επειδή είναι πίνακας διακύμανσης μπορεί να γραφτεί στη μορφή **Ω=B'B** όπου **B** ένας κατάλληλος ορθογώνιος πίνακας και επομένως έχουμε πως

$$\mathbf{\Sigma = L\Omega L}'+\mathbf{\Psi = LB}'\mathbf{BL} + \mathbf{\Psi = L}^{*}\mathbf{L}^{*'}+\mathbf{\Psi}$$

και επομένως παρατηρούμε πως καταλήγουμε σε ένα ορθογώνιο μοντέλο.

Στην πράξη αν θέλουμε να εκτιμήσουμε συσχετισμένους παράγοντες, αυτό μπορεί να γίνει χρησιμοποιώντας μια μη ορθογώνια περιστροφή, που όπως είδαμε και προηγουμένως θα οδηγήσει σε παράγοντες με συσχέτιση μεταξύ τους.

Από εκεί και πέρα αν θέλουμε να δώσουμε μια συγκεκριμένη μορφή στον πίνακα **Ω** παρουσιάζονται αρκετά προβλήματα στην εκτίμηση. Για παράδειγμα το βασικό ορθογώνιο παραγοντικό μοντέλο υποθέτει ότι οι παράγοντες είναι ασυσχέτιστοι αλλά και πως έχουν την ίδια διακύμανση (ίση με 1 για όλους τους παράγοντες). Για να προσαρμόσουμε λοιπόν ένα μοντέλο με διαφορετικές διακυμάνσεις σε κάθε παράγοντα τα πράγματα διαφοροποιούνται κάπως, αν και όπως είπαμε πριν αυτό στην πράξη δεν θα γίνει κατά τη διάρκεια της εκτίμησης αλλά κατά τη διάρκεια της περιστροφής.

8.10Συμπεράσματα και Σχόλια

Η παραγοντική ανάλυση από τον ορισμό του μοντέλου της έχει να κάνει με συνεχή δεδομένα, και οι παράγοντες που υποθέτουμε πως υπάρχουν είναι και αυτοί συνεχείς. Στη βιβλιογραφία έχουν αναπτυχθεί μέθοδοι που να χρησιμοποιούν την ιδέα της παραγοντικής ανάλυσης αλλά να μπορούν να δουλέψουν και με άλλου τύπου δεδομένα. Στον πίνακα 8.3 μπορεί κανείς να δει σχετικά μοντέλα. Δεν θα επεκταθούμε περισσότερο σε αυτά.

Μεταβλητές	Παράγοντες	Μέθοδος
Συνεχείς	Συνεχείς	Παραγοντική Ανάλυση
Διακριτές	Συνεχείς	Latent Trait Analysis
Συνεχείς	Διακριτές	Latent Profile Analysis
Διακριτές	Διακριτές	Latent Class Analysis

Πίνακας 8.3. Μέθοδοι σχετικές με την παραγοντική ανάλυση ανάλογα με τη μορφή των δεδομένων

Τελειώνοντας αυτή την περιγραφή της μεθόδου της παραγοντικής ανάλυσης θα πρέπει να σημειώσουμε τα εξής

- Γενικά όταν οι λύσεις με διαφορετικές μεθόδους ή με την ίδια μέθοδο και διαφορετικό αριθμό παραγόντων διαφέρουν πολύ αυτό αποτελεί ισχυρή ένδειξη για ακαταλληλότητα του μοντέλου

- Μπορεί κάποιος να κάνει παραγοντική ανάλυση έχοντας μόνο τον πίνακα διακύμανσης συνδιακύμανσης, και όχι τα πλήρη δεδομένα. Προφανώς σε αυτή την περίπτωση δεν μπορεί να υπολογίσει τις καινούριες μεταβλητές (τους παράγοντες). Αυτό πάντως είναι πολύ χρήσιμο καθώς μας επιτρέπει να κάνουμε παραγοντική ανάλυση με κατηγορικά δεδομένα και χρήση κάποιου αντίστοιχου πίνακας συνδιακύμανσης (π.χ. κάποιον πίνακα συσχετίσεων με συσχετίσεις για κατηγορικά δεδομένα). Σε αυτή την περίπτωση η μέθοδος μεγίστης πιθανοφάνειας δεν πρέπει να χρησιμοποιείται αφού είναι ξεκάθαρο ότι τα δεδομένα δεν είναι κανονικά.

8.11 Εφαρμογή της Μεθόδου

Το παράδειγμα με το οποίο θα δουλέψουμε αφορά 406 αυτοκίνητα και τα χαρακτηριστικά τους συμπεριλαμβανομένων τεχνικών χαρακτηριστικών αλλά και ήπειρο προέλευσης. Τα δεδομένα αυτά έχει χρησιμοποιηθεί από την Αμερικάνικη Στατιστική Ένωση (American Statistical Association) σαν δεδομένα σύγκρισης διαφόρων στατιστικών πακέτων και επομένως μπορείτε να το βρείτε (δυστυχώς όχι απαραίτητα πλήρες) σε πολλά στατιστικά πακέτα. Οι μεταβλητές που θα μας απασχολήσουν εμφανίζονται στον πίνακα 8.4.

Αγγλική ονομασία	Ελληνική μετάφραση	Ονομασία που θα χρησιμοποιούμε
Miles per gallon	Κατανάλωση σε μίλα ανά γαλόνι βενζίνης	Κατανάλωση
Engine Displacement (cub. Inches)	Μέγεθος μηχανής σε κυβικές ίντσες (ίντσα ≈ 2.5 εκατ)	Μέγεθος μηχανής
Horsepower	Ιπποδύναμη	Ιπποδύναμη
Vehicle Weight (lbs)	Βάρος οχήματος σε λίβρες	Βάρος οχήματος
Time to accelerate from 0 to 60 miles per hour	Χρόνος επιτάχυνσης από τα 0 στα 60 μίλια την ώρα	Επιτάχυνση
Model Year (year –1900)	Χρονιά κατασκευής (κρατώντας μόνο τα 2 τελευταία ψηφία)	Χρονιά κατασκευής
Number of cylinders	Αριθμός κυλίνδρων	Αριθμός κυλίνδρων

Πίνακας 8.4. Οι μεταβλητές που θα χρησιμοποιηθούν στην ανάλυση.

Σκοπός λοιπόν της ανάλυσης είναι να δούμε αν και κατά πόσο τα παραπάνω δεδομένα μπορούν να ερμηνευτούν με τη χρήση του ορθογώνιου παραγοντικού μοντέλου και κατά πόσο μπορούμε να ερμηνεύσουμε τους παράγοντες που προκύπτουν.

Το ορθογώνιο παραγοντικό μοντέλο που θα προσαρμόσουμε έχει τη μορφή

$$\text{καταναλωση} = L_{11}F_1 + L_{12}F_2 + ... + L_{1k}F_k + \varepsilon_1$$
$$\text{μεγεθος μηχανης} = L_{21}F_1 + L_{22}F_2 + ... + L_{2k}F_k + \varepsilon_2$$
$$...$$
$$\text{κυλινδροι} = L_{71}F_1 + L_{72}F_2 + ... + L_{7k}F_k + \varepsilon_7$$

και σκοπός μας είναι να εκτιμήσουμε τις επιβαρύνσεις L_{ij} και τις διακυμάνσεις των τυχαίων όρων ε_i .

8.11.1 Καταλληλότητα των δεδομένων

Όπως πάντα πριν ξεκινήσουμε την ανάλυση των δεδομένων πρέπει να δούμε κάποια περιγραφικά στοιχεία για αυτά. Πιθανότατα κάποια γραφήματα να ήταν πιο διαφωτιστικά αλλά για οικονομία χώρου δεν έχουμε περιλάβει γραφήματα. Στον πίνακα 8.5 μπορεί κανείς να δει την μέση τιμή και την τυπική απόκλιση των δεδομένων. Επειδή υπήρχαν 15 αυτοκίνητα χωρίς πλήρη δεδομένα τα εξαιρέσαμε από την ανάλυση. Το ενδιαφέρον στοιχείο που προκύπτει από τον πίνακα 8.5 δεν είναι τόσο οι μέσες τιμές που έτσι κι αλλιώς δεν είναι άμεσα συγκρίσιμες αλλά το γεγονός πως οι διακυμάνσεις διαφέρουν αρκετά και επομένως δεν μοιάζει λογικό να προχωρήσει κανείς σε ανάλυση με τον πίνακα διακύμανσης. Ο πίνακας συσχετίσεων μοιάζει πιο λογικός. Θυμηθείτε πως αυτή η διαφορά έχει σημασία μόνο αν χρησιμοποιηθεί η μέθοδος των κυρίων συνιστωσών αφού η μέθοδος μεγίστης πιθανοφάνειας θα δώσει τα ίδια αποτελέσματα ανεξάρτητα από τον πίνακα που θα χρησιμοποιήσουμε (στην ουσία ανεξάρτητα από το αν τυποποιήσουμε ή όχι τα δεδομένα μας).

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Miles per Gallon	23.48	7.78	391
Engine Displacement (cu. inches)	194.13	104.63	391
Horsepower	104.24	38.28	391
Vehicle Weight (lbs.)	2973.10	845.83	391
Time to Accelerate from 0 to 60 mph (sec)	15.53	2.76	391
Model Year (modulo 100)	75.99	3.68	391
Number of Cylinders	5.47	1.70	391

Πίνακας 8.5. Περιγραφικά Στατιστικά για τις μεταβλητές μας.

Η δράση λοιπόν ξεκινά με τον πίνακα συσχετίσεων που βλέπουμε στον πίνακα 8.6 μαζί με την ορίζουσα του καθώς και κάποια άλλα στατιστικά που μπορείτε να δείτε στους πίνακες 8.7 και 8.8. Κοιτάζοντας λοιπόν τον πίνακα συσχετίσεων (πίνακας 8.6) υπάρχουν ενδείξεις πως οι συσχετίσεις ανάμεσα στις μεταβλητές είναι ικανοποιητικά μεγάλες. Αυτό δεν έχει να κάνει μόνο με το γεγονός πως όλες είναι στατιστικά σημαντικές. Κάτω από κάθε συσχέτιση υπάρχει το p-value για τον έλεγχο της μηδενικής υπόθεσης πως η τιμή του συντελεστή συσχέτισης στον πληθυσμό είναι 0, έναντι της εναλλακτικής πως είναι διάφορη του 0. Βλέπουμε λοιπόν πως σε όλες τις περιπτώσεις είναι στατιστικά σημαντική η συσχέτιση. Αυτό από μόνο του δεν μας αρκεί. Παρατηρούμε πως η μικρότερη συσχέτιση σε απόλυτη τιμή είναι 0.296 ανάμεσα στην επιτάχυνση και τη χρονιά κατασκευής.

Correlation Matrix[a]

	Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)	Model Year (modulo 100)	Number of Cylinders
Miles per Gallon	1.000	-.805	-.776	-.831	.431	.577	-.776
		.000	.000	.000	.000	.000	.000
Engine Displacement	-.805	1.000	.898	.934	-.548	-.367	.951
	.000		.000	.000	.000	.000	.000
Horsepower	-.776	.898	1.000	.863	-.701	-.411	.842
	.000	.000		.000	.000	.000	.000
Vehicle Weight (lbs.)	-.831	.934	.863	1.000	-.425	-.303	.897
	.000	.000	.000		.000	.000	.000
Time to Accelerate	.431	-.548	-.701	-.425	1.000	.296	-.511
	.000	.000	.000	.000		.000	.000
Model Year (modulo 100)	.577	-.367	-.411	-.303	.296	1.000	-.342
	.000	.000	.000	.000	.000		.000
Number of Cylinders	-.776	.951	.842	.897	-.511	-.342	1.000
	.000	.000	.000	.000	.000	.000	

a. Determinant = 1.306E-04

Πίνακας 8.6 Πίνακας συσχετίσεων για τα δεδομένα του παραδείγματος μας

Επίσης ένα άλλο ενδιαφέρον συμπέρασμα είναι πως η μεταβλητή που έχει τις μικρότερες συσχετίσεις με τις υπόλοιπες είναι η χρονιά κατασκευής. Από τον πίνακα 8.7 μπορεί κανείς να δει το στατιστικό Keiser-Meyer-Olkin το οποίο είναι αρκετά υψηλό (0.813) και αυτό υποδεικνύει ότι οι συσχετίσεις ανάμεσα στα δεδομένα μας είναι αρκετά υψηλές. Επίσης, όπως ήταν αναμενόμενο μετά τη συζήτηση, ο έλεγχος σφαιρικότητας του Bartlett απορρίπτει τη μηδενική υπόθεση πως ο πίνακας συσχέτισης είναι ο μοναδιαίος (τιμή της ελεγχοσυνάρτησης 3459.6, βαθμοί ελευθερίας 21, p=7). Η τιμή της ορίζουσας του δειγματικού πίνακα συσχέτισης εμφανίζεται στο κάτω άκρο του πίνακα 8.6 και είναι 0.000131. Παρατηρείστε πως η ορίζουσα σαν τιμή δεν μας λεει και πολλά πράγματα. Αλλά επιτρέπει συγκρίσεις με άλλα σετ δεδομένων.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.813
Bartlett's Test of Sphericity	Approx. Chi-Square	3459.634
	df	21
	Sig.	.000

Πίνακας 8.7 . Το στατιστικό Kaiser-Meyer-Olkin και ο έλεγχος σφαιρικότητας του Bartlett

Anti-image Matrices

	Miles per	Engine Displac	Horsep ower	Vehicle Weight	Time to Acceler	Model Year	Numbe r of
Miles per Gallon	.844[a]	-.057	.007	.460	-.037	-.590	.053
Engine Displacement (cu. inches)	-.057	.837[a]	-.264	-.415	.079	.098	-.674
Horsepower	.007	-.264	.839[a]	-.379	.647	.167	.119
Vehicle Weight (lbs.)	.460	-.415	-.379	.795[a]	-.434	-.392	-.079
Time to Accelerate from 0 to 60 mph (sec)	-.037	.079	.647	-.434	.707[a]	.077	.063
Model Year (modulo 100)	-.590	.098	.167	-.392	.077	.631[a]	-.023
Number of Cylinders	.053	-.674	.119	-.079	.063	-.023	.876[a]

a. Measures of Sampling Adequacy(MSA)

Πίνακας 8.8. Ο πίνακας μερικών συσχετίσεων με αλλαγμένα πρόσημα. Τα διαγώνια στοιχεία είναι MSA των μεταβλητών

Όλα λοιπόν τα στοιχεία δείχνουν πως τα δεδομένα μας είναι κατάλληλα για παραγοντική ανάλυση και πως μπορούμε να προχωρήσουμε. Είναι όμως όλες οι μεταβλητές κατάλληλες να χρησιμοποιηθούν στο μοντέλο; Για να το εξετάσουμε αυτό χρησιμοποιούμε την τιμή MSA την οποία μπορείτε να διαβάσετε για κάθε μεταβλητή στη διαγώνιο του πίνακα 8.8. Ο πίνακας 8.8 ονομάζεται Anti-Image πίνακας και περιέχει στα μη διαγώνια στοιχεία την τιμή του συντελεστή μερικής συσχέτισης των δύο μεταβλητών όταν εξουδετερώσουμε την επίδραση των υπολοίπων με αντίστροφο όμως πρόσημο. Στην καλύτερη περίπτωση για μας ο πίνακας αυτός θα έπρεπε να είναι διαγώνιος. Παρατηρείστε πως για τη μεταβλητή χρονιά κατασκευής η τιμή είναι η μικρότερη (0.631) κάτι που είχαμε παρατηρήσει και από τον πίνακα συσχετίσεων και δηλώνει πως αυτή η μεταβλητή είναι λιγότερο 'σχετική' με τις υπόλοιπες. Παρόλα αυτά όλες οι τιμές κρίνονται ικανοποιητικές και δεν υπάρχει λόγος να διώξουμε κάποια μεταβλητή.

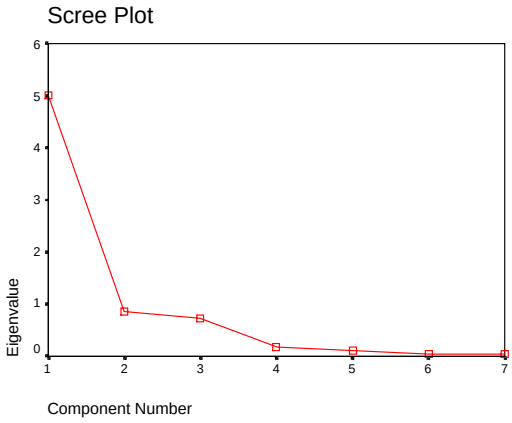
8.11.2 Επιλογή αριθμού παραγόντων

Η επιλογή του αριθμού των παραγόντων δεν μπορεί να γίνει πριν από την εκτίμηση του μοντέλου εκτός αν υπάρχουν ισχυρά στοιχεία ότι ο αριθμός των παραγόντων είναι συγκεκριμένος. Αυτό σημαίνει πως η επιλογή του αριθμού των παραγόντων είναι μια δυναμική διαδικασία και προϋποθέτει επαναληπτικά την εκτίμηση και αξιολόγηση του μοντέλου. Τα περισσότερα κριτήρια που χρησιμοποιούνται για την επιλογή του αριθμού των παραγόντων μοιάζουν με αυτά που χρησιμοποιούνται και στην ανάλυση σε κύριες συνιστώσες και βασίζονται στις ιδιοτιμές του πίνακα συσχετίσεων. Επομένως μπορεί κάποιος να χρησιμοποιήσει τον κανόνα του Kaiser, το ποσοστό της διακύμανσης που εξηγείται ή το scree plot (αναφερόμαστε σε αυτά καθώς αυτά προσφέρονται συνήθως από τα στατιστικά πακέτα) αλλά και όλα τα υπόλοιπα κριτήρια που αναφέραμε στην ανάλυση σε κύριες συνιστώσες. Επιπροσθέτως μπορεί κανείς να χρησιμοποιήσει κριτήρια βασισμένα πάνω στο μοντέλο καθεαυτό όπως τα κατάλοιπα του εκτιμημένου πίνακα συσχετίσεων με τον δειγματικό πίνακα συσχετίσεων ή κριτήρια βασισμένα στην πιθανοφάνεια. Σε αυτές όμως τις περιπτώσεις ο αριθμός των παραγόντων του μοντέλου ανάγει το πρόβλημα σε πρόβλημα επιλογής μοντέλου, όπου κάποιος πρέπει να προσαρμόσει πολλά μοντέλα και να κρατήσει αυτό που θεωρεί καλύτερο με βάση κάποιο κριτήριο. Κάτι τέτοιο δεν χρειαζόταν στην περίπτωση της χρήσης κριτηρίων βασισμένων στις ιδιοτιμές.

Total Variance Explained

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	5.015	71.638	71.638
2	.870	12.434	84.072
3	.721	10.305	94.377
4	.185	2.647	97.023
5	.120	1.710	98.733
6	.054	.773	99.506
7	.035	.494	100.000

Πίνακας 8.9. Οι ιδιοτιμές του πίνακα συσχέτισης των δεδομένων



Γράφημα 8.1. Το Scree Plot για τα δεδομένα μας

Ο πίνακας 8.9 περιέχει τις ιδιοτιμές και το ποσοστό της διακύμανσης που κάθε ιδιοτιμή ερμηνεύει. Έχετε υπόψη σας ότι αυτή η ερμηνεία (ποσοστό διακύμανσης που ερμηνεύει) είναι σωστή μόνο αν χρησιμοποιηθεί η μέθοδος των κυρίων συνιστωσών, καθώς και ότι με τις υπόλοιπες μεθόδους εκτίμησης η διακύμανση που ερμηνεύεται από κάθε παράγοντα διαφέρει. Γίνεται λοιπόν κατανοητό πως το πρόβλημα επιλογής αριθμού παραγόντων δεν είναι άσχετο με την επιλογή μεθόδου εκτίμησης. Παρόλα αυτά κριτήρια βασισμένα στις ιδιοτιμές χρησιμοποιούνται συχνά στην πράξη άσχετα με τη μέθοδο εκτίμησης που διαλέγει κανείς. Με αυτή την προσέγγιση, 2 παράγοντες εξηγούν πάνω από το 80% της διακύμανσης.

Από το scree plot πάλι δεν είναι ξεκάθαρο πόσους παράγοντες θα κρατήσουμε. Θα επανέλθουμε στο πρόβλημα εύρεσης του αριθμού των παραγόντων σε λίγο. Προς το παρόν διαλέγουμε να εκτιμήσουμε 2 παράγοντες με τη χρήση της μεθόδου των κυρίων συνιστωσών και να δουλέψουμε με αυτούς.

8.11.3 Εκτίμηση των παραμέτρων

Μέθοδος Κυρίων Συνιστωσών

Με τη μέθοδο των κυρίων συνιστωσών η εκτιμήτρια του πίνακα **L** είναι ο πίνακας

$$\hat{\mathbf{L}} = \left[\sqrt{\lambda_1}e_1 \quad \sqrt{\lambda_2}e_2\right],$$

όπου λ_i, e_i είναι οι ιδιοτιμές και τα ιδιοδιανύσματα στήλη που αντιστοιχεί σε κάθε μια ιδιοτιμή. Στην περίπτωση μας παίρνουμε μόνο τα δύο πρώτα ζεύγη ιδιοτιμών και ιδιοδιανυσμάτων επειδή διαλέξαμε δύο παράγοντες. Στον πίνακα 8.10 μπορεί κανείς να δει τα στοιχεία του πίνακα $\hat{\mathbf{L}}$, δηλαδή τις επιβαρύνσεις των παραγόντων, που προκύπτουν για το μοντέλο με δύο παράγοντες.

Component Matrix[a]

	Component	
	1	2
Miles per Gallon	-.890	.188
Engine Displacement (cu. inches)	.961	.168
Horsepower	.947	.078
Vehicle Weight (lbs.)	.926	.216
Time to Accelerate from 0 to 60 mph (sec)	-.646	.015
Model Year (modulo 100)	-.510	.848
Number of Cylinders	.931	.187

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Πίνακας 8.10. Πίνακας με τις επιβαρύνσεις των παραγόντων

Από τον πίνακα μπορεί κανείς να δει πως εκφράζεται κάθε μια μεταβλητή με τη χρήση των 2 παραγόντων που χρησιμοποιήσαμε. Έτσι έχουμε πως

$$\text{Κατανάλωση} = -0.890\ F_1 + 0.188\ F_2$$

$$\text{Μέγεθος μηχανής} = 0.961\ F_1 + 0.168\ F_2$$

$$\text{Ιπποδύναμη} = 0.947\ F_1 + 0.078\ F_2$$

$$\text{Βάρος οχήματος} = 0.926\ F_1 + 0.216\ F_2$$

$$\text{Επιτάχυνση} = -0.646\ F_1 + 0.015\ F_2$$

$$\text{Χρονιά κατασκευής} = -0.510\ F_1 + 0.848\ F_2$$

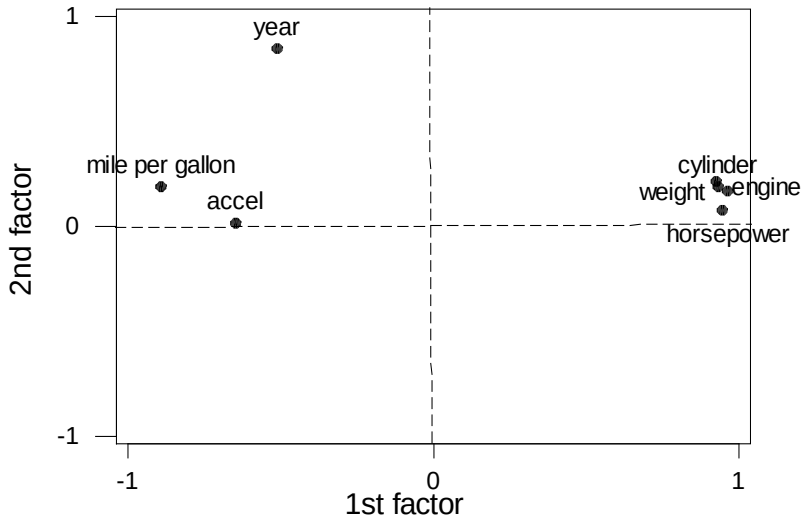
$$\text{Αριθμός κυλίνδρων} = 0.931\ F_1 + 0.187\ F_2$$

Κοιτάζοντας λοιπόν τις σχέσεις των μεταβλητών και των παραγόντων μπορεί κάποιος να δει πως οι μεταβλητές Κατανάλωση, Επιτάχυνση και Χρονιά κατασκευής έχουν αρνητικά πρόσημα ενώ οι υπόλοιπες θετικά πρόσημα για τον πρώτο παράγοντα, επομένως κάποιος θα μπορούσε να διακινδυνεύσει κάποια ερμηνεία για τον πρώτο παράγοντα με βάση αυτή την παρατήρηση. Σας υπενθυμίζουμε πως τα πρόσημα των ιδιοδιανυσμάτων δεν είναι μοναδικά καθώς αν πολλαπλασιάσουμε έναν ιδιοδιάνυσμα με –1 αυτό συνεχίζει να είναι λύση της εξίσωσης από την οποία ορίζονται τα ιδιοδιανύσματα. Επομένως άλλα στατιστικά πακέτα μπορεί να έχουν αντίθετα πρόσημα (π.χ. το MINITAB για τα ίδια δεδομένα δίνει όλο αρνητικά πρόσημα για το δεύτερο παράγοντα)

Πιο εύκολα μπορεί να καταλάβει κανείς πως σχετίζονται οι μεταβλητές αν κάνει ένα γράφημα των παραγόντων ως εξής: σε κάθε άξονα αντιστοιχούμε έναν παράγοντα και για κάθε μεταβλητή την αναπαριστούμε με συντεταγμένη τις επιβαρύνσεις κάθε παράγοντα. Μπορούμε να χρησιμοποιήσουμε τρεις παράγοντες σε τρισδιάστατο γράφημα αλλά συνήθως είναι δύσκολο να χειριστεί κάποιος περισσότερες από δύο διαστάσεις. Για παράδειγμα το γράφημα 8.2 μας δείχνει τους δύο πρώτους παράγοντες. Παρατηρείστε πως οι μεταβλητές βάρος, ιπποδύναμη, μέγεθος μηχανής και αριθμός κυλίνδρων είναι πολύ κοντά η μία στην άλλη και σε αντιδιαστολή με τις υπόλοιπες τρεις.

Πόσο καλό όμως είναι το παραγοντικό μοντέλο που προσαρμόσαμε; Στον πίνακα 8.11 μπορεί κανείς να δει τις εταιρικότητες, δηλαδή τις διακυμάνσεις που εξηγούν οι παράγοντες που προσαρμόσαμε. Αυτό είναι αναγκαστικά ένας αριθμός από 0 έως 1 και είναι το ποσοστό της διακύμανσης κάθε μεταβλητής που εξηγείται από τον αριθμό των παραγόντων που προσαρμόσαμε. Η πρώτη στήλη (initial) είναι 1 αν έχουμε χρησιμοποιήσει τη μέθοδο των κυρίων συνιστωσών. Παρατηρούμε επομένως πως με το μοντέλο που προσαρμόσαμε ερμηνεύουμε το 83% της κατανάλωσης. Το ποσοστό αυτό δεν είναι παρά το άθροισμα τετραγώνων των επιβαρύνσεων των παραγόντων σε αυτή τη μεταβλητή, δηλαδή για τη μεταβλητή κατανάλωση είναι $(-0.89)^2 + (0.188)^2 = 0.828$. Βλέπουμε λοιπόν πως το μοντέλο δεν καταφέρνει να εξηγήσει παρά μόνο το 42% της

μεταβλητής επιτάχυνση και αυτό είναι ίσως μια ένδειξη πως πρέπει να προσθέσουμε και άλλον παράγοντα αν θέλουμε να αυξήσουμε την ερμηνεία για αυτή τη μεταβλητή. Εδώ πρέπει να τονίσουμε πως



Γράφημα 8.2. Οι επιβαρύνσεις των παραγόντων για κάθε μια μεταβλητή

- Ο πίνακας communalities διαφέρει από μέθοδο εκτίμησης σε μέθοδο εκτίμησης. Η διαφορά αφορά κυρίως τη στήλη initial, όπου υπάρχουν μονάδες για τη μέθοδο κυρίων συνιστωσών και ένας αριθμός από το 0 μέχρι το 1 για όλες τις υπόλοιπες μεθόδους.

- Αν είχαμε χρησιμοποιήσει τον πίνακα διακύμανσης, τότε ο πίνακας θα περιείχε κάποιες επιπλέον στήλες, καθώς θα μας παρουσίαζε τόσο τις διακυμάνσεις στα πραγματικά τους μεγέθη όσο και σαν ποσοστά.

- Αν αφαιρέσουμε τη δεύτερη στήλη από τη μονάδα (και όχι από τη στήλη initial που απλά τυγχάνει να είναι 1 επειδή χρησιμοποιήσαμε τη μέθοδο κυρίων συνιστωσών) έχουμε τις εκτιμήσεις των ιδιαιτεροτήτων ψ_i για κάθε μεταβλητή, δηλαδή του κομματιού εκείνου της διακύμανσης κάθε μεταβλητής που δεν μπορεί να εξηγήσει το παραγοντικό μοντέλο. Έτσι από τον πίνακα 8.11 βλέπουμε πως η ιδιαιτερότητα για την μεταβλητή κατανάλωση είναι 0.172 (=1-0.828).

Communalities

	Initial	Extraction
Miles per Gallon	1.000	.828
Engine Displacement (cu. inches)	1.000	.952
Horsepower	1.000	.902
Vehicle Weight (lbs.)	1.000	.905
Time to Accelerate from 0 to 60 mph (sec)	1.000	.418
Model Year (modulo 100)	1.000	.978
Number of Cylinders	1.000	.902

Extraction Method: Principal Component Analysis.

Πίνακας 8.11. Οι εταιρικότητες των μεταβλητών για το παραγοντικό μοντέλο που προσαρμόσαμε

Ένα ακόμα μέτρο του πόσο καλό είναι το μοντέλο που προσαρμόσαμε είναι η σύγκριση του εκτιμώμενου πίνακα διακυμάνσεων ανάμεσα στις αρχικές μεταβλητές. Έχοντας κάνει τη θεμελιώδη για το μοντέλο υπόθεση πως οι συσχετίσεις ανάμεσα στις μεταβλητές οφείλονται αποκλειστικά και μόνο στο ότι αυτές μοιράζονται κάποιους κοινούς παράγοντες, μπορεί κάποιος να δει ότι ο εκτιμημένος πίνακας διακύμανσης των μεταβλητών είναι

$$\hat{\mathbf{S}} = \hat{\mathbf{L}}\hat{\mathbf{L}}'$$

όπου $\hat{\mathbf{L}}$ είναι ο πίνακας επιβαρύνσεων. Προσέξτε πως αυτός είναι πίνακας διακύμανσης και όχι συσχέτισης. Αν έχουμε βασίσει την ανάλυση σε έναν πίνακα συσχέτισης (όπως στην περίπτωση μας) ο εκτιμώμενος πίνακας είναι πίνακας διακύμανσης με την έννοια πως δεν έχει μονάδες στη διαγώνιο. Μπορεί κάποιος να δείξει πως η εκτιμώμενη συνδιακύμανση ανάμεσα στις αρχικές μεταβλητές X_i, X_j είναι

$$\hat{s}_{ij} = \sum_{k=1}^{m} L_{ik} L_{jk}$$

όπου L_{jk} είναι η επιβάρυνση του k παράγοντα στη j μεταβλητή και m είναι το πλήθος των παραγόντων του μοντέλου μας. Ο πίνακας 8.12 περιέχει αυτόν τον εκτιμημένο πίνακα. Στη διαγώνιο υπάρχουν οι εταιρικότητες. Αν το μοντέλο ήταν τέλειο δεν θα έπρεπε να υπάρχουν διαφορές (κατάλοιπα) ανάμεσα στον πραγματικό πίνακα (τον πίνακα συσχετίσεων στην περίπτωση μας) και στον εκτιμημένο πίνακα. Στο κάτω μέρος του πίνακα μπορεί κανείς να δει αυτά τα κατάλοιπα, δηλαδή τη διαφορά του πραγματικού πίνακα μείον τον εκτιμημένο. Δεν υπάρχει σαφές κριτήριο με βάση το οποίο να αποφασίζει κανείς αν οι εκτιμήσεις ήταν καλές. Παρόλα αυτά ανάμεσα σε δύο διαφορετικά μοντέλα μπορεί κανείς να πάρει κάποια συνάρτηση των καταλοίπων (π.χ. άθροισμα τετραγώνων) και να κρίνει ποιο από τα δύο μοντέλα ήταν καλύτερο.

Reproduced Correlations

		Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)	Model Year (modulo 100)	Number of Cylinders
Reproduced Correlation	Miles per Gallon	.828[b]	-.824	-.828	-.784	.578	.613	-.793
	Engine Displacement (cu. inches)	-.824	.952[b]	.923	.927	-.619	-.347	.926
	Horsepower	-.828	.923	.902[b]	.894	-.611	-.417	.896
	Vehicle Weight (lbs.)	-.784	.927	.894	.905[b]	-.595	-.289	.903
	Time to Accelerate from 0 to 60 mph (sec)	.578	-.619	-.611	-.595	.418[b]	.342	-.599
	Model Year (modulo 100)	.613	-.347	-.417	-.289	.342	.978[b]	-.316
	Number of Cylinders	-.793	.926	.896	.903	-.599	-.316	.902[b]
Residual[a]	Miles per Gallon		.0190	.0518	-.0469	-.1473	-.0360	.0172
	Engine Displacement (cu. inches)	.0190		-.0246	.0072	.0705	-.0200	.0246
	Horsepower	.0518	-.0246		-.0308	-.0905	.0057	-.0536
	Vehicle Weight (lbs.)	-.0469	.0072	-.0308		.1700	-.0143	-.0056
	Time to Accelerate from 0 to 60 mph (sec)	-.1473	.0705	-.0905	.1700		-.0462	.0880
	Model Year (modulo 100)	-.0360	-.0200	.0057	-.0143	-.0462		-.0256
	Number of Cylinders	.0172	.0246	-.0536	-.0056	.0880	-.0256	

Extraction Method: Principal Component Analysis.

a. Residuals are computed between observed and reproduced correlations. There are 7 (33.0%) nonredundant residuals with absolute values > 0.05.

b. Reproduced communalities

Πίνακας 8.12. Ο εκτιμώμενος πίνακας διακύμανσης και ο πίνακας καταλοίπων

Δεν θα προχωρήσουμε σε περιστροφή της λύσης που πήραμε ούτε και θα δοκιμάσουμε άλλα μοντέλα. Θα σταματήσουμε εδώ την περιγραφή μας για τη μέθοδο των κυρίων συνιστωσών παρατηρώντας πως

- Αν αποφασίσουμε να προσθέσουμε έναν ακόμα παράγοντα οι επιβαρύνσεις των δύο παραγόντων που είχαμε ήδη δεν θα αλλάξουν. Αυτό ισχύει μόνο στην περίπτωση εκτίμησης με τη μέθοδο των κυρίων συνιστωσών. Αυτό είναι εξαιρετικά χρήσιμο καθώς η προσθήκη νέων παραγόντων είναι σχετικά εύκολη και δεν αλλάζει την όποια ερμηνεία έχουμε ήδη δώσει σε αυτούς τους παράγοντες.

- Δεν μπορούμε να προχωρήσουμε σε ελέγχους σημαντικότητας καθώς η μέθοδος είναι 'μη-παραμετρική' δεν βασίζεται δηλαδή σε καμιά υπόθεση για τον πληθυσμό των δεδομένων μας. Έτσι δεν μπορούμε να κρίνουμε αν και πόσο καλό είναι το παραγοντικό μοντέλο ή ποιο μοντέλο δίνει καλύτερα αποτελέσματα.

Μέθοδος μεγίστης πιθανοφάνειας

Ας δοκιμάσουμε τώρα να χρησιμοποιήσουμε και τη μέθοδο μεγίστης πιθανοφάνειας για να εκτιμήσουμε το παραγοντικό μοντέλο. Η μέθοδος αυτή έχει κάποια πλεονεκτήματα και κάποια μειονεκτήματα ως προς τη μέθοδο των κυρίων συνιστωσών που είδαμε.

Για τα δεδομένα μας διαλέξαμε να πάρουμε 2 παράγοντες, κυρίως για να μπορέσουμε να κάνουμε σύγκριση ανάμεσα στις δύο μεθόδους. Μπορεί κανείς να επαληθεύσει πως για τα δεδομένα μας μπορούμε να εκτιμήσουμε το πολύ 3 παράγοντες με τη μέθοδο μεγίστης πιθανοφάνειας

Ο αλγόριθμος για να ξεκινήσει χρειάζεται αρχικές τιμές για τις ιδιαιτερότητες και τις εταιρικότητες. Οι αρχικές τιμές των ιδιαιτεροτήτων για κάθε μεταβλητή είναι οι συντελεστές προσδιορισμού της παλινδρόμησης κρατώντας ως εξαρτημένη τη μεταβλητή και ανεξάρτητες όλες τις υπόλοιπες. Αυτές τις αρχικές τιμές τις βλέπουμε στη στήλη Initial του πίνακα communalities (πίνακας 8.14). Θυμηθείτε πως στην περίπτωση της εκτίμησης με τη μέθοδο των κυρίων συνιστωσών η αρχική τιμή ήταν 1.

Ο αλγόριθμος εργάζεται επαναληπτικά χρησιμοποιώντας κάθε φορά τις ιδιαιτερότητες ως γνωστές, εκτιμά τις επιβαρύνσεις και στη συνέχεια χρησιμοποιώντας τις επιβαρύνσεις ως γνωστές επανεκτιμά τις ιδιαιτερότητες, αυτή η διαδικασία επαναλαμβάνεται μέχρι οι παράμετροι να πάψουν να αλλάζουν. Συνήθως τα στατιστικά πακέτα δίνουν στο χρήστη τη δυνατότητα να καθορίσει τα κριτήρια τερματισμού του αλγορίθμου.

Factor Matrix[a]

	Factor	
	1	2
Miles per Gallon	-.454	-.697
Engine Displacement (cu. inches)	.574	.802
Horsepower	.721	.608
Vehicle Weight (lbs.)	.453	.844
Time to Accelerate from 0 to 60 mph (sec)	-.999	3.173E-02
Model Year (modulo 100)	-.305	-.249
Number of Cylinders	.536	.789

Extraction Method: Maximum Likelihood.

a. 2 factors extracted. 7 iterations required.

Πίνακας 8.13. Ο πίνακας με τις επιβαρύνσεις των παραγόντων

Ο πίνακας 8.13 είναι ο πίνακας με τις επιβαρύνσεις όταν τις εκτιμήσουμε με τη μέθοδο μεγίστης πιθανοφάνειας. Βλέπουμε πως τώρα μπορούμε να εκφράσουμε τις μεταβλητές μας ως εξής

Κατανάλωση $= -0.454\ F_1 - 0.697\ F_2$

Μέγεθος μηχανής $= 0.574\ F_1 + 0.802\ F_2$

Ιπποδύναμη $= 0.721\ F_1 + 0.608\ F_2$

Βάρος οχήματος $= 0.453\ F_1 + 0.844\ F_2$

Επιτάχυνση $= -0.999\ F_1 + 0.032\ F_2$

Χρονιά κατασκευής $= -0.305\ F_1 - 0.249\ F_2$

Αριθμός κυλίνδρων $= 0.536\ F_1 + 0.789\ F_2$

Συγκρίνοντας τα αποτελέσματα με αυτά που πήραμε με την προηγούμενη μέθοδο εκτίμησης μπορεί κανείς να δει πως υπάρχουν σημαντικές διαφορές οι οποίες δεν είναι μόνο αριθμητικές. Ως προς τον πρώτο παράγοντα τα πρόσημα έχουν παραμείνει ίδια αλλά βέβαια οι επιβαρύνσεις έχουν αλλάξει σε μερικές μεταβλητές και μάλιστα αρκετά. Για παράδειγμα, για την επιτάχυνση έχουμε τώρα –0.999 έναντι –0.324 που είχαμε πριν. Παρατηρείστε πως στο δεύτερο παράγοντα οι επιβαρύνσεις είναι πολύ διαφορετικές και ως προς τα πρόσημα. Δηλαδή έχουμε μια μεγάλη διαφορά στους παράγοντες (επομένως και στην όποια ερμηνεία τους) ανάμεσα στις 2 μεθόδους. Αυτό αποτελεί μια ισχυρή ένδειξη πως το μοντέλο δεν είναι σωστό.

Παρατήρηση: Το γεγονός πως δύο διαφορετικές μέθοδοι εκτίμησης δίνουν αρκετά διαφορετικά αποτελέσματα για τα ίδια δεδομένα και το ίδιο μοντέλο, συνήθως σημαίνει πως το μοντέλο δεν είναι σωστό για τα δεδομένα, καθώς όλες οι μέθοδοι εκτίμησης ξέρουμε πως

δουλεύουν καλά αν το μοντέλο είναι σωστό, και άρα αν το μοντέλο ήταν σωστό και οι δύο μέθοδοι θα το έβρισκαν. Πρέπει να τονιστεί πως αυτό δεν συμβαίνει μόνο στην παραγοντική ανάλυση αλλά και σε άλλες στατιστικές μεθόδους. Για παράδειγμα στη γραμμική παλινδρόμηση κάποιος χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων των καταλοίπων και τη μέθοδο των ελαχίστων απόλυτων αποκλίσεων των καταλοίπων, αν το μοντέλο είναι σωστό (δηλαδή ισχύουν οι υποθέσεις του γραμμικού μοντέλου) θα πάρει παρόμοια αποτελέσματα, αν όμως το μοντέλο δεν είναι σωστό μπορεί να πάρει ριζικά διαφορετικά αποτελέσματα).

Communalities[a]

	Initial	Extraction
Miles per Gallon	.808	.692
Engine Displacement (cu. inches)	.951	.973
Horsepower	.897	.889
Vehicle Weight (lbs.)	.926	.917
Time to Accelerate from 0 to 60 mph (sec)	.632	.999
Model Year (modulo 100)	.474	.155
Number of Cylinders	.906	.910

Extraction Method: Maximum Likelihood.

a. One or more communalitiy estimates greater than 1.0 were encountered during iterations. The resulting solution should be interpreted with caution.

Πίνακας 8.14. Οι εταιρικότητες για το μοντέλο με 2 παράγοντες εκτιμημένο με τη μέθοδο μεγίστης πιθανοφάνειας.

Στον πίνακα 8.14 μπορούμε να δούμε τη διακύμανση κάθε μεταβλητής που καταφέραμε να εξηγήσουμε με 2 παράγοντες. Ο αριθμός προκύπτει όπως και πριν ως το άθροισμα τετραγώνων των επιβαρύνσεων όλων των παραγόντων. Αυτό που είναι ενδιαφέρον να παρατηρήσει κανείς είναι πως με δύο παράγοντες, αλλά με διαφορετικές μεθόδους, παίρνω διαφορετικά αποτελέσματα ως προς το ποσοστό της διακύμανσης κάθε μεταβλητής. Με τη μέθοδο μεγίστης πιθανοφάνειας εξηγώ μόλις το 15% της χρονιάς κατασκευής ενώ με την προηγούμενη μέθοδο είχα εξηγήσει το 98%. Επίσης ενώ πριν εξηγούσα μόλις το 42% της επιτάχυνσης τώρα εξηγώ σχεδόν όλη τη διακύμανση (99.9%).

Επίσης παρατηρείστε και την προειδοποίηση κάτω από τον πίνακα. Επειδή ο αλγόριθμος σε κάθε βήμα εκτιμά τις παραμέτρους μέχρι να συγκλίνει, σε κάποιες επαναλήψεις υπάρχει περίπτωση να φτάσει σε μη επιτρεπτές τιμές. Αυτό το στατιστικό πακέτο μας το πληροφορεί με κατάλληλο μήνυμα. Τέτοια μηνύματα είναι προειδοποιήσεις πως το μοντέλο δεν είναι σωστό. Στη βιβλιογραφία η περίπτωση κατά την επαναληπτική διαδικασία εκτίμησης να βρούμε προσωρινά μια εταιρικότητα ίση με τη μονάδα, αναφέρεται ως Heywood περίπτωση ενώ αν βρούμε τιμή μεγαλύτερη της μονάδας ονομάζεται ultra-Heywood περίπτωση. Στην πραγματικότητα αυτό είναι μια ένδειξη πως το μοντέλο δεν είναι σωστό (ο αριθμός των παραγόντων είναι λάθος, ή οι αρχικές τιμές είναι κακές).

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.015	71.638	71.638	2.639	37.705	37.705	3.983	56.904	56.904
2	.870	12.434	84.072	2.896	41.377	79.082	1.552	22.178	79.082
3	.721	10.305	94.377						
4	.185	2.647	97.023						
5	.120	1.710	98.733						
6	.054	.773	99.506						
7	.035	.494	100.000						

Extraction Method: Maximum Likelihood.

Πίνακας 8.15. Το ποσοστό της διακύμανσης που εξηγούν οι 2 παράγοντες

Ενδιαφέρον παρουσιάζει και ο πίνακας 8.15. Στην αριστερή στήλη μπορεί κανείς να δει τις ιδιοτιμές και το ποσοστό της διακύμανσης που θα εξηγούσε κάθε κύρια συνιστώσα με βάση όσα είπαμε προηγούμενα. Στην πραγματικότητα αυτή η στήλη δεν μας λεει τίποτα για τη μέθοδο απλά μας υπενθυμίζει κάποια αποτελέσματα από την προηγούμενη μέθοδο εκτίμησης (κύριες συνιστώσες). Τώρα στο δεξί μέλος το ποσοστό της διακύμανσης κάθε παράγοντα διαφέρει από αυτό που οι ιδιοτιμές δίνουν. Για τον k παράγοντα, η διακύμανση που εξηγεί αυτός ο παράγοντας είναι το άθροισμα τετραγώνων των επιβαρύνσεων του για όλες τις μεταβλητές δηλαδή $\sum_{i=1}^{p} L_{ik}{}^2$. Παρατηρείστε πως αθροίζω ως προς τη στήλη και όχι ως προς τη γραμμή όπως έκανα για να βρω το ποσοστό της διακύμανσης για κάθε μεταβλητή. Από τον πίνακα βλέπουμε πως οι παράγοντες δεν είναι σε αύξουσα σειρά ως προς την διακύμανση που εξηγούν. Η τρίτη στήλη έχει να κάνει με το ποσοστό της διακύμανσης μετά την περιστροφή που θα δούμε σε λίγο. Παρατηρείστε πως το ποσοστό κάθε παράγοντα έχει αλλάξει μετά την περιστροφή, αν και το συνολικό ποσοστό παρέμεινε το ίδιο. Επίσης πως το μοντέλο με 2 παράγοντες εξηγεί το 79% με τη μέθοδο μεγίστης πιθανοφάνειας και το 84% με τη μέθοδο των κυρίων συνιστωσών.

Μια σημαντική διαφορά με τη μέθοδο κυριών συνιστωσών είναι πως κάθε φορά που προσθέτουμε έναν παράγοντα οι επιβαρύνσεις των προηγούμενων αλλάζουν. Για παράδειγμα στον πίνακα 8.16 έχουμε τις επιβαρύνσεις για τις λύσεις με 1, 2 και 3 παράγοντες. Παρατηρείστε πως οι τιμές αλλάζουν από μοντέλο σε μοντέλο, αλλά παρόλα αυτά τα πρόσημα και ενδεχομένως και η ερμηνεία παραμένουν ίδιες. Δηλαδή για τον πρώτο παράγοντα δείτε ότι τα πρόσημα και στα τρία μοντέλα είναι ίδια (και συμφωνούν με αυτά της μεθόδου κυρίων συνιστωσών). Σε άλλες περιπτώσεις μπορεί να αλλάζει και η ερμηνεία τους και όχι μόνο οι τιμές των επιβαρύνσεων.

	Αριθμός Παραγόντων					
	3			2		1
Μεταβλητή	Παράγοντας			Παράγοντας		Παράγοντας
	1	2	3	1	2	1
Κατανάλωση	-0.856	-0.515	0.033	-0.454	-0.697	-0.823
Μέγεθος μηχανής	0.818	0.235	0.509	0.574	0.802	0.991
Ιπποδύναμη	0.884	0.059	0.319	0.721	0.608	0.907
Βάρος οχήματος	0.762	0.374	0.441	0.453	0.844	0.943
Επιτάχυνση	-0.834	0.550	0.011	-0.999	0.032	-0.553
Χρονιά κατασκευής	-0.516	-0.248	0.243	-0.305	-0.249	-0.378
Αριθμός κυλίνδρων	0.779	0.243	0.497	0.536	0.789	0.956

Πίνακας 8.16. Οι επιβαρύνσεις των παραγόντων για διαφορετικό αριθμό παραγόντων

8.11.4 Αξιολόγηση του μοντέλου

Η χρήση της μεθόδου μεγίστης πιθανοφάνειας μας επιτρέπει τη στατιστική αξιολόγηση του μοντέλου. Συγκεκριμένα μπορούμε να ελέγξουμε τη μηδενική υπόθεση

H_0: $\mathbf{\Sigma} = \mathbf{LL}'+\mathbf{\Psi}$ έναντι της εναλλακτικής

H_1: δεν υπάρχει περιορισμός στον πίνακα $\mathbf{\Sigma}$

Η ελεγχοσυνάρτηση είναι η

$$LR = n(tr\mathbf{D} - \ln|\mathbf{D}| - p)$$

όπου $\mathbf{D} = (\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}})^{-1}\mathbf{S}$. Η μηδενική υπόθεση είναι ισοδύναμη με την υπόθεση ότι το μοντέλο προσαρμόζει καλά τα δεδομένα. Η τιμή του LR συγκρίνεται με την τιμή της χ^2 κατανομής με $s=[(p-k)^2 - (p+k)]/2$ βαθμούς ελευθερίας και αν είναι μεγαλύτερη απορρίπτουμε την καλή προσαρμοστικότητα του μοντέλου. Για τα δεδομένα μας βλέπουμε στον πίνακα 8.17 πως απορρίπτουμε ακόμα και για το μοντέλο με 3 παράγοντες κάτι που σημαίνει πως δεν είναι σωστό το μοντέλο.

Αριθμός παραγόντων k	Τιμή ελεγχοσυνάρτησης	Βαθμοί ελευθερίας	p-value
1	499.618	14	0.000
2	282.514	8	0.000
3	108.140	3	0.000

Πίνακας 8.17. Έλεγχοι καλής προσαρμογής για διάφορα μοντέλα

Στην πράξη μια σωστή πρακτική είναι να προσθέτουμε παράγοντες μέχρι το μοντέλο να γίνει στατιστικά σημαντικό. Προσέξτε πως σε αυτή την περίπτωση περαιτέρω παράγοντες θα οδηγήσουν και πάλι ίσως σε στατιστικά σημαντικό μοντέλο και άρα χρειαζόμαστε κάποιο κριτήριο επιλογής του καλύτερου μοντέλου. Αναφέραμε προηγουμένως κάποια από αυτά.

Μέχρι τώρα έχουμε δει αρκετές φορές πως το μοντέλο με 2 παράγοντες που έχουμε υποθέσει δεν φαίνεται να περιγράφει καλά τα δεδομένα. Στην πραγματικότητα είδαμε πως ούτε το μοντέλο με 3 παράγοντες δεν περιγράφει καλά τα δεδομένα. Η ερώτηση είναι τι μπορεί να φταιει για αυτό. Αφενός ίσως η υπόθεση της κανονικότητας (απαραίτητη για τη μέθοδο μεγίστης πιθανοφάνειας) δεν ισχύει. Στην πράξη στη βιβλιογραφία έχει δειχτεί πώς μικρές αποκλίσεις από την υπόθεση της κανονικότητας δεν μπορούν να θεωρηθούν ως αιτία της απόρριψης του μοντέλου. Αν μάλιστα ληφθεί υπόψη πως οι συσχετίσεις που είχαμε δει στην αρχή της ανάλυσης ήταν μάλλον υψηλές, η απόρριψη του ορθογώνιου παραγοντικού μοντέλου μπορεί να οφείλεται σε μια σειρά από λόγους όπως

- Μη γραμμικότητα της σχέσης παραγόντων και μεταβλητών
- Ύπαρξη συσχέτισης μεταξύ των παραγόντων (δηλαδή το μοντέλο δεν είναι ορθογώνιο)
- Η βασική υπόθεση του μοντέλου είναι πως η συσχέτιση ανάμεσα σε δύο μεταβλητές οφείλεται στην ύπαρξη κάποιων κοινών παραγόντων, επομένως αν κάτι τέτοιο δεν ισχύει μπορεί να οδηγήσει παρά την ύπαρξη σημαντικών συσχετίσεων σε απόρριψη του μοντέλου.

Παρά την απόρριψη του μοντέλου θα συνεχίσουμε να εργαζόμαστε με αυτά τα δεδομένα για να δείξουμε κάποια άλλα στοιχεία του παραγοντικού μοντέλου. Σε αυτό συνηγορεί η περιγραφική φύση του παραγοντικού μοντέλου καθώς συνήθως απλά θέλουμε να δούμε τη δομή που υπάρχει στα δεδομένα μας και συνεπώς η στατιστική συμπερασματολογία περνάει σε δεύτερη μοίρα.

Επίσης έχει δειχτεί στη βιβλιογραφία πως η εκτίμηση με τη μέθοδο μεγίστης πιθανοφάνειας είναι ισοδύναμη με τη μεγιστοποίηση της ορίζουσας του πίνακα των μερικών συσχετίσεων και επομένως η υπόθεση της κανονικότητας χρειάζεται περισσότερο για σκοπούς στατιστικής συμπερασματολογίας παρά για να ισχύουν τα αποτελέσματα της. Θα συνεχίσουμε λοιπόν με την περιστροφή.

8.11.5 Περιστροφή

Η περιστροφή είναι ανεξάρτητη της μεθόδου εκτίμησης και σκοπό έχει να αυξήσει την ερμηνευτική ικανότητα του μοντέλου. Βασίζεται στο αποτέλεσμα πως αν **L** είναι ένας πίνακας που περιέχει τις επιβαρύνσεις και **G** ένας ορθογώνιος πίνακας τότε ισχύει πως $\mathbf{LG(LG)'} = \mathbf{LGG'L'} = \mathbf{LL'}$ κι επομένως και ο πίνακας **LG** είναι μια λύση, δηλαδή μια εκτίμηση του πίνακα των επιβαρύνσεων. Μαθηματικά ο πίνακας **G** ορίζει έναν ορθογώνιο μετασχηματισμό. Αν δεν χρησιμοποιήσουμε την ιδιότητα της ορθογωνιότητα του πίνακα **G** τότε μπορούμε να βρούμε μη ορθογώνιους μετασχηματισμούς, οι οποίοι οδηγούν σε μη ορθογώνιους άξονες. Στατιστικά αυτό σημαίνει πως οι παράγοντες δεν είναι πια ασυσχέτιστοι.

Για την περίπτωση μας θα χρησιμοποιήσουμε ορθογώνιους μετασχηματισμούς και συγκεκριμένα τις μεθόδους

•Varimax: που προσπαθεί να ελαχιστοποιήσει τον αριθμό των μεταβλητών που έχουν μεγάλες επιβαρύνσεις για κάθε παράγοντα

•Quartimax: που προσπαθεί να ελαχιστοποιήσει τον αριθμό των παραγόντων που εξηγούν μια μεταβλητή

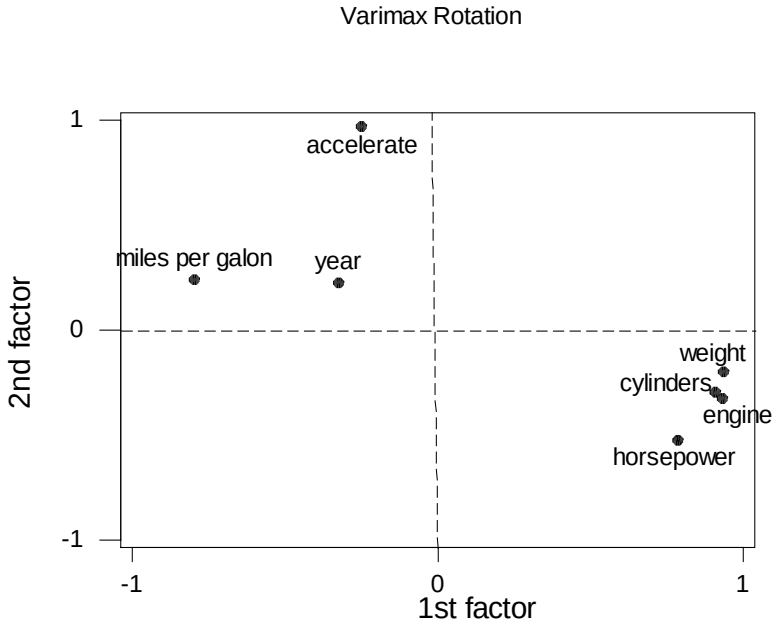
Αν κοιτάξουμε τον πίνακα 8.13 που περιέχει τις επιβαρύνσεις με τη μέθοδο μεγίστης πιθανοφάνειας, η varimax μέθοδος προσπαθεί να κάνει τα στοιχεία κάθε στήλης του πίνακα μικρά ενώ η quartimax μέθοδος προσπαθεί να κάνει τα στοιχεία κάθε γραμμής όσο γίνεται πιο μικρά. Η επιλογή της μιας ή της άλλης εξαρτάται από τους σκοπούς της ανάλυσης. Αν σκοπός μας είναι να δούμε ποιες μεταβλητές επιδρούν σε κάθε παράγοντα για να αναγνωρίσουμε τον παράγοντα η varimax περιστροφή είναι προτιμότερη. Αν σκοπός μας είναι να δούμε απλά σε ποιο παράγοντα είναι κάθε μεταβλητή πιο σημαντική, κι ενδεχομένως να δούμε οι μεταβλητές πως τείνουν να είναι μαζί στους παράγοντες, τότε προτιμότερη είναι η περιστροφή quartimax.

Ο πίνακας Rotated Factor Matrix περιέχει τις επιβαρύνσεις των παραγόντων μετά την περιστροφή. Στον πίνακα 8.18 μπορεί κανείς να δει τις επιβαρύνσεις και με τις δύο μεθόδους

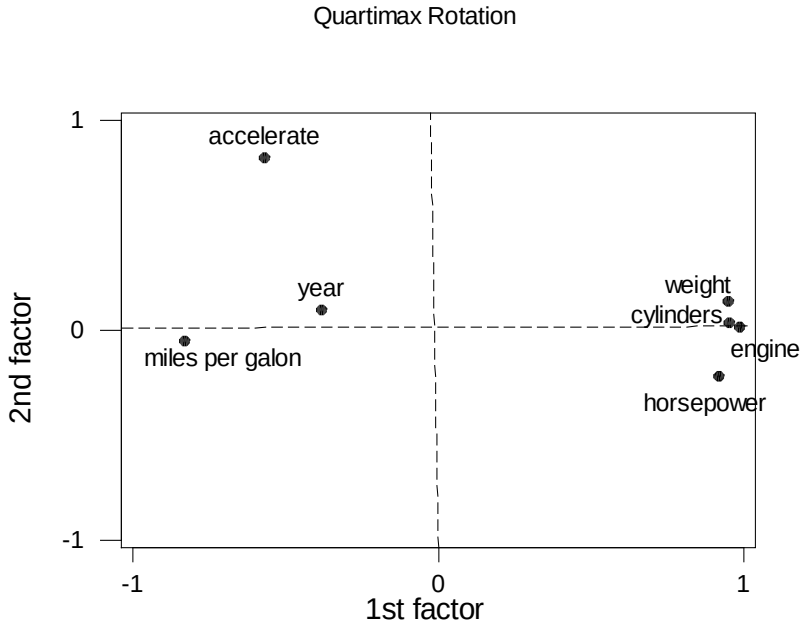
	Varimax		**Quartimax**	
	Παράγοντας		Παράγοντας	
	1	2	1	2
Κατανάλωση	-0.796	0.241	-0.830	-0.049
Μέγεθος μηχανής	0.931	-0.327	0.986	0.014
Ιπποδύναμη	0.785	-0.523	0.917	-0.220
Βάρος οχήματος	0.937	-0.199	0.948	0.137
Επιτάχυνση	-0.248	0.968	-0.568	0.823
Χρονιά κατασκευής	-0.324	0.223	-0.381	0.097
Αριθμός κυλίνδρων	0.907	-0.295	0.953	0.037

Πίνακας 8.18. Οι επιβαρύνσεις των παραγόντων μετά την περιστροφή

Για να δει κανείς τη διαφορά των δύο μεθόδων μπορεί να κοιτάξει τα γραφήματα 8.3α και 8.3β. Η quartimax μέθοδος προσπαθεί να 'μαζέψει' τις μεταβλητές κοντά σε μια από τις δύο γραμμές που δείχνουν τους παράγοντες, αυτό σημαίνει πως για αυτό τον παράγοντα η μεταβλητή έχει μικρή επιβάρυνση και άρα δεν είναι σημαντική. Από την άλλη η μέθοδος varimax προσπαθεί να απομακρύνει όσο γίνεται τις μεταβλητές ώστε σε κάθε παράγοντα μόνο λίγες μεταβλητές να έχουν μεγάλες επιβαρύνσεις. Παρατηρείστε πάντως πως η βασική ιδέα για το πως ομαδοποιούνται ή διαφοροποιούνται οι μεταβλητές παραμένει η ίδια. Το ίδιο και οι αποστάσεις των σημείων μεταξύ τους.



Γράφημα 8.3α. Οι επιβαρύνσεις μετά την περιστροφή (varimax rotation)



Γράφημα 8.3β. Οι επιβαρύνσεις μετά την περιστροφή (quartimax rotation)

Factor Transformation Matrix

Factor	1	2
1	.279	-.960
2	.960	.279

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

Πίνακας 8.19. Ο πίνακας μετασχηματισμού που χρησιμοποιήθηκε για τη varimax περιστροφή

Ο πίνακας Factor Transfomation Matrix (πίνακας 8.19) είναι ο πίνακας με τον οποίο πολλαπλασιάσαμε τον αρχικό πίνακα επιβαρύνσεων για να οδηγηθούμε στον τελικό πίνακα επιβαρύνσεων. Δηλαδή προκύπτει πως

$$\text{Rotated Factor Matrix} = \text{Factor Matrix} \times \text{Factor Transformation Matrix}$$

Και επομένως έχουμε (για τη varimax περιστροφή)

$$\begin{bmatrix} -0.454 & -0.697 \\ 0.574 & 0.802 \\ 0.721 & 0.608 \\ 0.453 & 0.844 \\ -0.999 & 0.032 \\ -0.305 & -0.249 \\ 0.536 & 0.789 \end{bmatrix} = \begin{bmatrix} -0.796 & 0.241 \\ 0.931 & -0.327 \\ 0.785 & -0.523 \\ 0.937 & -0.199 \\ -0.248 & 0.968 \\ -0.324 & -0.223 \\ 0.907 & -0.295 \end{bmatrix} \times \begin{bmatrix} 0.279 & -0.96 \\ 0.96 & 0.279 \end{bmatrix}$$

Επίσης από τον πίνακα 8.15 παρατηρείστε πως πια αλλάζει και το ποσοστό της διακύμανσης που κάθε παράγοντας εξηγεί.

Όπως είπαμε και στην αρχή η περιστροφή έχει σκοπό να μας βοηθήσει να 'δουμε' καλύτερα τι σημαίνει κάθε παράγοντας αφού ελπίζουμε πως θα μας ξεχωρίσει καλύτερα τις μεταβλητές.

Είναι πολύ ενδιαφέρον να παρατηρήσουμε πως στην περίπτωση των δύο παραγόντων, ο πίνακας G που ορίζει έναν ορθογώνιο μετασχηματισμό παίρνει τη μορφή

$$\mathbf{G} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

όπου θ είναι η γωνία περιστροφής των αξόνων. Έτσι στην περίπτωση της varimax περιστροφής που είδαμε προηγουμένως βρίσκουμε πως $\cos(73.74^{\circ}) = 0.279$ και άρα περιστρέψαμε την αρχική λύση κατά 73.74 μοίρες.

8.11.6 Δημιουργία των factor scores

Όπως είπαμε στην αρχή ένας από τους σκοπούς της παραγοντικής ανάλυσης είναι η δημιουργία καινούριων μεταβλητών οι οποίες να συμπυκνώνουν όσο γίνεται τη διακύμανση των αρχικών μεταβλητών, ελπίζοντας πως έτσι μπορούμε να μειώσουμε τις διαστάσεις του προβλήματος. Δηλαδή θέλουμε για κάθε παρατήρηση να δημιουργήσουμε καινούριες μεταβλητές, τόσες όσοι και οι παράγοντες στο μοντέλο που χρησιμοποιήσαμε, ώστε να δημιουργήσουμε καινούρια δεδομένα για κάθε παρατήρηση με σκοπό περαιτέρω στατιστική επεξεργασία. Οι καινούριες μεταβλητές όταν έχουμε χρησιμοποιήσει ένα ορθογώνιο παραγοντικό μοντέλο θα είναι ασυσχέτιστες (θυμηθείτε πως αυτή είναι μια από τις υποθέσεις του παραγοντικού μοντέλου). Προσοχή χρειάζεται όταν ναι μεν ξεκινήσουμε από ένα ορθογώνιο παραγοντικό μοντέλο αλλά στη συνέχεια χρησιμοποιήσουμε μη ορθογώνια περιστροφή και επομένως οι παράγοντες που θα προκύψουν μετά την περιστροφή είναι συσχετισμένοι.

Σκοπός δηλαδή είναι να εκφράσουμε κάθε παράγοντα ως γραμμικό συνδυασμό των αρχικών μας μεταβλητών, και συγκεκριμένα

$$F_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$
$$F_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$
$$\ldots$$
$$F_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kp}X_p$$

ή υπό μορφή πινάκων **F=AX,** όπου ο πίνακας **A** ονομάζεται factor score coefficient matrix και διαφέρει από τον πίνακα των επιβαρύνσεων. Για να βρούμε τον πίνακα αυτό μπορούμε να χρησιμοποιήσουμε διάφορες μεθόδους όπως περιγράφτηκε.

Για το μοντέλο που χρησιμοποιήσαμε, δηλαδή δύο παράγοντες εκτιμημένους με τη μέθοδο μεγίστης πιθανοφάνειας ο πίνακας 8.20 περιέχει τον πίνακα με τους συντελεστές των σκορ (factor score coefficient matrix). Στην πραγματικότητα είναι ο **A**' και όχι ο **A**. Από αυτόν μπορούμε να διαβάσουμε πως ο πρώτος παράγοντας μπορεί να αναπαρασταθεί ως

F_1 = -0.047 Κατανάλωση + 0.624 Μέγεθος Μηχανής + 0.115 Ιπποδύναμη +
0.212 Βάρος + 0.380 Επιτάχυνση – 0.006 Χρονιά κατασκευής +
+ 0.182 Κύλινδροι

Factor Score Coefficient Matrix

	Factor	
	1	2
Miles per Gallon	-.047	-.012
Engine Displacement (cu. inches)	.624	.160
Horsepower	.115	.027
Vehicle Weight (lbs.)	.212	.056
Time to Accelerate from 0 to 60 mph (sec)	.380	1.128
Model Year (modulo 100)	-.006	-.001
Number of Cylinders	.182	.047

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

Factor Scores Method: Regression.

Πίνακας 8.20. Ο πίνακας των συντελεστών των factor scores.

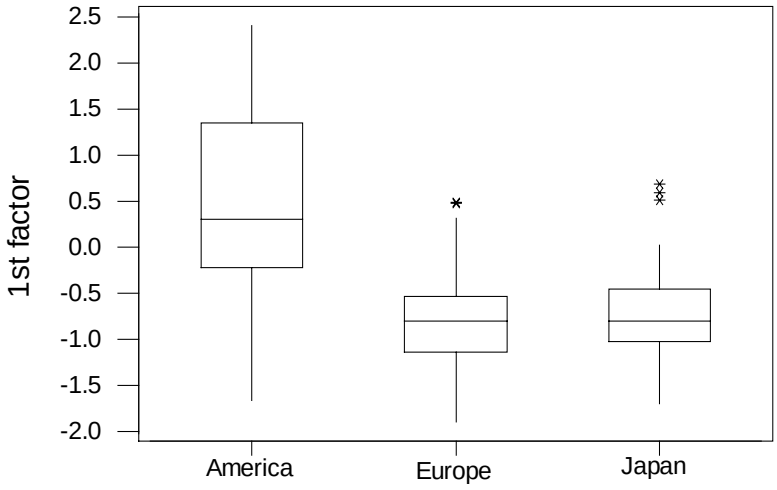
Αξίζει να παρατηρήσουμε πως στην περίπτωση που χρησιμοποιούμε τη μέθοδο των κυρίων συνιστωσών ο πίνακας factor score coefficient μπορεί να προκύψει απλά ως ο πίνακας των συντελεστών από την απλή ανάλυση σε κύριες συνιστώσες. Επειδή όμως θα πρέπει εξ' ορισμού όλοι οι παράγοντες έχουν διακύμανση 1 και όχι όπως οι κύριες συνιστώσες διακυμάνσεις που έχουν διακύμανση ίση με την ιδιοτιμή και είναι σε φθίνουσα τάξη μεγέθους, διαιρούμε με την τετραγωνική ρίζα της ιδιοτιμής, για να κάνουμε όλους τους παράγοντες να έχουν ίδια διακύμανση. Αυτή την τεχνική χρησιμοποιούν κάποια πακέτα όπως το MINITAB. Από την άλλη το SPSS ακόμα και όταν χρησιμοποιήσουμε τη μέθοδο των κυρίων συνιστωσών για την εκτίμηση, αυτό χρησιμοποιεί τη μέθοδο της παλινδρόμησης αν και οι τρεις μέθοδοι που προαναφέραμε δίνουν τα ίδια αποτελέσματα.

8.11.7 Χρήση των σκορ

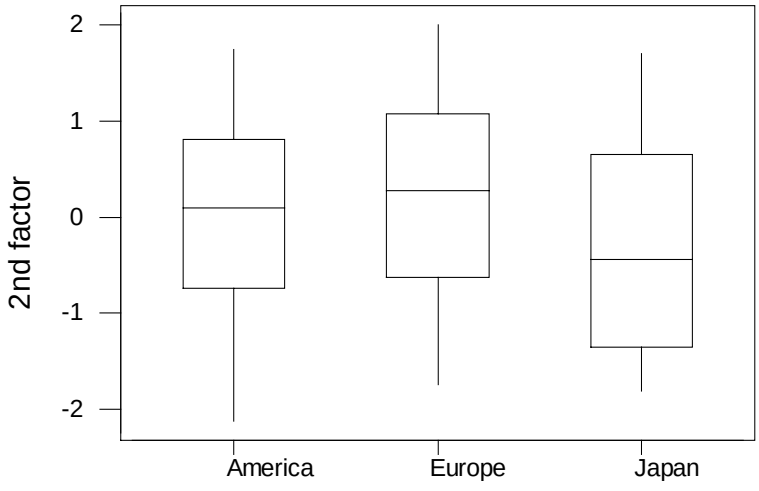
Τα σκορ λοιπόν που έχουμε πια αποθηκεύσει σε καινούριες μεταβλητές μπορούν να χρησιμοποιηθούν για να συνεχίσει κανείς την ανάλυση. Στην ουσία έχουμε πια ποσοτικοποιήσει τους παράγοντες που υποθέσαμε ότι εξηγούν τις συσχετίσεις των μεταβλητών μας. Για τα δεδομένα μας αποθηκεύσαμε τους δύο παράγοντες των παραγόντων για όλες τις παρατηρήσεις μπορεί κανείς να τις δει στα γραφήματα που ακολουθούν. Από το αρχείο δεδομένων γνωρίζουμε για κάθε αυτοκίνητο την ήπειρο προέλευσης τους. Τα αυτοκίνητα είναι ταξινομημένα σε Αμερικάνικα, Ευρωπαϊκά και Γιαπωνέζικα. Στα γραφήματα 8.4α, 8.4β και 8.5 βλέπουμε Boxplots για τις τρεις ομάδες αυτοκινήτων και τις τιμές τους στους δύο παράγοντες. Είναι ξεκάθαρη η διαφορά στον πρώτο παράγοντα όπου τα αυτοκίνητα από την Αμερική έχουν αρκετά μεγαλύτερες τιμές.

Πρέπει να αναφερθεί πως αν κάποιος δει τα περιγραφικά στατιστικά για αυτές τις καινούριες μεταβλητές εύκολα διαπιστώνει πως έχουν μέση τιμή 0 και διακύμανση 1 (εκτός

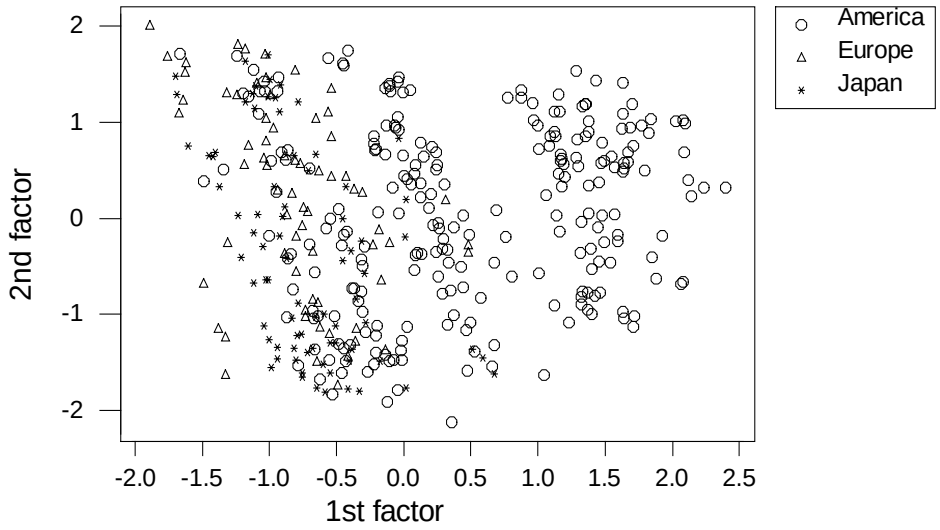
ίσως από πολύ μικρές αριθμητικές αποκλίσεις). Το αν θα είναι όμως ασυσχέτιστες ή όχι εξαρτάται από τη μέθοδο υπολογισμού που χρησιμοποιήσαμε αλλά και την περιστροφή που επιλέξαμε.



Γράφημα 8.4α. Boxplot για τις τιμές στον πρώτο παράγοντα με βάση την ήπειρο προέλευσης



Γράφημα 8.4β. Boxplot για τις τιμές στο δεύτερο παράγοντα με βάση την ήπειρο προέλευσης



Γράφημα 8.5. Διάγραμμα σημείων για τι τιμές όλων των αυτοκινήτων στους δύο πρώτους παράγοντες. Παρατηρείστε πως τα περισσότερα αμερικάνικα αυτοκίνητα ξεχωρίζουν στη δεξιά μεριά του γραφήματος

Παρατηρείστε στο γράφημα 8.5 το οποίο παρουσιάζει έναν νέφος σημείων όλων των παρατηρήσεων πως τα αμερικάνικα αυτοκίνητα μαζεύονται στο δεξί μέλος της εικόνας. Δηλαδή αυτό που βλέπουμε είναι πως χρησιμοποιώντας την παραγοντική ανάλυση η πληροφορία που συμπυκνώνουν οι 2 πρώτοι παράγοντες είναι αρκετή για να αποκτήσουμε την ικανότητα να ξεχωρίζουμε τα αυτοκίνητα. Θυμηθείτε πως μια ερμηνεία που είχαμε δώσει στον πρώτο παράγοντα ήταν ένα κοντράστ ανάμεσα στο μέγεθος του αυτοκινήτου και τις επιδόσεις του. Τα αμερικάνικα αυτοκίνητα, όπως είναι γνωστό άλλωστε στους φίλους του αυτοκινήτου, είναι συνήθως ογκώδη και για αυτό ξεχωρίζουν από τα υπόλοιπα. Το παραπάνω αποτελεί μια απλή εφαρμογή των καινούριων δεδομένων που προέκυψαν από την παραγοντική ανάλυση.

Όπως είπαμε και προηγουμένως το παραγοντικό μοντέλο απέτυχε στο συγκεκριμένο σετ δεδομένων. Χρησιμοποιώντας το απλά για περιγραφικούς σκοπούς κάποιος μπορεί να αποκτήσει μια εικόνα σχετικά με τον τρόπο που συσχετίζονται οι μεταβλητές ακόμα και αν το μοντέλο δεν είναι καλό. Έχετε υπόψη σας πως η αξιολόγηση του μοντέλου ως μη καλό έγινε με τη χρήση στατιστικής συμπερασματολογίας επειδή χρησιμοποιήσαμε τη μέθοδο μεγίστης πιθανοφάνειας. Άλλες μέθοδοι δεν επιτρέπουν τέτοιου είδους αξιολόγηση. Για αυτό πολλές φορές το παραγοντικό μοντέλο χρησιμοποιείται απλά για περιγραφικούς σκοπούς αγνοώντας κάθε μορφή ελέγχου της καταλληλότητας του μοντέλου.

Τελειώνοντας θα πρέπει να τονίσουμε πως οι αρχικές υποθέσεις του παραγοντικού μοντέλου δεν μπορούν να ελεγχθούν από τα δεδομένα με σαφείς τρόπους, εκτός ίσως από την υπόθεση της πολυμεταβλητής κανονικότητας για την οποία έχουμε συζητήσει τη

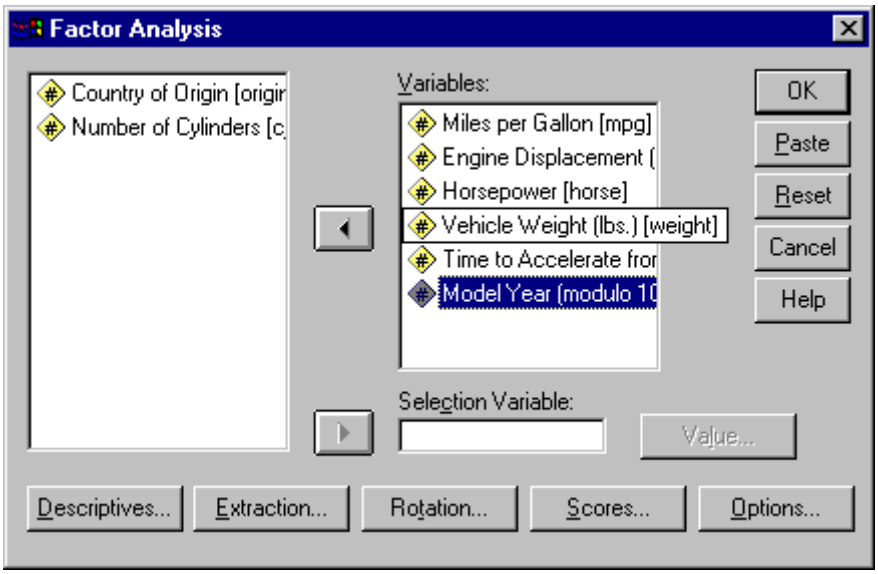
δυσκολία που παρουσιάζει ως προς τη διάγνωση της. Οι υποθέσεις αυτές λοιπόν ήταν μάλλον τεχνικές με σκοπό να επιτρέψουν το χτίσιμο του μοντέλου και την εκτίμηση των παραμέτρων του. Παρόλα αυτά απόρριψη του μοντέλου σημαίνει συνήθως πως οι υποθέσεις αυτές δεν φαίνονται ρεαλιστικές για τα δεδομένα μας.

8.12Παραγοντική Ανάλυση με τη Χρήση του Στατιστικού Πακέτου SPSS for Windows

Στο Στατιστικό πακέτο SPSS for Windows version 9.0 μπορεί κάποιος να κάνει παραγοντική ανάλυση διαλέγοντας διαδοχικά

Analyse > Data Reduction > Factor Analysis

Τότε εμφανίζεται το παράθυρο που φαίνεται στην εικόνα 8.1



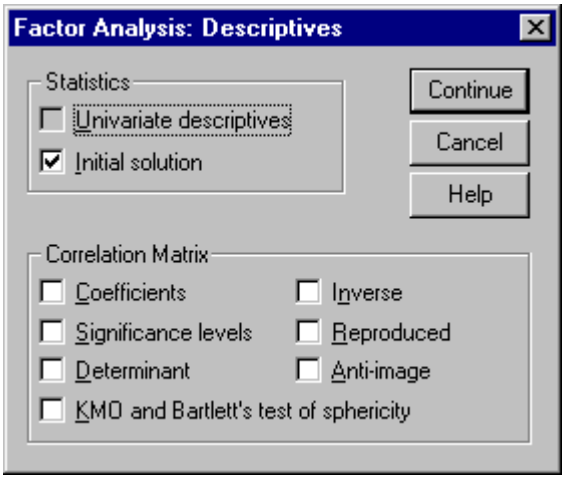
Εικόνα 8.1. Το αρχικό παράθυρο της παραγοντικής ανάλυσης

Από το παράθυρο αυτό μπορούμε να διαλέξουμε τις μεταβλητές που είναι αριστερά και μεταφέροντας αυτές στο διπλανό παράθυρο τις επιλέγουμε να χρησιμοποιηθούν για την ανάλυση. Στο παράδειγμα μας έχουμε διαλέξει τις μεταβλητές mpg, engine, horse, weight, accel και year. Οι επιλογές που μας προσφέρονται αντιστοιχούν στα πλήκτρα στο κάτω μέρος του παραθύρου καθένα από τα οποία ενεργοποιεί διάφορες επιλογές σχετικά με το μοντέλο. Διαλέγοντας το πλήκτρο ***Descriptives*** εμφανίζεται το παράθυρο της εικόνας 8.2.

Οι επιλογές αφορούν διάφορα περιγραφικά μέτρα που είναι χρήσιμα για την ανάλυση μας. Μπορούμε να επιλέξουμε όποια από αυτά θέλουμε. Συγκεκριμένα οι επιλογές μας δίνουν

Επιλογή	Αποτέλεσμα
Statistics	
Univariate Statistics	περιγραφικά στατιστικά για κάθε μεταβλητή
Initial Solution	Η αρχική λύση. Αν διαλέξουμε τη μέθοδο κυρίων συνιστωσών παίρνουμε μια στήλη με μονάδες. Αυτό δεν ισχύει στην περίπτωση της μεθόδου μεγίστης πιθανοφάνειας όπου οι αρχικές τιμές είναι ο συντελεστής παλινδρόμησης της κάθε μεταβλητής με επεξηγηματικές μεταβλητές όλες τις υπόλοιπες
Correlation Matrix	
Coefficients	ο πίνακας συσχετίσεων
Significance Levels	ο πίνακας με τη στατιστική σημαντικότητα κάθε συσχέτισης ξεχωριστά.
Determinant	την ορίζουσα του πίνακα συσχετίσεων. Τιμές κοντά στο 0 σημαίνουν την ύπαρξη συσχετίσεων
KMO and Bartlett's test of sphericity	ο έλεγχος σφαιρικότητας του Bartlett και η Kaiser-Meyer-Olkin στατιστική συνάρτηση για την καταλληλότητα των δεδομένων
Inverse	Ο αντίστροφος του πίνακα συσχετίσεων
Reproduced	ο εκτιμημένος πίνακας συσχετίσεων σύμφωνα με το μοντέλο. Τα διαγώνια στοιχεία είναι οι εταιρικότητες (communalities) ενώ τα στοιχεία κάτω από τη διαγώνιο είναι η διαφορά της εκτιμημένης συσχέτισης με την πραγματική
Anti-image	Περιέχει τις αρνητικές τιμές του πίνακα μερικών συσχετίσεων, ενώ τα διαγώνια στοιχεία του πίνακα είναι τα MSA των μεταβλητών

Με το πλήκτρο ***Extraction*** διαλέγουμε τη μέθοδο εκτίμησης των παραγόντων από την επιλογή ***Method*** . Το πακέτο μας προσφέρει διάφορες μεθόδους μεταξύ αυτών και τη μέθοδο κυρίων συνιστωσών (που είναι η μέθοδος που χρησιμοποιεί αν δεν επιλέξουμε εμείς κάποια) και τη μέθοδο μεγίστης πιθανοφάνειας. Μπορούμε να επιλέξουμε επίσης αν θα δουλέψουμε με τον πίνακας συσχετίσεων η τον πίνακα διακύμανσης συνδιακύμανσης.

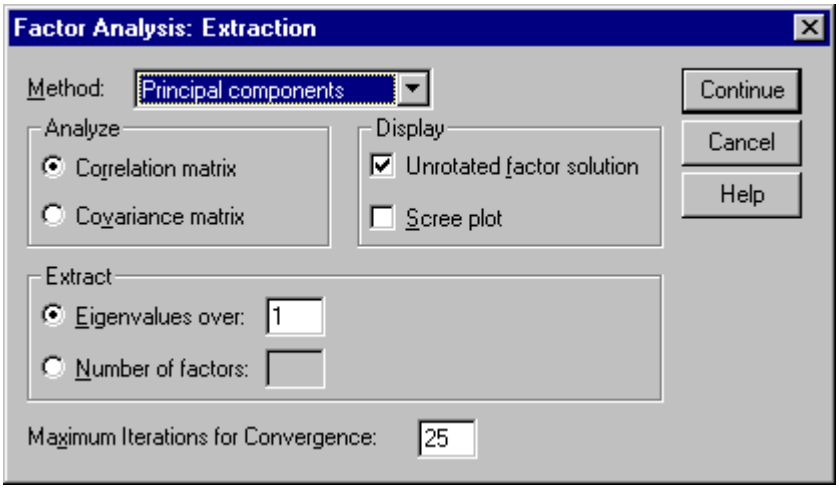


Εικόνα 8.2. Οι επιλογές από το παράθυρο ***Descriptives***

.

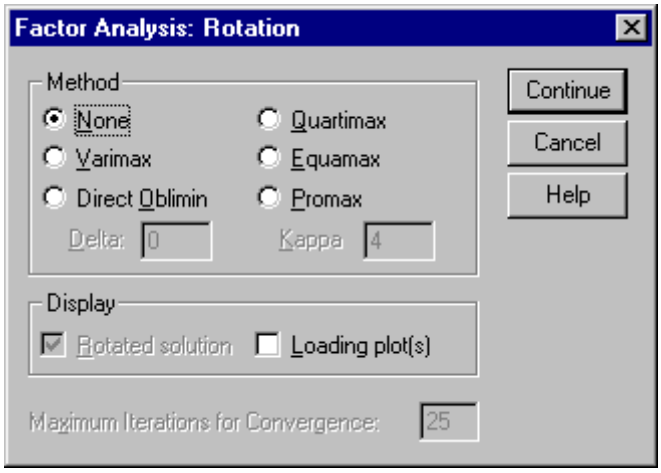
Με τη μέθοδο μεγίστης πιθανοφάνειας είναι αδιάφορο ποια μέθοδο θα επιλέξουμε και η επιλογή δεν ενεργοποιείται . Μπορούμε να διαλέξουμε επίσης να μας εμφανιστεί η λύση πριν την περιστροφή καθώς και το scree plot. Παρατηρείστε ότι καθώς για να τρέξουμε μια παραγοντική ανάλυση πρέπει να ξέρουμε τον αριθμό των παραγόντων, πολλές φορές χρειάζεται να τρέξουμε μια ανάλυση ώστε να βρούμε τα στοιχεία που θα μας κάνει να βρούμε την πληροφορία σχετικά με τον αριθμό των παραγόντων. Για να επιλέξουμε πόσους παράγοντες μπορούμε είτε να διαλέξουμε απευθείας τον αριθμό είτε να διαλέξουμε μια τιμή και έτσι να πάρουμε τόσους παράγοντες όσες και οι ιδιοτιμές το πίνακα που χρησιμοποιούμε (συσχετίσεων ή συνδιακύμανσης) που είναι μεγαλύτερες από τη μέση τιμή όλων των ιδιοτιμών. Σημειώστε πως

- αν χρησιμοποιούμε τον πίνακα συσχετίσεων η μέση τιμή όλων των ιδιοτιμών είναι 1
- αν χρησιμοποιούμε τη μέθοδο μεγίστης πιθανοφάνειας το πακέτο χρησιμοποιεί τον πίνακα συσχετίσεων και τις ιδιοτιμές του.
- Το κριτήριο δηλαδή που βασίζεται στις ιδιοτιμές μπορούμε να το χρησιμοποιήσουμε και για τη μέθοδο μεγίστης πιθανοφάνειας, άσχετα αν δεν είναι σχετικό με αυτή τη μέθοδο
- Στην πράξη η χρήση της παραγοντικής ανάλυσης είναι μια επαναληπτική διαδικασία όπου προσαρμόζουμε διάφορα μοντέλα και επιλέγουμε αυτό που θεωρούμε καλύτερο



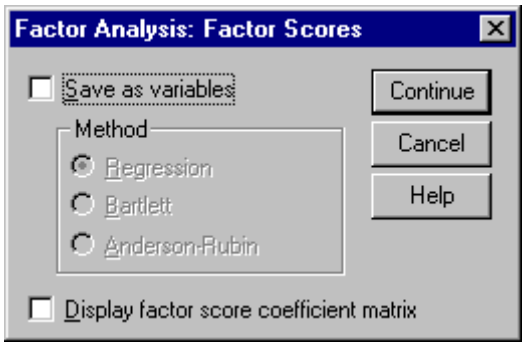
Εικόνα 8.3. Επιλέγοντας μέθοδο και αριθμό παραγόντων

Από το πλήκτρο ***Rotation*** μπορούμε να επιλέξουμε τη μέθοδο περιστροφής που θέλουμε. Αν δεν επιλέξουμε κάποια δεν θα γίνει περιστροφή και θα πάρουμε απλά τη λύση χωρίς περιστροφή. Οι μέθοδοι που προσφέρονται είναι αυτοί που αναφέρθηκαν. Αν επιλεγεί μη ορθογώνια περιστροφή (oblique ή Promax) πρέπει να συμπληρώσουμε την επιλογή ***Delta*** (μια τιμή κοντά στο 1 προτείνεται στη βιβλιογραφία) και αν επιλέξουμε τη μέθοδο Promax πρέπει να συμπληρώσουμε την επιλογή ***Kappa.*** Τέλος επιλέγοντας ***Loading plots*** θα πάρουμε τα γραφήματα των παραγόντων που έχουμε επιλέξει. Αν οι παράγοντες είναι 3 θα πάρουμε 3-διάστατο γράφημα και όχι 3 2-διάστατα γραφήματα που θα ήταν πιο βολικό. Δυστυχώς για να κάνουμε κάτι τέτοιο πρέπει να δουλέψουμε από το Syntax και όχι από τα menu. Το πακέτο θα μας εμφανίσει τον πίνακα του μετασχηματισμού που χρησιμοποιήθηκε για την περιστροφή κάτω από τον τίτλο Transfromation Matrix.



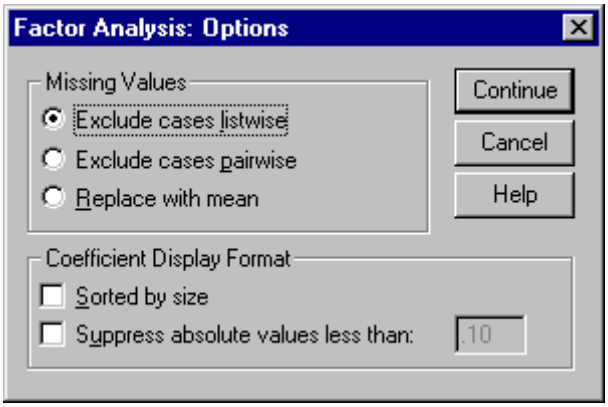
Εικόνα 8.4. Επιλογή μεθόδου περιστροφής

Στην εικόνα 8.5 βλέπουμε τις επιλογές από το πλήκτρο ***Scores***. Επιλέγοντας ***Save as variables*** το πακέτο θα δημιουργήσει τόσες νέες μεταβλητές όσοι κι οι παράγοντες που επέλεξα χρησιμοποιώντας τυποποιημένα μοναδικά ονόματα της μορφής fac1_1 κλπ. Αν ξανατρέξουμε πάλι μια παραγοντική ανάλυση και ζητήσουμε πάλι τα factor scores τότε το όνομα θα είναι fac2_1 και ούτω καθεξής. Αν έχω εκτιμήσει τους παράγοντες με τη μέθοδο των κυρίων συνιστωσών τα factor scores μπορούν να υπολογιστούν με ακρίβεια. Αλλιώς θα πρέπει να επιλέξω κάποια από τις 3 μεθόδους (Regression, Bartlett και Anderson-Rubin) Τέλος η επιλογή ***Display factor score coefficient matrix*** θα μου εμφανίζει τον πίνακα με τους συντελεστές. Προσοχή αυτός ο πίνακας περιέχει τους συντελεστές με τους οποίους μπορώ να εκφράσω έναν παράγοντα ως γραμμικό συνδυασμό των μεταβλητών.



Εικόνα 8.5. Επιλογές για τον υπολογισμό των factor scores.

Τέλος από το παράθυρο ***Options*** που βλέπουμε στην εικόνα 6 μπορούμε να επιλέξουμε κάποιες άλλες επιλογές σχετικά με το χειρισμό των missing values και με τον τρόπο που θέλουμε να εμφανίζονται τα αποτελέσματα των εκτιμημένων παραγόντων. Έτσι μπορούμε να δούμε τις μεταβλητές με σειρά απόλυτης τιμής (από τη μεγαλύτερη και άρα πιο σημαντική για τον παράγοντα προς τη μικρότερη) ή ακόμα και να μη εμφανιστούν μικροί συντελεστές. Αυτό είναι χρήσιμο όταν έχουμε πολλές μεταβλητές επειδή έτσι μπορούμε να διαβάσουμε γρήγορα τα αποτελέσματα.



Εικόνα 8.6. Διάφορες άλλες επιλογές.

Τελειώνοντας θα πρέπει να τονίσουμε τα εξής

- όταν χρησιμοποιούμε τη μέθοδο κυρίων συνιστωσών το πρόγραμμα χρησιμοποιεί τον όρο Components αντί για Factors σε όλα τα αποτελέσματα
- το πρόγραμμα εμφανίζει διάφορους πίνακες αποτελεσμάτων. Για να αποφευχθεί τυχόν λάθος οι πίνακες αυτοί και τα περιεχόμενα τους δίνονται ως εξής

Πίνακας	Περιεχόμενα
Factor Matrix	Οι επιβαρύνσεις για κάθε παράγοντα και μεταβλητή
Rotated Factor Matrix	Οι επιβαρύνσεις για κάθε παράγοντα και μεταβλητή μετά την περιστροφή
Factor Transformation Matrix	Ο πίνακας που χρησιμοποιήθηκε για την περιστροφή
Factor Score Coefficient Matrix	Ο πίνακας με τα σκορ των παραγόντων
Factor Score Covariance Matrix	Ο πίνακας συνδιακύμανσης των σκορ. Επειδή το μοντέλο είναι ορθογώνιο και άρα οι παράγοντες ασυσχέτιστοι πρέπει να είναι διαγώνιος. Αυτό δεν ισχύει αν έχουμε κάνει μη ορθογώνια περιστροφή οπότε υπάρχει συσχέτιση.

Αν και γενικά το SPSS εμφανίζει κάτω από κάθε πίνακα αποτελεσμάτων πληροφορίες σχετικά με το ποια μέθοδο εκτίμησης χρησιμοποιήσαμε και τον αριθμό των παραγόντων, οι 2 μέθοδοι διαφέρουν αρκετά στα output που δίνει το πακέτο και επομένως μπορεί κανείς εύκολα να διακρίνει ποια μέθοδος χρησιμοποιήθηκε.

Επίσης αν και αρκετά συχνά αναφέρεται πως μπορεί κάποιος να χρησιμοποιήσει το SPSS και συγκεκριμένα τις διαδικασίες της παραγοντικής ανάλυσης για να κάνει ανάλυση σε κύριες συνιστώσες κάτι τέτοιο δεν είναι δυνατόν να γίνει και χρειάζονται μετατροπές στα αποτελέσματα ώστε να βρει κανείς τις κύριες συνιστώσες.

9 ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ

9.1 Εισαγωγή

Η ανάλυση κατά συστάδες είναι μια μέθοδος που σκοπό έχει να κατατάξει σε ομάδες τις υπάρχουσες παρατηρήσεις χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Με άλλα λόγια η ανάλυση κατά συστάδες εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθμό μεταβλητών με σκοπό να δημιουργήσει ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους.

Μια επιτυχημένη ανάλυση θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς αλλά παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο.

Η σημαντική διαφορά της μεθόδου από τη διακριτική ανάλυση, η οποία θα εξετασθεί σε επόμενο κεφάλαιο, είναι πως στη διακριτική ανάλυση γνωρίζουμε κάποια ομαδοποίηση ως προς κάποιο χαρακτηριστικό των παρατηρήσεων και θέλουμε να φτιάξουμε κάποιον κανόνα που θα μας βοηθήσει να κατατάξουμε καινούριες παρατηρήσεις. Βλέπουμε λοιπόν πως καθώς οι 2 μέθοδοι έχουν κάποια κοινά χαρακτηριστικά ως προς τον τρόπο που λειτουργούν, μπορούν να λειτουργήσουν συμπληρωματικά.

Παραδείγματα εφαρμογών της ανάλυσης σε συστάδες είναι τα ακόλουθα:

- Οι βιολόγοι ενδιαφέρονται να κατατάξουν διαφορετικά είδη ζώων σε ομάδες με βάση κάποια χαρακτηριστικά τους,
- Στο μάρκετινγκ το ενδιαφέρον είναι πως μπορούν να ομαδοποιηθούν οι πελάτες σύμφωνα με τα στοιχεία που υπάρχουν σχετικά με τις αγοραστικές τους συνήθειες και τα δημογραφικά χαρακτηριστικά τους. Κάτι τέτοιο είναι πολλαπλά χρήσιμο κυρίως για διαφημιστικούς λόγους, για παράδειγμα κάποια προϊόντα απευθύνονται σε συγκεκριμένη αγοραστική ομάδα
- Στην αρχαιολογία ενδιαφέρεται κανείς να κατατάξει τα ευρήματα μιας ανασκαφής σε ομάδες που για παράδειγμα αντανακλούν διαφορετικές χρονικές περιόδους. Για

να το επιτύχει αυτό προσπαθεί να χρησιμοποιήσει μια σειρά από μετρήσεις σχετικές με τα ευρήματα ώστε με βάση αυτές τις μετρήσεις να ομαδοποιήσει τα ευρήματα

- Οι σχεδιαστές ηλεκτρονικών σελίδων ενδιαφέρονται να βρουν και να ομαδοποιήσουν τη συμπεριφορά των χρηστών του Internet ανάλογα με τον τρόπο με τον οποίο σερφάρουν ανάμεσα σε διαφορετικές σελίδες. Επομένως η συμπεριφορά τους όπως καταγράφεται με τη διαδοχική εναλλαγή σελίδων προσφέρει δεδομένα με σκοπό την ομαδοποίηση των χρηστών.

Εκτός από τα παραδείγματα που μόλις αναφέραμε η ανάλυση σε συστάδες βρίσκει πληθώρα εφαρμογών σχεδόν σε κάθε επιστήμη και επομένως αποτελεί ένα πολυτιμότατο εργαλείο στα χέρια όλων των επιστημονικών κλάδων.

Δύο βασικές έννοιες για την ανάλυση κατά συστάδες αλλά όχι μόνο, είναι οι έννοιες της απόστασης και της ομοιότητας. Μπορείτε εύκολα να διαπιστώσετε πως αυτές οι δύο έννοιες είναι αντίθετες, παρατηρήσεις που είναι όμοιες θα έχουν μεγάλη ομοιότητα και μικρή απόσταση. Οι έννοιες αυτές ουσιαστικά ποσοτικοποιούν αυτό που στην καθημερινή γλώσσα εννοούν. Δηλαδή παρατηρήσεις που μοιάζουν πολύ μεταξύ τους, έχουν με απλά λόγια σχετικά όμοιες τιμές, θα πρέπει να έχουν πολύ μεγάλη τιμή για το μέτρο της ομοιότητας που θα χρησιμοποιήσουμε και πολύ μικρή απόσταση. Θα μιλήσουμε εκτενέστερα για τις έννοιες αυτές στην επόμενη ενότητα. Οι έννοιες αυτές είναι πολύ χρήσιμες καθώς μας επιτρέπουν να μετρήσουμε πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους και επομένως να τις τοποθετήσουμε στην ίδια ομάδα. Επομένως σκοπός της ανάλυσης σε συστάδες είναι να δημιουργήσουμε ομάδες μέσα στις οποίες οι παρατηρήσεις απέχουν λίγο ενώ παρατηρήσεις διαφορετικών ομάδων απέχουν μεταξύ τους αρκετά.

Τελειώνοντας αυτή την μικρή εισαγωγή θα πρέπει να παρατηρήσουμε πως υπάρχουν αρκετές διαφορετικές προσεγγίσεις για το πως μπορούμε να ομαδοποιήσουμε τα δεδομένα μας. Κάποιες από αυτές στηρίζονται αποκλειστικά σε ad-hoc επιχειρήματα ενώ άλλες στηρίζονται στη θεωρία πιθανοτήτων.

Οι βασικότερες και πιο διαδεδομένες προσεγγίσεις είναι:

- Ιεραρχικές μέθοδοι: Ξεκινάμε με κάθε παρατήρηση να είναι από μόνη της μια ομάδα. Σε κάθε βήμα ενώνουμε τις 2 παρατηρήσεις που έχουν πιο μικρή απόσταση. Αν 2 παρατηρήσεις έχουν ενωθεί σε προηγούμενο βήμα ενώνουμε μια προϋπάρχουσα ομάδα με μια παρατήρηση μέχρι να φτιάξουμε μια ομάδα. Κοιτώντας τα αποτελέσματα διαλέγουμε στις πόσες ομάδες θα σταματήσουμε.

- K-Means. Ο αριθμός των ομάδων είναι γνωστός από πριν. Με έναν επαναληπτικό αλγόριθμο μοιράζουμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην παρατήρηση.

- Στατιστικές μέθοδοι: Και οι δύο μέθοδοι που είπαμε στηρίζονται καθαρά σε αλγοριθμικές λύσεις και δεν προϋποθέτουν κάποιο μοντέλο. Υπάρχουν αρκετές μέθοδοι στατιστικές όπου ξεκινώντας από κάποιες υποθέσεις κατατάσσουμε τις παρατηρήσεις. Δυστυχώς αυτές οι μέθοδοι έχουν αρκετά υπολογιστικά προβλήματα και για αυτό δεν προσφέρονται από πολλά στατιστικά πακέτα που χρησιμοποιούνται στην πράξη.

Θα πρέπει να τονιστεί πως η εξάπλωση και η ευρεία χρήση των υπολογιστών σε θέματα ανάλυσης σε συστάδες, σε συνδυασμό με την ευρεία διαθεσιμότητα δεδομένων (πχ ένα super market καταγράφει καθημερινά σε δυαδική μορφή όλες τις αγορές που έγιναν) έχει οδηγήσει τα τελευταία χρόνια σε μια καινούρια θεώρηση των προβλημάτων της ανάλυσης που κυρίως βασίζεται στο πως θα αναλυθούν και θα ομαδοποιηθούν τα δεδομένα τεράστιων βάσεων δεδομένων. Αυτό έχει ως συνέπεια να δίνεται ολοένα και μεγαλύτερο βάρος στον υπολογιστικό φόρτο των μεθόδων και να ερευνώνται μέθοδοι για να μπορέσει κανείς να διαχειριστεί μεγάλο όγκο δεδομένων. Το τίμημα που συνήθως πληρώνουμε είναι μικρότερη ακρίβεια και ορθότητα των αποτελεσμάτων.

Τελειώνοντας αυτή την εισαγωγή θα πρέπει να τονίσουμε ότι μερικές φορές η ανάλυση σε ομάδες μπορεί να έχει και άλλους σκοπούς εκτός από την απλή ομαδοποίηση των δεδομένων. Έτσι η ανάλυση σε ομάδες μπορεί να χρησιμοποιηθεί για

- να αποκτηθεί κάποια γνώση σχετικά με τα δεδομένα, αν για παράδειγμα παρουσιάζουν ομοιότητες, ποιες μεταβλητές μοιάζουν να έχουν διακριτική ικανότητα κλπ

- τη διερεύνηση σχέσεων στα δεδομένα, συνήθως έχοντας ένα σετ δεδομένων στα χέρια μας έχουμε μια πολύ ασαφή εικόνα για το τι περιέχουν τα δεδομένα τι είδους σχέσεις υπάρχουν κλπ.

- τη μείωση των διαστάσεων του προβλήματος. Ειδικά στη σύγχρονη εποχή το πλήθος των δεδομένων που συγκεντρώνεται είναι τεράστιο χωρίς αυτό να σημαίνει ότι και η πληροφορία που περιέχεται είναι εξίσου τεράστια. Υπάρχουν επικαλύψεις, μεταβλητές χωρίς ιδιαίτερο ενδιαφέρον κλπ. Επομένως ομαδοποιώντας τα δεδομένα αποκτούμε μια εικόνα σχετικά με τις μεταβλητές που παρουσιάζουν ενδιαφέρον και επικεντρωνόμαστε σε αυτές

- δημιουργία και έλεγχο υποθέσεων σχετικά με τα δεδομένα. Πολλές φορές ο ερευνητής υποψιάζεται την ύπαρξη κάποιων ομάδων με βάση κάποιο θεωρητικό μοντέλο που έχει στο μυαλό του (πχ κάποια είδη του ζωικού βασιλείου μοιάζουν

μεταξύ τους επομένως ο ερευνητής θέλει να διαπιστώσει κατά πόσο μπορεί να τα κατατάξει στην ίδια ομάδα).

- για πρόβλεψη καινούριων τιμών. Έχοντας δημιουργήσει ομάδες από παρατηρήσεις σε πολλές εφαρμογές ενδιαφερόμαστε στο να κατατάξουμε καινούριες παρατηρήσεις. Για παράδειγμα μια τράπεζα έχει κατατάξει τους πελάτες της σε καλούς, μέτριους και κακούς και θέλει να κατατάσσει καινούριους πελάτες σε αυτές τις κατηγορίες με βάση τα χαρακτηριστικά τους.

9.2 Η Απόσταση

9.2.1 Η έννοια της απόστασης

Η απόσταση είναι μια θεμελιώδης έννοια στην πολυμεταβλητή ανάλυση και όχι μόνο για την ανάλυση δεδομένων. Σκοπός της απόστασης είναι να μετρήσει πόσο απέχουν δύο παρατηρήσεις, να ποσοτικοποιήσει δηλαδή αν μοιάζουν ή όχι οι παρατηρήσεις.

Για παράδειγμα ας υποθέσουμε πως ενδιαφερόμαστε για δύο μεταβλητές το βάρος και το ύψος, δηλαδή για κάθε παρατήρηση έχουμε μετρήσεις για αυτές τις δύο μεταβλητές. Αν συμβολίσουμε τις δύο παρατηρήσεις ως $y=(y_1,y_2)$ και $x=(x_1,x_2)$ τότε μια πρώτη προσέγγιση για την επιλογή μιας απόστασης ανάμεσα στις δύο παρατηρήσεις θα ήταν η ευκλείδεια απόσταση

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$$

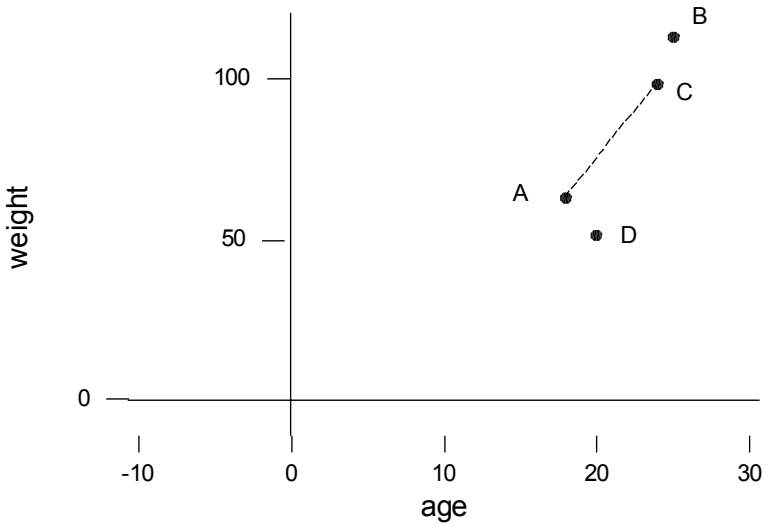
την οποία μπορούμε να γενικεύσουμε για την περίπτωση που έχουμε παρατηρήσεις σε p μεταβλητές, δηλαδή $y=(y_1,y_2,...,y_p)$ και $x=(x_1,x_2,...,x_p)$ τότε η αντίστοιχη απόσταση μπορεί να οριστεί ως

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

Αν και η ευκλείδεια απόσταση είναι θεμελιώδης στα μαθηματικά και την καθημερινή μας ζωή, στη στατιστική τα πράγματα είναι λίγο διαφορετικά. Για να δούμε γραφικά τι συμβαίνει στην περίπτωση των δύο μεταβλητών (που επειδή βρισκόμαστε στο δισδιάστατο χώρο μπορούμε να έχουμε μια οπτική περιγραφή). Ας υποθέσουμε πως έχουμε 4 φοιτητές για τους οποίους ξέρουμε την ηλικία και το βάρος τους. Συγκεκριμένα έχουμε

	Ηλικία	Βάρος
A	18	64
B	25	115
C	24	100
D	20	52

Χρησιμοποιώντας τις τιμές των δύο μεταβλητών ως συντεταγμένες τα παραπάνω δεδομένα απεικονίζονται στο παρακάτω γράφημα.



Γράφημα 2.1. Γραφική απεικόνιση μαθητών ως προς τα χαρακτηριστικά βάρους και ηλικίας

Επομένως η απόσταση ανάμεσα στον Α και τον C φοιτητή είναι

$(18\text{-}24)^2 + (64\text{-}100)^2 = 1332$

Όμως παρατηρείστε πως η κλίμακα είναι διαφορετική στις παρατηρήσεις. Η ηλικία παίρνει τιμές στο διάστημα 18-25 ενώ το βάρος στο διάστημα 52-115 και συνεπώς λόγω της διαφοράς κλίμακας η απόσταση καθορίζεται σε ένα πολύ μεγάλο βαθμό από το βάρος. Για παράδειγμα αν το βάρος μετριόταν σε γραμμάρια τότε η διαφορά θα ήταν ακόμα πιο μεγάλη. Συγκεκριμένα η απόσταση θα ήταν πια

$(18\text{-}24)^2 + (64000\text{-}100000)^2 = 1269000036$

και ουσιαστικά η επίδραση της ηλικίας θα ήταν αμελητέα μπροστά στο βάρος.

Επομένως η ευκλείδεια απόσταση δεν μοιάζει να είναι ένα καλό μέτρο απόστασης.

Γενικεύοντας, θα μιλάμε για απόσταση όταν έχουμε μια συνάρτηση που μετρά το πόσο απέχουν (διαφέρουν) μεταξύ τους δύο παρατηρήσεις. Στην πραγματικότητα μια συνάρτηση $f(x, y)$ είναι απόσταση αν ισχύουν οι παρακάτω ιδιότητες

1. $f(x,y) = f(y,x) \geq 0$ (συμμετρική ιδιότητα)
2. $f(x,y) \leq f(x,z) + f(y,z)$ (τριγωνική ιδιότητα)
3. $f(x,y) \neq 0 \Leftrightarrow x \neq y$
4. $f(x,x) = 0$

Η πιο σημαντική ιδιότητα είναι η τριγωνική η οποία, όμως, δεν ικανοποιείται από πολλά μέτρα που χρησιμοποιούνται στην πράξη.

Επίσης πολλοί συγγραφείς παρατηρούν πως δεν είναι απαραίτητο η απόσταση να είναι συμμετρική. Για παράδειγμα οι εξαγωγές μιας χώρας σε μια άλλη δεν είναι απαραίτητα συμμετρικές, παρόλα αυτά θα μπορούσε κάποιος να χρησιμοποιήσει έναν τέτοιο πίνακα αποστάσεων για να δουλέψει με μια σειρά από χώρες και προϊόντα.

Η ευκλείδεια απόσταση που είδαμε αμέσως πριν ικανοποιεί αυτές τις ιδιότητες αλλά από στατιστική άποψη δεν είναι επαρκής. Ένας γνωστός τρόπος ώστε να φέρουμε κάθε μεταβλητή σε συγκρίσιμη κλίμακα είναι να διαιρέσουμε κάθε μεταβλητή με την τυπική της απόκλιση κι επομένως αφού όλες οι μεταβλητές πια θα αναφέρονται σε μονάδες τυπικής απόκλισης έχουμε εξαλείψει το πρόβλημα. Δηλαδή αν s_i^2 συμβολίσουμε τη διακύμανση της *i* μεταβλητής τότε η απόσταση που θα χρησιμοποιήσουμε έχει τη μορφή:

$$d(x,y) = \sqrt{\sum_{i=1}^{p} \left(\frac{x_i - y_i}{s_i^2} \right)^2} \tag{9.1}$$

Από στατιστικής άποψης η απόσταση αυτή είναι πιο ενδιαφέρουσα και επιτρέπει πιο καλές συγκρίσεις ανάμεσα στις μεταβλητές. Το μόνο μειονέκτημα όμως που έχει είναι πως δεν λαμβάνει υπόψη της τις συνδιακυμάνσεις ανάμεσα στις μεταβλητές. Αν δύο μεταβλητές είναι πολύ συσχετισμένες τότε η απόσταση των παρατηρήσεων ουσιαστικά οφείλονται μόνο σε μια από αυτές αφού η άλλη μεταβλητή απλά ακολουθεί την πρώτη εξαιτίας της συσχέτισης. Επομένως θα ήταν χρήσιμη μια απόσταση που να λάμβανε υπόψη της τις συνδιακυμάνσεις. Τέτοια απόσταση είναι η απόσταση του Mahalanobis που υπολογίζεται ως

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

όπου S ο δειγματικός πίνακας διακυμάνσεων.

Παρατηρείστε πως αν ο S είναι διαγώνιος η απόσταση του Mahalanobis γίνεται ίδια με την απόσταση (9.1).

Αυτό που μόλις συζητήσαμε είναι πως μπορεί κανείς να κατασκευάσει μια απόσταση που να έχει στατιστική ερμηνεία. Στη βιβλιογραφία υπάρχει μια τεράστια ποικιλία από

μέτρα αποστάσεων (αν και στην πραγματικότητα κάποια από αυτά δεν είναι αποστάσεις με την αυστηρή έννοια επειδή δεν ικανοποιούν τις ιδιότητες που αναφέραμε πριν). Επίσης πολλά μέτρα είναι μέτρα ομοιότητας (similarity) και όχι απόστασης αλλά όπως είπαμε και πριν ένα μέτρο απόστασης μπορεί εύκολα να μετασχηματιστεί σε μέτρο ομοιότητας και το αντίστροφο.

Στη συνέχεια θα περιγράψουμε κάποια από αυτά και συγκεκριμένα θα συζητήσουμε για τέτοια μέτρα ανάλογα με το είδος των δεδομένων που έχουμε.

9.2.2 Μέτρα απόστασης

Θα δούμε τώρα κάποια μέτρα απόστασης που χρησιμοποιούνται στην πράξη. Τα μέτρα αυτά έχουν χωριστεί σε ομάδες ανάλογα με το είδος των δεδομένων στα οποία μπορούν να εφαρμοσθούν. Κάποια από αυτά είναι ομοιότητες, αλλά θα περιγράψουμε αποστάσεις και ομοιότητες ενιαία.

Συνεχή Δεδομένα

Η περίπτωση των συνεχών δεδομένων είναι ίσως η απλούστερη αλλά και η περισσότερο διαδεδομένη. Υπάρχουν πολλές αποστάσεις που έχουν χρησιμοποιηθεί για να μετρήσουν την απόσταση ανάμεσα σε συνεχή δεδομένα. Θα περιγράψουμε μερικές από αυτές. Θα πρέπει να παρατηρήσουμε πως δεν ικανοποιούν απαραίτητα τις ιδιότητες που είδαμε πριν από λίγο

- *Ευκλείδεια απόσταση*

Όπως είπαμε και προηγουμένως η ευκλείδεια απόσταση αποτελεί την πιο απλή και την πιο γνωστή απόσταση ανάμεσα σε συνεχή δεδομένα. Μερικές χρήσιμες ιδιότητές της είναι οι εξής:

- Η ευκλείδεια απόσταση εξαρτάται πολύ από την κλίμακα μέτρησης κι επομένως αλλάζοντας την κλίμακα μπορούμε να πάρουμε ολότελα διαφορετικές αποστάσεις.
- Επίσης μεταβλητές με μεγάλες απόλυτες τιμές έχουν πολύ μεγαλύτερο βάρος και σχεδόν καθορίζουν την απόσταση ανάμεσα σε παρατηρήσεις.

Η ερμηνεία της απόστασης είναι πολύ εύκολο να αποδοθεί γεωμετρικά. Στην πραγματικότητα η απόσταση αγνοεί τις στατιστικές ιδιότητες των παρατηρήσεων όπως για παράδειγμα τη μεταβλητότητα κάθε μεταβλητές.

Δεδομένου ότι παίρνουμε τετραγωνικές αποκλίσεις outliers έχουν μεγάλη επίδραση στον υπολογισμό της απόστασης.

- *City-block (Manhattan) distance*

Η απόσταση Manhattan μοιάζει πολύ με την ευκλείδεια απόσταση με τη διαφορά ότι αντί για τετραγωνικές αποκλίσεις χρησιμοποιούμε απόλυτες αποκλίσεις. Επομένως η απόσταση ορίζεται ως

$$d(x,y)=\sum_{i=1}^{p}\left|x_i-y_i\right|.$$

Συνήθως λόγω της ομοιότητας της με την ευκλείδεια απόσταση δίνει περίπου ίδια αποτελέσματα εκτός από την περίπτωση που υπάρχουν outliers όπου επειδή τους δίνει μικρότερο βάρος (εξαιτίας της απόλυτης τιμής) μπορεί να οδηγήσει σε πιο ανθεκτικά αποτελέσματα.

Θα πρέπει να σημειώσει κανείς πως και αυτή η απόσταση αγνοεί τις στατιστικές ιδιότητες των δεδομένων.

- *Απόσταση Minkowski (or Lq norm)*

Η απόσταση Minkowski κατά κάποιον τρόπο γενικεύει την Ευκλείδεια απόσταση και την απόσταση Manhattan.

Η απόσταση ορίζεται ως

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

Η τιμή της παραμέτρου *q* μπορεί να χρησιμοποιηθεί για να δώσει ιδιαίτερο βάρος σε κάποιες αποκλίσεις. Προφανώς αν *q=1* προκύπτει η απόσταση Manhattan ενώ αν *q=2* η ευκλείδεια απόσταση.

Μια γενίκευση είναι η Power distance, που ορίζεται ως

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

όπου τα r και q είναι παράμετροι που ορίζει ο ερευνητής. Θα πρέπει να σημειωθεί πως ο παραπάνω γενικός ορισμός επιτρέπει τη χρήση πολλών άλλων αποστάσεων που έχουν χρησιμοποιηθεί για στατιστικούς σκοπούς.

- *Chebychev distance*

Η απόσταση Chebychev, σε αντίθεση με τις υπόλοιπες αποστάσεις που είδαμε δεν χρησιμοποιεί όλες τις αποκλίσεις αλλά μόνο τη μεγαλύτερη εξ αυτών. Συγκεκριμένα η απόσταση ορίζεται ως

$$d(x,y)=\max\left\{\left(x_i-y_i\right),i=1,\ldots p\right\}.$$

Η απόσταση αυτή είναι χρήσιμη όταν κανείς θέλει να θεωρήσεις δύο παρατηρήσεις διαφορετικές αν έχουν διαφορές τουλάχιστον σε μια μεταβλητή.

Επειδή η απόσταση χρησιμοποιεί μόνο τη μεγαλύτερη απόκλιση εξαρτάται πολύ από τις διαφορές στην κλίμακα των μεταβλητών και επομένως αν οι κλίμακες είναι διαφορετικές ουσιαστικά θα αντικατοπτρίζει τη διαφορά στη μεταβλητή με την μεγαλύτερη κλίμακα.

Όλες οι παραπάνω αποστάσεις έχουν το μειονέκτημα ότι δεν λαμβάνουν υπόψη τους τις όποιες διαφορές στην κλίμακα των μεταβλητών όπως επίσης και τις διαφορές στις διακυμάνσεις τους. Επίσης τυχόν συσχετίσεις ανάμεσα στις μεταβλητές δεν λαμβάνονται υπόψη και έτσι κατά κάποιον τρόπο αν υπάρχουν συσχετισμένες μεταβλητές η απόσταση ανάμεσα σε δύο παρατηρήσεις μπορεί να είναι πλασματική.

Ένα μέτρο απόστασης που εκ κατασκευής βασίζεται σε στατιστικές έννοιες και λαμβάνει υπόψη διακυμάνσεις και συνδιακυμάνσεις είναι η απόσταση Mahalanobis που ορίσαμε προηγουμένως.

Τέλος θα πρέπει να πούμε πως εκτός από τα μέτρα απόστασης που μόλις είδαμε μπορεί κανείς να ορίσει και μέτρα συσχέτισης όπως για παράδειγμα ο γνωστός συντελεστής συσχέτισης μόνο που τώρα δεν χρησιμοποιούμαι διαφορετικές παρατηρήσεις αλλά διαφορετικές μεταβλητές. Δηλαδή αθροίζουμε ως προς όλες τις μεταβλητές και όχι ως προς όλες τις παρατηρήσεις. Πρέπει να τονιστεί πως 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 μια μονάδα σε κάθε μεταβλητή τότε η τιμή του συντελεστή είναι η μεγαλύτερη δυνατή και ίση με 1. Παρόλα αυτά οι παρατηρήσεις διαφέρουν σε όλες τις μεταβλητές και συνεπώς δεν είναι όμοιες. Συνεπώς η χρήση του συντελεστή αν και τυπικά μπορεί να γίνει στην πράξη δεν δουλεύει καλά. Από την άλλη είναι ενδιαφέρον το γεγονός πως ως ένα αρκετά διαδεδομένο στατιστικό μέτρο η ερμηνεία του είναι αρκετά εύκολη.

Δυαδικά δεδομένα

Έστω ότι τα δεδομένα μας αφορούν μια σειρά από μεταβλητές για κάθε παρατήρηση για τις οποίες έχουμε δυαδική κατάσταση, δηλαδή η τιμή 1 δηλώνει την παρουσία του χαρακτηριστικού και 0 την απουσία. Για παράδειγμα σε ιατρικά δεδομένα οι μεταβλητές μπορεί να είναι διάφορα συμπτώματα (η ύπαρξη ή όχι αυτών), όπως υποθερμία, διάρροια, πυρετός κλπ.

Για τέτοιας μορφής δεδομένα υπάρχει μια ποικιλία από αποστάσεις που μπορεί κανείς να χρησιμοποιήσει.

Στην πράξη κατασκευάζουμε έναν πίνακα ανομοιότητας (ή ομοιότητας αντίστοιχα) με βάση τον οποίο υπολογίζουμε την απόσταση.

Συγκεκριμένα για τον υπολογισμό ενός μέτρου ομοιότητας s(x,y) ή ανομοιότητας d(x,y)

Ανάμεσα στην x και τη y παρατήρηση χρησιμοποιούμε τον παρακάτω πίνακα συνάφειας:

		Παρατήρηση y		
		1	0	
Παρατήρηση x	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	

όπου a, b, c, d δηλώνουν το πλήθος των συνδυασμών (1,1) (1,0) (0,1) (0,0) αντίστοιχα. Το κελί (0,0) υποδηλώνει τα χαρακτηριστικά τα οποία είναι απόντα και στις δύο παρατηρήσεις ενώ το κελί (1,0) τα χαρακτηριστικά τα οποία είναι παρόντα στην παρατήρηση x και απόντα στην παρατήρηση y, κοκ.

Πριν μιλήσουμε για τα μέτρα απόστασης και ομοιότητας πρέπει να παρατηρήσουμε πως οι μεταβλητές μπορούν να χωριστούν σε δύο κατηγορίες. Η πρώτη κατηγορία περιλαμβάνει τις συμμετρικές μεταβλητές όπου οι τιμές 0 και 1 έχουν την ίδια σημασία, για παράδειγμα αν η μεταβλητή μας είναι ο κλάδος που ανήκει μια επιχείρηση και συγκεκριμένα αν ανήκει στη βιομηχανίας (1 αν ανήκει και 0 αν δεν ανήκει). Η περίπτωση των ασύμμετρων μεταβλητών αφορά την περίπτωση που μια εκ των δύο κατηγοριών δεν είναι χρήσιμη για την εξαγωγή συμπερασμάτων. Τέτοια πληροφορία συνήθως περιέχετε σε ιατρικά παραδείγματα όπου η απουσία κάποιων συμπτωμάτων δεν είναι το ίδιο σημαντική ως πληροφορία με την παρουσία αυτών.

Στην περίπτωση συμμετρικών μεταβλητών όλα τα κελιά έχουν την ίδια βαρύτητα. Στην περίπτωση ασύμμετρων μεταβλητών το κύριο βάρος πέφτει στο κελί (1,1) δηλαδή στην κοινή παρουσία κάποιων χαρακτηριστικών.

Ονομασία	**s(x,y)**	**d(x,y)**
Simple matching coefficient (Sokal and Michener, 1958)	$\frac{a+d}{a+b+c+d}$	$\frac{b+c}{a+b+c+d}$
Rogers and Tarimoto (1960)	$\frac{a+d}{(a+d)+2(b+c)}$	$\frac{2(b+c)}{(a+d)+2(b+c)}$
Sokal and Sneath (1963)	$\frac{2(a+d)}{2(a+d)+(b+c)}$	$\frac{(b+c)}{2(a+d)+(b+c)}$

Πίνακας 9.1. Αποστάσεις χρήσιμες για συμμετρικές δυαδικές μεταβλητές

Στον πίνακα 9.1 μπορεί κανείς να δει κάποιες ομοιότητες (αποστάσεις) για συμμετρικά δεδομένα. Παρατηρείστε πως μια ομοιότητα μπορεί πολύ εύκολα να μετασχηματιστεί σε απόσταση χρησιμοποιώντας απλά το συμπληρωματικό της.

Το πρώτο μέτρο, γνωστό και ως συντελεστής ομοιότητας, ουσιαστικά μετράει τον αριθμό των μεταβλητών για τις οποίες οι δύο παρατηρήσεις συμφωνούν (κελιά (0,0) και (1,1)). Τα δύο επόμενα μέτρα χρησιμοποιούν την ίδια πληροφορία αλλά δίνουν διαφορετικό βάρος στα κελιά (0,0) και (1,1) από ότι στα κελιά (0,1) και (1,0). Ουσιαστικά δηλαδή χρησιμοποιούν με διαφορετικά βάρη την πληροφορία για τα κελιά που δηλώνουν συμφωνία.

Παρατηρείστε πως αν το κελί (0,0) (κοινή απουσία των χαρακτηριστικών) δεν είναι πραγματικά ενδιαφέρον (πχ σπάνια χαρακτηριστικά) τότε ο συντελεστής αυτός δεν είναι ιδιαίτερα χρήσιμος. Σε αυτή την περίπτωση η χρήση ασύμμετρων συντελεστών όπως αυτοί του πίνακα 9.2 είναι πιο εύχρηστοι

Ονομασία	s(x,y)	d(x,y)
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 and Sneath (1963)	$\frac{a}{a+2(b+c)}$	$\frac{2(b+c)}{a+2(b+c)}$

Πίνακας 9.2. Αποστάσεις χρήσιμες για ασύμμετρες δυαδικές μεταβλητές

Παρατηρείστε πως οι συντελεστές του πίνακα 9.2 δεν λαμβάνουν υπόψη τους το κελί (0,0) (η συχνότητα d του κελιού αυτού δεν χρησιμοποιείται καθόλου). Ο συντελεστής του Jackard είναι ίδιος με τον απλό συντελεστή ομοιότητας αλλά αγνοώντας το κελί (0,0). Αντίστοιχα οι υπόλοιποι συντελεστές δίνουν μεγαλύτερο η μικρότερο βάρος στα κελιά συμφωνίας ή ασυμφωνίας.

Όλοι οι συντελεστές που είδαμε παίρνουν τιμές στο διάστημα (0,1)

Παράδειγμα 9.1

Έστω 4 χρήστες του Ιντερνετ. Για κάθε χρήστη έχουμε μια σειρά από ιστοσελίδες και δίνουμε την τιμή 1 αν ο χρήστης επισκέφτηκε την ιστοσελίδα και 0 αν όχι. Συνολικά τα δεδομένα είναι τα εξής

		ιστοσελίδα									
		1	2	3	4	5	6	7	8	9	10
Χρήστης	Α	1	1	1	0	0	0	1	0	0	1
	Β	0	1	1	0	1	1	0	0	0	0
	Γ	0	0	0	0	0	1	1	0	0	0
	Δ	0	1	1	0	0	1	0	0	0	0

Πίνακας 9.3. Υποθετικά δεδομένα

Με βάση τον παραπάνω πίνακα ο χρήστης Α επισκέφτηκε τις ιστοσελίδες {1,2,3,7,10} ενώ ο Δ μόνο τις {2,3,6}. Οι αποστάσεις που είδαμε πριν από λίγο μπορούν να χρησιμοποιηθούν για να μετρήσουν την απόσταση ανάμεσα στους χρήστες. Έτσι η απόσταση ανάμεσα στον Α και το Β θα υπολογιστεί με τη χρήση του πίνακα ομοιότητας

		Α		
		1	0	
Β	1	2	2	4
	0	3	3	4
		5	5	

Βλέπουμε λοιπόν πως υπάρχουν 2 ιστοσελίδες που επισκέφτηκαν και οι δύο ενώ υπάρχουν άλλες 3 τις οποίες δεν επισκέφτηκε κανείς από τους δύο. Με βάση αυτό τον πίνακα ομοιότητας τα μέτρα ομοιότητας δίνονται στον παρακάτω πίνακα

Δείκτης Ομοιότητας	τιμή
Simple matching coefficient	0.50
Rogers and Tarimoto	0.33
Sokal and Sneath	0.66
Jaccard	0.286
Dice	0.444
Sokal and Sneath II	0.167

Πίνακας 9.4. Ομοιότητες για τους χρήστες Α και Β

Για τους 4 χρήστες και χρησιμοποιώντας τον συντελεστή ομοιότητας ο πίνακας ομοιότητας είναι ο εξής

	Α	Β	Γ	Δ
Α	1			
Β	0.5	1		
Γ	0.5	0.6	1	
Δ	0.6	0.9	0.7	1

Πίνακας 9.5. Πίνακας ομοιοτήτων για τα δεδομένα του παραδείγματος

Δεδομένα σε ονομαστική κλίμακα

Όταν οι μεταβλητές αναφέρονται σε ονομαστική κλίμακα (πχ μάρκα αυτοκινήτου, χρώμα μαλλιών κλπ) τότε είναι σχετικά δύσκολο να υπολογίσουμε απόσταση. Αυτό οφείλεται στο γεγονός πως ουσιαστικά αν οι δυο παρατηρήσεις έχουν την ίδια τιμή τότε αυτό είναι χρήσιμη πληροφορία αλλά αν δεν έχουν την ίδια τιμή τότε υπάρχουν πολλοί τρόποι που αυτό μπορεί να συμβεί και όλοι αυτοί έχουν το ίδιο βάρος.

Συνήθως η απόσταση που χρησιμοποιείται είναι ο *συντελεστής ομοιότητας* (simple matching approach) ο οποίος ορίζεται ως

$$s(x,y)=\frac{u}{p} \qquad \text{και} \qquad d(x,y)=\frac{p-u}{p},$$

όπου u είναι ο αριθμός των μεταβλητών που έχουν την ίδια τιμή και p ο συνολικός αριθμός μεταβλητών

Ο παραπάνω συντελεστής είναι ισοδύναμος με την κατασκευή ψευδομεταβλητών μια για κάθε επίπεδο κάθε ονομαστικής μεταβλητής και τον υπολογισμό πια του αντίστοιχου συντελεστή για δυαδικά δεδομένα.

Με την ίδια λογική κανείς μπορεί να χρησιμοποιήσει ψευδομεταβλητές σε συνδυασμό με κάποιο άλλο μέτρο ομοιότητας για δυαδικά δεδομένα όπως περιγράψαμε παραπάνω.

Μεταβλητές σε κλίμακα κατάταξης

Στην περίπτωση κατηγορικών μεταβλητών σε κλίμακα κατάταξης συνήθως αυτό που γίνεται είναι να θεωρήσουμε τις μεταβλητές ως συνεχείς και να χρησιμοποιήσουμε μια κατάλληλη απόσταση. Συνήθως τέτοιας μορφής δεδομένα χρησιμοποιούνται στη ψυχομετρία όπου ζητείται σε κάποιον να απαντήσεις σε ερωτήσεις αποδίδοντας βαθμό σε κάποια κλίμακα (πχ πλήρης συμφωνία – 0, πλήρης διαφωνία 9) . Σε τέτοιες περιπτώσεις πρέπει να φροντίζει κανείς να χρησιμοποιεί ίδια κλίμακα ώστε να μην υπάρχει πρόβλημα. Εναλλακτικά αυτό που μπορεί να γίνει είναι να μετασχηματίσουμε την κλίμακα να παίρνει τιμές στο διάστημα (0,1).

Μεταβλητές διαφόρων τύπων

Μέχρι τώρα ασχοληθήκαμε με τις αποστάσεις στην περίπτωση που όλες οι μεταβλητές μας άνηκαν στην ίδια κατηγορία. Προφανώς κάτι τέτοιο δεν είναι ρεαλιστικό και συνήθως με πραγματικά δεδομένα οι μεταβλητές μπορούν να αφορούν διαφορετικούς τύπους μεταβλητών, όπως για παράδειγμα ηλικία (συνεχής), φύλο (δυαδική), οικογενειακή κατάσταση (ονομαστική κατηγορική) κλπ. Στην περίπτωση αυτή μιλάμε για δεδομένα μεικτού τύπου (mixed mode variables).

Υπάρχουν αρκετοί διαφορετικοί τρόποι να αντιμετωπίσουμε τέτοιας μορφής δεδομένα. Θα μας απασχολήσουν τρεις διαφορετικές περιπτώσεις. Οι δύο πρώτες έχουν περιορισμένη χρήση καθώς στηρίζονται σε μάλλον απλές διαδικασίες που δεν είναι ρεαλιστικές. Η τρίτη μέθοδος είναι αυτή που χρησιμοποιείτε περισσότερο αν και, δυστυχώς, δεν προσφέρεται στα περισσότερα στατιστικά πακέτα και ο χρήστης πρέπει να καταφύγει σε πολύπλοκες διαδικασίες για να την υλοποιήσει

- Στην πρώτη περίπτωση προχωράμε στην ανάλυση χρησιμοποιώντας ομοειδής μεταβλητές, με αυτή την προσέγγιση κάνουμε ομαδοποιήσεις για κάθε τύπο μεταβλητών ξεχωριστά ελπίζοντας ότι οι ομάδες που θα βρούμε θα είναι περίπου οι ίδιες. Δυστυχώς στην πράξη δεν είναι απαραίτητο να συμβεί αυτό, συνήθως τα αποτελέσματα από μια τέτοια ανάλυση οδηγούν σε διαφορετικές κατατάξεις.

- Εναλλακτικά μπορεί κάποιος να ορίσει ψευδομεταβλητές για όλους του τύπους των δεδομένων με αποτέλεσμα να καταλήξει σε ένα σύνολο από δυαδικές μεταβλητές. Στην περίπτωση συνεχών μεταβλητών αυτό προϋποθέτει μια διακριτοποίηση τους, δηλαδή αρχικά τις μετατρέπουμε σε μικρότερα διακριτά διαστήματα και στη συνέχεια ορίζουμε ψευδομεταβλητές για τη διακριτοποιημένη μεταβλητή. Προφανώς αυτή η προσέγγιση μπορεί να οδηγήσει σε σοβαρό χάσιμο πληροφορίας κατά τη διάρκεια των μετασχηματισμών αυτών και επομένως πρέπει να είναι κανείς πολύ προσεκτικός.

Στην πράξη η πιο διαδεδομένη μεθοδολογία βασίζεται στον υπολογισμό μια απόστασης κατάλληλης για μικτής μορφής δεδομένα. Η απόσταση αυτή ορίστηκε από τον Gower το 1971 και έχει τη μορφή

$$s(x,y)=\frac{\sum_{i=1}^{p} w_i(x,y)\cdot s_i(x,y)}{\sum_{i=1}^{p} w_i(x,y)},$$

όπου $s_i(x,y)$ υποδηλώνει την ομοιότητα ανάμεσα στις παρατηρήσεις x και y για την i μεταβλητή. Η ιδέα είναι να χρησιμοποιήσει κανείς διαφορετικές αποστάσεις για κάθε τύπο μεταβλητής, όμως αυτές οι αποστάσεις να είναι στην ίδια κλίμακα ώστε να είναι συγκρίσιμες και να συνεισφέρουν το ίδιο στη συνολική απόσταση. Συγκεκριμένα

- Αν η μεταβλητή είναι κατηγορική (είτε ονομαστικής κλίμακας είτε σε κλίμακα κατάταξης) $s_i(x,y)=\begin{cases}1, \text{ for matches}\\ 0, \text{ otherwise}\end{cases}$

- Αν η μεταβλητή είναι συνεχής $s_i(x,y)=1-\frac{|x_i-y_i|}{R_i}$, όπου R_i είναι το εύρος της μεταβλητής. Ουσιαστικά, διαιρούμε με το εύρος ώστε να είμαστε σίγουροι πως η τιμή ανήκει στο διάστημα [0,1], κάτι το οποίο ισχύει για την περίπτωση των κατηγορικών μεταβλητών που είδαμε παραπάνω.

- Τα βάρη $w_i(x,y)$ παίρνουν την τιμή 0 ή 1 ανάλογα με το αν η σύγκριση ανάμεσα στις δύο παρατηρήσεις για τη συγκεκριμένη μεταβλητή έχει έννοια. Όταν η σύγκριση έχει νόημα η τιμή είναι 1 ενώ στην αντίθετη περίπτωση είναι 0. Περιπτώσεις με βάρος 0 έχουμε όταν μια εκ των δύο τιμών δεν υπάρχει (missing values) είτε όταν οι μεταβλητές είναι κατηγορικές και ασύμμετρες με κοινή απουσία του χαρακτηριστικού, η οποία δεν έχει ενδιαφέρον.

Παράδειγμα 9.1 (συν.)

Έστω πάλι τα δεδομένα του παραδείγματος 9.1 αλλά τώρα έχουμε και πληροφορίες σχετικές με δημογραφικά χαρακτηριστικά. Συγκεκριμένα τα δεδομένα έχουν τη μορφή

		ιστοσελίδα											
		1	2	3	4	5	6	7	8	9	10	Ηλικία	Πόλη
Χρήστης	Α	1	1	1	0	0	0	1	0	0	1	18	Αθήνα
	Β	0	1	1	0	1	1	0	0	0	0	21	Αθήνα
	Γ	0	0	0	0	0	1	1	0	0	0	23	Πάτρα
	Δ	0	1	1	0	0	1	0	0	0	0	41	Τρίπολη

Τώρα έχουν προστεθεί άλλες δύο μεταβλητές, μια συνεχής (η ηλικία) και μια σε ονομαστική κλίμακα (η πόλη). Χρησιμοποιώντας το δείκτη του Gower βρίσκουμε πως η απόσταση ανάμεσα στο χρήστη Α και το χρήστη Β είναι η εξής:

Για τη συνεχή μεταβλητή (και περιοριζόμενοι στις 4 παρατηρήσεις για τα δεδομένα μας) βρίσκουμε πως το εύρος είναι 23 κι επομένως

$$s_{11} = \frac{|18-21|}{23} = 0.130$$

Για την ονομαστική μεταβλητή επειδή και οι δύο είναι από την ίδια πόλη έχουμε $s_{12} = 1$

Ενώ για τις δυαδικές έχουμε 1 για τις 2 ιστοσελίδες που έχουν επισκεφτεί και οι δύο και 0 αλλού.

Επομένως η απόσταση θα είναι

$$s(A,B) = \frac{2+1+0.130}{12} = 0.275$$

Πριν κλείσουμε αυτή την ενότητα σχετικά με μέτρα απόστασης θα πρέπει να τονίσουμε πως η έννοια της απόστασης καθώς και ο πίνακας αποστάσεων χρησιμοποιούνται και σε πολλές άλλες στατιστικές τεχνικές όπως η πολυδιάστατη κλιμάκωση (multidimensional scaling) και η ανάλυση αντιστοιχιών (correspondence analysis).

9.3 Προβλήματα που πρέπει να αντιμετωπίσει ο ερευνητής

Πριν προχωρήσουμε στην περιγραφή των συγκεκριμένων μεθόδων ανάλυσης θα πρέπει να αναφερθούμε σε κάποια από τα προβλήματα που αφορούν όλες τις μεθόδους ανάλυσης κατά συστάδες. Σε οποιαδήποτε μέθοδο θα πρέπει να τονιστεί ότι δυστυχώς υπάρχουν πολλά σημεία στα οποία ο ερευνητής μπορεί να λειτουργήσει υποκειμενικά, με αποτέλεσμα από τα ίδια δεδομένα να εξαχθούν ακόμα και αντικρουόμενα αποτελέσματα. Από την άλλη μια γενική αλήθεια είναι πως όταν στα δεδομένα υπάρχουν πραγματικά ομοιογενείς ομάδες τότε οποιαδήποτε μέθοδος θα καταφέρει να τις αναγνωρίσει. Επομένως οι αντιφατικές λύσεις είναι μάλλον μια ένδειξη ότι δεν υπάρχει η κατάλληλη δομή στα δεδομένα μας, δηλαδή δεν υπάρχουν ομοιογενείς ομάδες.

Μερικά προβλήματα που πρέπει να απαντηθούν σχετικά με την ανάλυση είναι

- **Ποιες μεταβλητές πρέπει να χρησιμοποιηθούν**

Στην πραγματικότητα δεν υπάρχει κάποιος τρόπος για να οδηγήσει στην επιλογή μεταβλητών πριν την ανάλυση. Στην πράξη η επιλογή των μεταβλητών αυτών αν και είναι πολύ σημαντική για την ανάλυση συνήθως δεν αντιμετωπίζεται με τη δέουσα σοβαρότητα.

Αν, λοιπόν, δεν υπάρχει κάποια εμπειρία ή κάποιος θεωρητικός λόγος για να επιλέξουμε κάποιες συγκεκριμένες μεταβλητές για την ανάλυση καταφεύγουμε στη χρήση όλων των διαθέσιμων μεταβλητών. Εναλλακτικά θα μπορούσε κανείς να διαλέξει μόνο εκείνες τις μεταβλητές που πιστεύουμε για κάποιους λόγους ότι έχουν τη δυνατότητα να δημιουργήσουν ομοιογενείς ομάδες.

Αφού κάνουμε την ανάλυση μπορούμε εκ των υστέρων να δούμε αν κάποιες μεταβλητές τελικά ήταν αδιάφορες με την έννοια ότι η τιμή τους είναι η ίδια για όλες τις ομάδες που δημιουργήσαμε κι επομένως δεν έχουν καμιά διακριτική ικανότητα. Αν μάλιστα πιστεύουμε ότι δεν υπάρχει λόγος η μεταβλητή αυτή να παραμείνει στην ανάλυση μπορούμε να την αφαιρέσουμε και να ξαναομαδοποιήσουμε τα δεδομένα με τις υπόλοιπες μεταβλητές. Αν η μεταβλητή που αφαιρέσαμε πραγματικά ήταν άχρηστη τότε τα αποτελέσματα δεν πρόκειται να αλλάξουν.

Ένα άλλο σημαντικό πρόβλημα έχει να κάνει με τυχόν μετασχηματισμούς των δεδομένων. Για παράδειγμα είδαμε προηγουμένως πως για μερικές αποστάσεις οι διαφορές στις κλίμακες μπορεί να έχουν σημαντική επίδραση στον υπολογισμό των αποστάσεων και άρα στην όλη ανάλυση. Από την άλλη μεριά όμως η τυποποίηση των μεταβλητών ώστε να έχουν ίδια μεταβλητότητα μπορεί να οδηγήσει σε χάσιμο πληροφορίας αφού οι διαφορές στην κλίμακα μπορεί να είναι σημαντικές.

Ειδικά στην περίπτωση πολυμεταβλητών συνεχών δεδομένων, η ύπαρξη συσχετίσεων σχετίζεται με το σχήμα των δεδομένων και οι διαφορές ανάμεσα στις ομάδες μπορεί να

αφορούν αυτό ακριβώς το χαρακτηριστικό (φανταστείτε για παράδειγμα ομάδες καθεμία με περίπου ίδιο διάνυσμα μέσων αλλά διαφορετικό πίνακα διακυμάνσεων. Τα υπερελλειψοειδή που αντιστοιχούν μπορεί να έχουν διαφορετικά σχήματα και αυτό οφείλεται σε διαφορές στις συνδιακυμάνσεις. Επομένως τυχόν μετασχηματισμός μπορεί να ακυρώσει αυτές τις διαφορές).

- **Ποια απόσταση/ ομοιότητα να χρησιμοποιήσουμε**

Η επιλογή της απόστασης έχει να κάνει με τη μέθοδο που θα χρησιμοποιήσουμε αλλά και τον τύπο των δεδομένων μας. Επίσης είναι σημαντικό να γνωρίζουμε το σκοπό της ανάλυσης αλλά και κάποια επιμέρους χαρακτηριστικά. Συνεπώς το πρόβλημα της επιλογής είναι αρκετά πολύπλοκο.

Η πολυπλοκότητα του αυξάνει καθώς η ανάλυση σε συστάδες είναι μια μέθοδος που στηρίζεται στη χρήση υπολογιστών και πιο συγκεκριμένα στατιστικών πακέτων , συνεπώς στις παραπάνω παρατηρήσεις θα πρέπει να προσθέσουμε και τη διαθεσιμότητα της απόστασης αυτής από το συγκεκριμένο στατιστικό πακέτο. Για παράδειγμα η απόσταση του Gower που είδαμε για μεικτού τύπου δεδομένα δεν προσφέρεται αυτούσια από πολλά πακέτα και ο χρήστης θα πρέπει να μετασχηματίσει κατάλληλα τα δεδομένα του πριν προχωρήσει.

- **Πόσες ομάδες θα φτιάξουμε**

Η ανάλυση σε συστάδες σκοπό έχει να φτιάξει ομοιογενείς ομάδες. Πόσες όμως θα είναι αυτές; Οποιοσδήποτε λογικός αριθμός θα μπορούσε να χρησιμοποιηθεί αλλά κατά πόσο μια λύση με πολλές μικρές ομάδες θα βοηθούσε τους σκοπούς της ανάλυσης; Όπως θα δούμε αργότερα κάποιες από τις μεθόδους απαιτούν ο αριθμός των ομάδων να είναι γνωστός εκ των προτέρων. Πως επομένως θα βρούμε τον αριθμό των ομάδων; Θα δούμε αργότερα με ποιους τρόπους θα μπορούσαμε να προσδιορίσουμε αυτό τον αριθμό. Πρέπει να τονιστεί πως και πάλι ο τρόπος προσδιορισμού (ή καλύτερα εκτίμησης) εξαρτάται και από τη μορφή των δεδομένων.

- **Ποια μέθοδο να χρησιμοποιήσω**

Το τελευταίο ερώτημα έχει να κάνει με την επιλογή ανάμεσα στις μεθόδους που έχουμε διαθέσιμες. Γενικά οι ιεραρχικές μέθοδοι δεν είναι καλή ιδέα να χρησιμοποιούνται για μεγάλο πλήθος δεδομένων καθώς απαιτούν πολύ χρόνο και υπολογιστική ισχύ. Επίσης υπάρχει η τάση να δημιουργούνται ομάδες με ανομοιογενές μέγεθος. Από την άλλη η μέθοδος K-means ενώ αποφεύγει αυτά τα προβλήματα, δουλεύει ικανοποιητικά με μεγάλα δείγματα και δημιουργεί ομάδες παραπλήσιου μεγέθους, εξαρτάται πολύ από τις αρχικές τιμές που θα χρησιμοποιήσουμε.

Τα προβλήματα σχετικά με τη μέθοδο γίνονται πιο έντονα σε σύγχρονες εφαρμογές όπου το πλήθος των δεδομένων είναι πολύ μεγάλο και επομένως οι υπολογιστικές ανάγκες είναι τεράστιες.

Έχοντας όλα τα παραπάνω στο μυαλό μας ας δούμε κάποιες από τις μεθόδους που χρησιμοποιούνται περισσότερο στην πράξη.

9.4 Η μέθοδος K-Means

Ο αλγόριθμος K-means ανήκει σε μια μεγάλη κατηγορία αλγορίθμων ομαδοποίησης που είναι γνωστοί ως αλγόριθμοι διαμέρισης (partitioning algorithms). Ουσιαστικά οι αλγόριθμοι είναι έτσι φτιαγμένοι ώστε να διαμερίζουν το πολυεπίπεδο που δημιουργούν τα δεδομένα σε περιοχές και να αντιστοιχούν μια περιοχή σε κάθε ομάδα.

9.4.1 Ο αλγόριθμος

Η μέθοδος θεωρεί πως ο αριθμός των ομάδων που θα προκύψουν είναι γνωστός εκ των προτέρων. Αυτό αποτελεί έναν περιορισμό της μεθόδου καθώς είτε πρέπει να τρέξουμε τον αλγόριθμο με διαφορετικές επιλογές ως προς το πλήθος των ομάδων είτε πρέπει με κάποιον άλλο τρόπο να έχουμε καταλήξει στον αριθμό των ομάδων.

Η μέθοδος δουλεύει επαναληπτικά. Χρησιμοποιεί την έννοια του κέντρου της ομάδας (centroid) και στη συνέχεια κατατάσσει τις παρατηρήσεις ανάλογα με την απόσταση τους από τα κέντρα όλων των ομάδων. Το κέντρο της ομάδας δεν είναι τίποτα άλλο από τη μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας, δηλαδή αντιστοιχεί στο διάνυσμα των μέσων.

Στη συνέχεια για κάθε παρατήρηση υπολογίζουμε την ευκλείδεια απόστασή της από τα κέντρα των ομάδων που έχουμε και κατατάσσουμε κάθε παρατήρηση στην ομάδα που είναι πιο κοντά (για την ακρίβεια 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 μέχρι ότου δεν υπάρχουν διαφορές ανάμεσα σε δύο διαδοχικές επαναλήψεις

Συνήθως η απόσταση που χρησιμοποιείται για να κατατάξει τις παρατηρήσεις είναι η ευκλείδεια. Αν θέλουμε να χρησιμοποιήσουμε άλλη απόσταση θα πρέπει να κάνουμε ειδικούς μετασχηματισμούς στα δεδομένα πριν τη χρησιμοποιήσουμε.

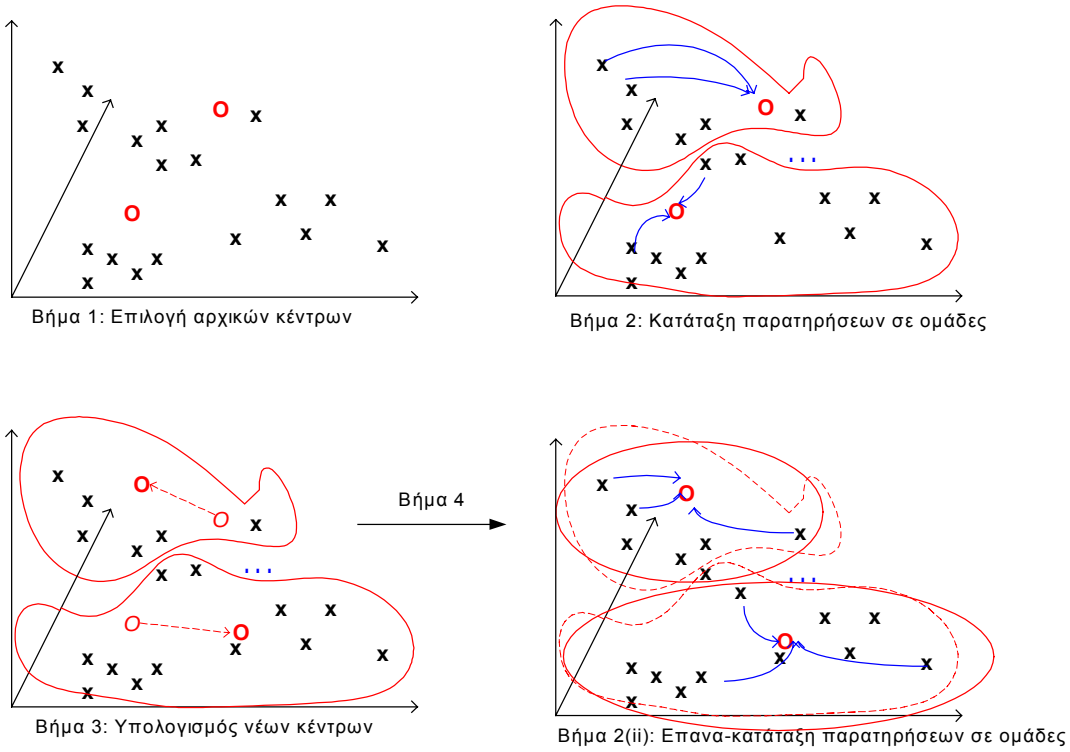
Όπως είπαμε και πριν ο αλγόριθμος αυτός δουλεύει ικανοποιητικά για μεγάλα σετ δεδομένων επειδή σε αυτή την περίπτωση δουλεύει πολύ πιο γρήγορα από την ιεραρχική

ομαδοποίηση. Αυτός είναι και ο λόγος που η μέθοδος μερικές φορές καλείται και γρήγορη ομαδοποίηση (Quick Clustering).

Αλγοριθμικά έχουμε ότι

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

Τα παρακάτω γραφήματα δείχνουν πως δουλεύει ο αλγόριθμος



Γράφημα 9.2 Περιγραφή της μεθόδου K-means

Οι παρατηρήσεις συμβολίζονται με 'x' και τα αρχικά κέντρα με 'o' στη εικόνα α. Για κάθε παρατήρηση μετράμε την απόσταση από κάθε ένα κέντρο και την κατατάσσουμε στην ομάδα με το πλησιέστερο κέντρο. Σχηματίζουμε δηλαδή τα δύο νέφη της εικόνας β.

Στη συνέχεια από όλες τις παρατηρήσεις που έχουμε κατατάξει υπολογίζουμε τα νέα κέντρα των ομάδων, τα διανύσματα των μέσων. Έτσι στη εικόνα γ βλέπουμε πως τα

κέντρα μετατοπίζονται. Με βάση αυτά τα νέα κέντρα ξεκινάμε από την αρχή κατατάσσοντας πάλι παρατηρήσεις κλπ.

Τα αρχικά κέντρα μπορούν είτε να οριστούν από το χρήστη είτε υπολογίζονται με κάποιο συγκεκριμένο αλγόριθμο. Για παράδειγμα ο αλγόριθμος που χρησιμοποιεί το SPSS είναι ο εξής και αφορά μια επανάληψη πριν την έναρξη του αλγορίθμου:

Αρχικά διάλεξε τις πρώτες *k* παρατηρήσεις ως τα αρχικά κέντρα. Στη συνέχεια για κάθε παρατήρηση κάνε το ακόλουθο:

Η παρατήρηση αντικαθιστά ένα από τα ήδη υπάρχοντα κέντρα, αν :

- η μικρότερη από τις αποστάσεις της από τα κέντρα που ήδη υπάρχουν είναι μεγαλύτερη από την απόσταση των δύο πιο κοντινών ήδη υπαρχόντων κέντρων. Δηλαδή έστω $c_j, j=1,...,k,$ τα υπάρχοντα κέντρα και $d(x,y)$ η απόσταση ανάμεσα στις παρατηρήσεις x και y. Τότε υπολογίζουμε για την i παρατήρηση τις αποστάσεις $d_j = d(x_i, c_j),\ j=1,...,k$. Ομοίως υπολογίζουμε τις αποστάσεις μεταξύ των κέντρων, δηλαδή τα $d_{ij} = d(c_i, c_j),\ \mathrm{i}, j=1,...,k,\ i \neq j$. Στη συνέχεια ελέγχουμε αν $\min_j(d_j) > \min_{i,j}(d_{ij})$. Αν ισχύει τότε το κέντρο που είναι κοντύτερα στην παρατήρηση αυτή αντικαθίσταται από την παρατήρηση.

- η παρατήρηση αντικαθιστά το κέντρο $c_j, j=1,...,k,$ αν $\min_j(d_j) > \min_i(d_{ij})$

αν δηλαδή η απόσταση της παρατήρησης από το κέντρο αυτό είναι μεγαλύτερη από τη μικρότερη απόσταση ανάμεσα στο συγκεκριμένο κέντρο και τα υπόλοιπα. Αυτός ο κανόνας λειτουργεί συμπληρωματικά με τον προηγούμενο.

Έτσι τα κέντρα που προκύπτουν όταν εξαντλήσουμε τις παρατηρήσεις αποτελούν τα αρχικά κέντρα για να ξεκινήσει ο αλγόριθμος.

Τα κριτήρια τερματισμού (βήμα 4) μπορούν να οριστούν από το χρήστη καθώς για μεγάλα σετ δεδομένων με πολύπλοκη δομή ο αλγόριθμος μπορεί να καθυστερήσει πολύ αν το κριτήριο τερματισμού είναι τόσο αυστηρό.

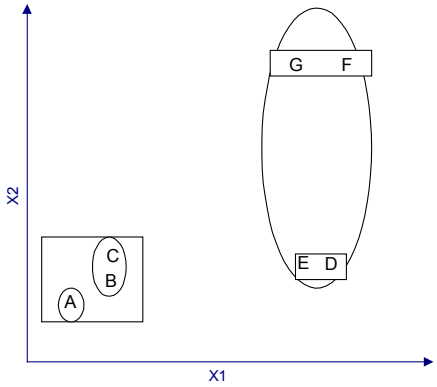
9.4.2 Χαρακτηριστικά του αλγορίθμου

Ας δούμε λίγο πιο αναλυτικά κάποια από τα χαρακτηριστικά του αλγορίθμου. Κατά αρχάς ο αλγόριθμος είναι ιδιαίτερα γρήγορος και στην πράξη σταματάμε συνήθως μετά από σχετικά λίγες επαναλήψεις. Αυτό τον κάνει ιδιαίτερα χρήσιμο για τις περιπτώσεις μεγάλων σετ δεδομένων. Επίσης δεν χρειάζεται να κρατά στη μνήμη πολλά στοιχεία και επομένως δεν χρειάζεται ιδιαίτερα μεγάλη υπολογιστική ισχύ.

Ο αλγόριθμος ουσιαστικά ελαχιστοποιεί το άθροισμα των τετραγωνικών αποστάσεων των παρατηρήσεων από τα κέντρα των ομάδων που ανήκουν. Συνήθως η λύση περιέχει ομάδες με περίπου όμοιο αριθμό παρατηρήσεων

Το μεγάλο μειονέκτημα του αλγορίθμου είναι ότι εξαρτάται από τις αρχικές τιμές οι οποίες αν δεν βρεθούν με καλό τρόπο μπορεί να οδηγήσουν σε ολότελα διαφορετική ομαδοποίηση. Για να το ξεπεράσουμε αυτό μια λύση είναι να τρέχουμε τον αλγόριθμο με διάφορες αρχικές τιμές ώστε να είμαστε σίγουροι πως δεν παγιδεύτηκε σε κάποια μη βέλτιστη λύση.

Το παράδειγμα που ακολουθεί εμφανίζει αυτό το πρόβλημα. Έστω πως έχουμε 7 παρατηρήσεις και θέλουμε να τις κατατάξουμε σε 3 ομάδες. Αν χρησιμοποιήσουμε ως αρχικές τιμές τα σημεία A,B,C τότε οι ομάδες που προκύπτουν συμβολίζονται με τις ελλείψεις. Οι ομάδες είναι {A} , {B,C}, {E,D,G,F}. Αν όμως ξεκινήσουμε από τα σημεία A, D και F τότε οι ομάδες που φτιάχνουμε συμβολίζονται με τα τετράγωνα και είναι {A,B,C}, {E,D}, {G,F}). Στην πρώτη περίπτωση οι διαφορές μέσα στις ομάδες είναι πολύ μεγαλύτερες από ότι στη δεύτερη.



Γράφημα 9.3. Ευαισθησία του αλγορίθμου k-means στην επιλογή αρχικών κέντρων

Ένα άλλο πρόβλημα έχει να κάνει με την επιλογή του αριθμού των ομάδων. Όπως είπαμε μια τακτική θα μπορούσε να είναι η ομαδοποίηση με διαφορετικό κάθε φορά αριθμό ομάδων και στο τέλος την επιλογή της ομάδας που είναι κατά κάποιον τρόπο βέλτιστη. Θα δούμε σε επόμενη ενότητα σφαιρικά το πρόβλημα αυτό.

Με βάση λοιπόν τα παραπάνω μερικές χρήσιμες στρατηγικές είναι οι ακόλουθες

- Η επιλογή των αρχικών κέντρων πρέπει να γίνεται έτσι ώστε αυτά να είναι όσο γίνεται πιο μακριά μεταξύ τους. Αυτό ουσιαστικά προσπαθεί να κάνει ο αλγόριθμος για την επιλογή αρχικών κέντρων που περιγράψαμε πριν από λίγο.

- Ένας τρόπος για να αποφύγουμε την υλοποίηση μεγάλου αριθμού διαφορετικών ομαδοποιήσεων είναι να μελετάμε τη λύση που ήδη έχουμε προσπαθώντας να ενώσουμε ή να διαλύσουμε ομάδες που θα μπορούσαν να μας βελτιώσουν τη μέση απόσταση των παρατηρήσεων από το κέντρο της ομάδας που ανήκουν. Για παράδειγμα αν δούμε μια

ομάδα με μερικές παρατηρήσεις πιο απομακρυσμένες θα ήταν μια καλή ιδέα να ξανατρέξουμε τον αλγόριθμο αυξάνοντας κατά ένα τον αριθμό των ομάδων.

- Είναι μάλλον σπάνιο να πετύχουμε την βέλτιστη λύση με μια μόνο επιλογή αριθμού ομάδων, συνεπώς θα πρέπει να δοκιμάσουμε διάφορες επιλογές και να χρησιμοποιήσουμε και τη διαίσθηση μας ώστε να πετύχουμε την καλύτερη ομαδοποίηση.

Τελειώνοντας θα πρέπει να παρατηρήσουμε πως η δυναμική του αλγορίθμου είναι πως με τις πρώτες λίγες επαναλήψεις πλησιάζει πολύ κοντά στην τελική λύση και στις υπόλοιπες επαναλήψεις οι όποιες διαφορές οφείλονται σε μετακινήσει κάποιων λίγων παρατηρήσεων που βρίσκονται ανάμεσα ουσιαστικά σε δύο ομάδες. Επομένως δεν είναι απαραίτητος ένας μεγάλος αριθμός επαναλήψεων καθώς η βασική δομή θα σχηματιστεί πολύ γρήγορα.

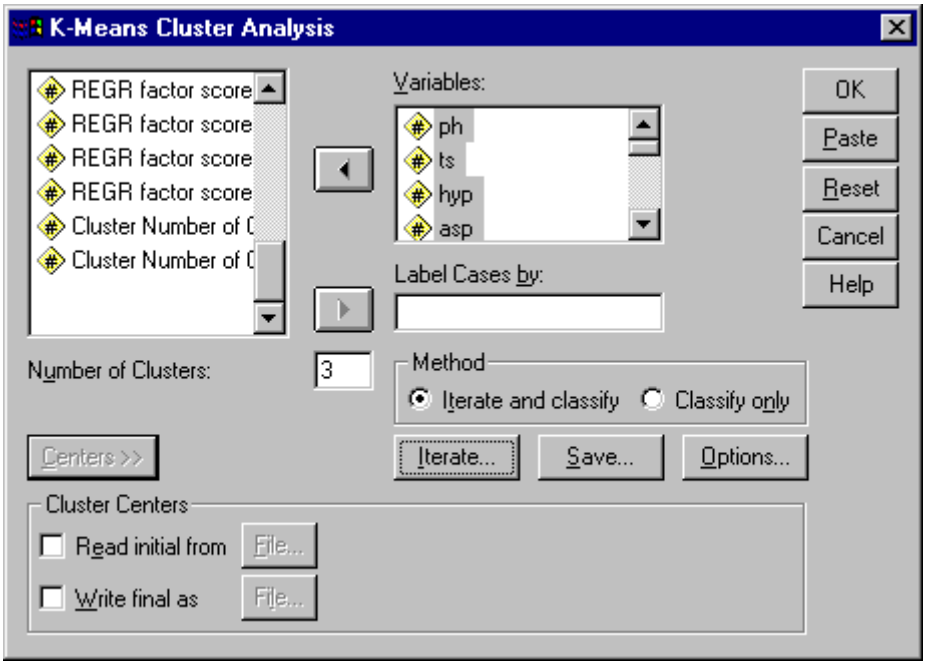
Επίσης ο αναγνώστης θα πρέπει να έχει υπόψη του πως συνήθως η μέθοδος βασίζεται στην ευκλείδεια απόσταση αλλά μπορεί να χρησιμοποιηθεί κάθε είδους απόσταση. Για μη συνεχή δεδομένα υπάρχει το πρόβλημα πως δεν είναι δυνατό να ορίσουμε το μέσο της ομάδας αλλά σε αυτή την περίπτωση μπορούμε να χρησιμοποιήσουμε αντίστοιχα μέτρα. Με το μέσο της ομάδας απλά θέλουμε να χρησιμοποιήσουμε κάποιο αντιπροσωπευτικό μέτρο για την ομάδα. Έτσι για παράδειγμα σε κατηγορικά δεδομένα 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 δεδομένα το κέντρο κάθε ομάδας μπορεί να αποτελείται από τις κορυφές των κατηγορικών μεταβλητών και τους μέσους των συνεχών.

9.4.3 K-means στο SPSS

Για να εκτελέσουμε μια K-Means ομαδοποίηση διαλέγουμε

Analyse > Classify > K-Means

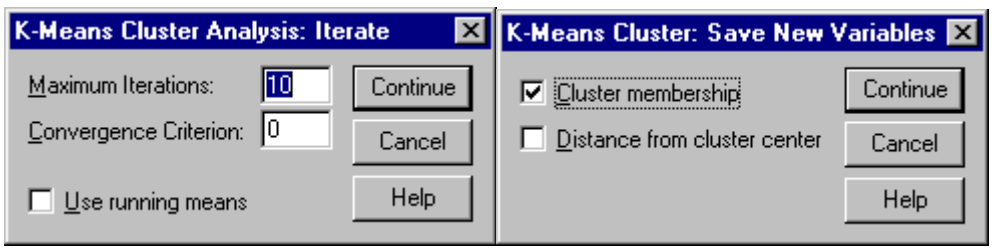
Έτσι, εμφανίζεται το παράθυρο της εικόνας 9.1 όπου πρέπει να διαλέξουμε τις μεταβλητές που θα χρησιμοποιήσουμε για ανάλυση. Στην εικόνα 9.1 έχουμε διαλέξει την επιλογή Centers η οποία ανοίγει και το κάτω μέρος του παραθύρου. Όπως βλέπεται μπορούμε να επιλέξουμε το πλήθος των ομάδων που επιθυμούμε (3 στην περίπτωση μας) και επίσης έχουμε τρεις επιλογές (Iterate, Save and Options) που ενεργοποιούν τα παράθυρα που βλέπουμε στις εικόνες, 9.2α, 9.2β και 9.3 αντίστοιχα. Τα αρχικά κέντρα των ομάδων μπορούν είτε να τεθούν από χρήστη είτε να βρεθούν από το πακέτο με βάση κάποιον αλγόριθμο.



Εικόνα 9.1. Το βασικό παράθυρο της K-means ομαδοποίησης

Στην εικόνα 9.2α επιλέγουμε τα κριτήρια τερματισμού του αλγορίθμου. Μπορούμε είτε να επιλέξουμε να σταματήσει έπειτα από συγκεκριμένο αριθμό επαναλήψεων (10 στην περίπτωση μας) είτε όταν η μεγαλύτερη απόσταση ανάμεσα σε διαδοχικά κέντρα όλων των ομάδων γίνει 0. Το 2ο κριτήριο αντιστοιχεί στην περίπτωση που οι ομάδες δεν αλλάξουν καθόλου μετά από μια επανάληψη. Την τιμή αυτή (0 στην περίπτωση μας) μπορούμε να την αλλάξουμε.

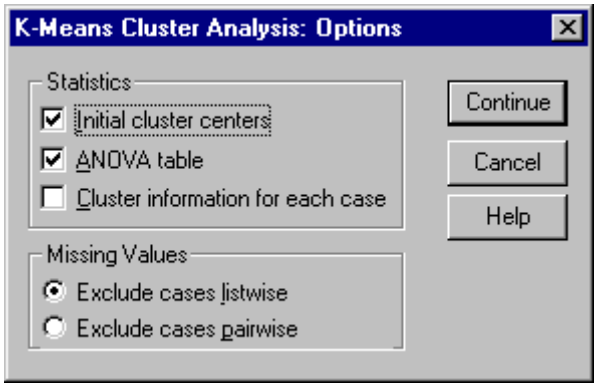
Στην εικόνα 9.2β επιλέγουμε τις καινούριες μεταβλητές που θέλουμε να δημιουργήσουμε. Έχουμε επιλέξει Cluster Membership που θα μας δημιουργήσει μια καινούρια στήλη όπου σε κάθε παρατήρηση θα δίνεται η τιμή της ομάδας που την κατατάξαμε. Αυτή η μεταβλητή είναι ιδιαίτερα χρήσιμη όταν μετά την ανάλυση προσπαθήσουμε να δούμε τα χαρακτηριστικά κάθε ομάδας αλλά και το αν κάποιες μεταβλητές προσφέρουν πληροφορία σχετικά με την ομαδοποίηση που κάναμε.



(α) (β)

Εικόνα 9.2. Τα παράθυρα για τα κριτήρια τερματισμού του επαναληπτικού αλγορίθμου και για τη δημιουργία καινούριων μεταβλητών με τα αποτελέσματα της ανάλυσης

Τέλος από το παράθυρο της εικόνας 9.3 μπορούμε να διαλέξουμε ποια αποτελέσματα θα εμφανιστούν στην οθόνη. Έχουμε επιλέξει να μας δώσει ο υπολογιστής τα αρχικά κέντρα (αυτό είναι χρήσιμο για να δούμε σε διαδοχικές εκτελέσεις του αλγορίθμου αν ξεκινώντας από διαφορετικά κέντρα πόσο αλλάζει το αποτέλεσμα) και να μας δώσει τους πίνακες ανάλυσης διακύμανσης για τις μεταβλητές που χρησιμοποιήσαμε ώστε να βρούμε ποιες περιέχουν όντως πληροφορία για την ομαδοποίηση που κάναμε.



Εικόνα 9.3. Διάφορες άλλες επιλογές

9.4.4 Εφαρμογή

Για να δούμε ένα παράδειγμα σε πραγματικά δεδομένα χρησιμοποιήσαμε 15 παρατηρήσεις που αφορούν παλιά βιβλία. Προκειμένου οι συντηρητές να διαπιστώσουν την κατάσταση στην οποία βρίσκονται τα βιβλία και αν χρειάζονται και τι είδους συντήρηση λαμβάνουν μια σειρά από μετρήσεις που αφορούν διάφορα αμινοξέα. Με βάση αυτές τις τιμές μπορεί κανείς να κατατάξει τα βιβλία σε διάφορες ομάδες. Για το συγκεκριμένο παράδειγμα μετρήθηκαν 23 αμινοξέα. Είναι ενδιαφέρον πως ο αριθμός των μεταβλητών μας είναι μεγαλύτερος από τον αριθμό των παρατηρήσεων κι επομένως πολλές στατιστικές τεχνικές θα αποτύγχαναν.

Για τα δεδομένα μας λοιπόν θέλουμε να κατατάξουμε τις παρατηρήσεις μας σε 3 ομάδες. Η επιλογή αυτού του αριθμού των ομάδων έγινε κυρίως γιατί στα δεδομένα υπάρχει μια τέτοιου είδους ομαδοποίηση σε 3 κατηγορίες αντικειμένων ανάλογα με την ανάγκη συντήρησης που έχουν. Αυτή η ομαδοποίηση προέκυψε από τον ερευνητή εμπειρικά. Έτσι με τη μέθοδο αυτή θα προσπαθήσουμε να επαληθεύσουμε την κατάταξη αυτή.

Οι μεταβλητές που χρησιμοποιήσαμε ήταν οι PH, TS, HYP, ASP, THR, SER, GLU, PRO, GLY , ALA, VAL, MET, ILE, LEU, TYR, PHE, HIS, HYL, LYS, ARG, ADA, BALA.

Οι πίνακες αποτελεσμάτων είναι οι ακόλουθοι

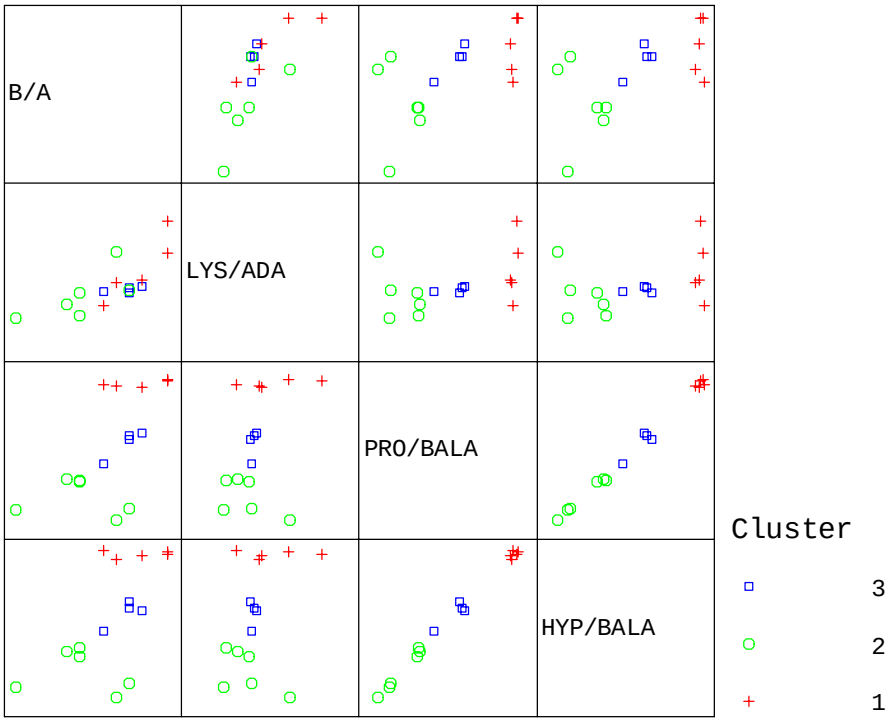
Initial Cluster Centers	Περιέχει τα αρχικά κέντρα των ομάδων, αυτά δηλαδή από όπου ξεκινά ο αλγόριθμος
Iteration History	Περιέχει πληροφορίες για το πως μετακινείται ο αλγόριθμος σε κάθε επανάληψη. Η τιμή που εμφανίζεται είναι η απόσταση ανάμεσα στο κέντρο της ομάδας στην τρέχουσα επανάληψη με το κέντρο της ομάδας κατά την προηγούμενη. Όταν η απόσταση αυτή μηδενιστεί σταματά ο αλγόριθμος.
Final Cluster Centers	Περιέχει τα κέντρα των ομάδων που βρέθηκαν αφού σταμάτησε ο αλγόριθμος, ο αλγόριθμος
ANOVA	Ο πίνακας περιέχει την ανάλυση διακύμανσης για το αν διαφέρουν οι μέσες τιμές ανάμεσα στις ομάδες. Μεταβλητές με καλή ικανότητα να ξεχωρίζουν τις παρατηρήσεις πρέπει να είναι στατιστικά σημαντικές. Πρέπει να ληφθεί υπόψη πως αυτές οι τιμές της στατιστικής σημαντικότητας έχουν μάλλον περιγραφικό σκοπό για να συγκρίνουμε μεταβλητές μεταξύ τους καθώς ο αλγόριθμος έχει κατάλληλα σχεδιαστεί να μεγιστοποιεί την ελεγχοσυνάρτηση F και επομένως η χρήση του είναι μάλλον ενδεικτική.
Number of Cases in each Cluster	Ο πίνακας παρουσιάζει πόσες παρατηρήσεις περιέχει κάθε ομάδα τελικά

Από τα αποτελέσματα παρατηρούμε πως μόνο η μεταβλητή TS φαίνεται να είναι διαφορετική ανάμεσα στις ομάδες. Αυτή είναι μια κακή ένδειξη ότι οι μεταβλητές μας δεν είναι πολύ καλές για το σκοπό που τις χρησιμοποιήσαμε, δεν μπορούν δηλαδή να ξεχωρίσουν ομάδες παρατηρήσεων. Επίσης παρατηρείστε πόσο λίγο έχουν αλλάξει τα κέντρα από την αρχική λύση στην τελική. Χρειάστηκαν μόνο2 επαναλήψεις. Αυτό δηλαδή που βλέπουμε είναι πως με αυτές τις μεταβλητές δεν είναι εύκολο να δημιουργήσουμε ομάδες. Αυτό είναι ένα πολύ κρίσιμο σημείο καθώς ο αλγόριθμος πάντα θα μας δώσει ομάδες παρατηρήσεων και επομένως ο ερευνητής πρέπει να αποφασίσει αν αυτές οι ομάδες περιέχουν κάποια πληροφορία ή όχι. Στην περίπτωση μας τα πράγματα είναι μάλλον άσχημα καθώς οι ομάδες δεν ξεχωρίζουν μεταξύ τους

Στη συνέχεια δοκιμάσαμε με κάποιες άλλες μεταβλητές και συγκεκριμένα με κάποιους συνδυασμούς των αρχικών μεταβλητών. Αυτοί ήταν B/A., LYS/ADA, PRO/BALA και HYP/BALA. Τα αποτελέσματα ήταν πολύ πιο ενθαρρυντικά. Οι ομάδες περιείχαν 5,4 και 6 παρατηρήσεις αλλά οι μεταβλητές είχαν μεγαλύτερη διακριτική ικανότητα. Το παρακάτω γράφημα μας δείχνει ένα πολλαπλό διάγραμμα σημείων ανά 2 μεταβλητές. Οι ομάδες παρουσιάζονται με διαφορετικά σύμβολα στο γράφημα. Παρατηρείστε πόσο καλά ξεχωρίζουν οι ομάδες.

Μεταβλητή	Αρχική Λύση			Τελική λύση			ANOVA	
	Cluster			Cluster			F	Sig.
	1	2	3	1	2	3		
PH	5.29	6.45	7.50	5.81	6.15	5.87	.182	.836
TS	60.00	72.00	85.00	61.67	73.80	83.29	63.955	.000
HYP	8.53	9.83	8.96	9.05	9.25	9.22	.122	.886
ASP	5.94	5.37	5.14	5.40	5.39	5.16	.482	.629
THR	2.26	2.21	1.93	2.02	2.00	2.01	.005	.995
SER	3.86	3.51	3.59	3.67	3.59	3.64	.381	.691
GLU	8.13	8.40	7.87	7.96	8.06	7.85	.977	.404
PRO	11.58	11.76	12.12	12.04	11.84	12.09	.326	.728
GLY	31.10	31.61	33.01	32.62	32.39	32.48	.032	.968
ALA	10.39	10.68	10.79	10.70	10.86	10.78	.607	.561
VAL	2.39	2.24	2.27	2.24	2.26	2.29	.061	.941
MET	.80	.66	.65	.69	.66	.67	.187	.832
ILE	1.61	1.35	1.39	1.36	1.39	1.37	.024	.976
LEU	3.00	2.93	2.79	2.70	2.80	2.77	.118	.890
TYR	.66	.32	.35	.40	.35	.33	.300	.746
PHE	1.56	1.33	1.40	1.36	1.38	1.37	.016	.984
HIS	.72	.50	.51	.51	.51	.53	.051	.950
HYL	.50	.41	.44	.46	.45	.47	.186	.832
LYS	2.16	1.86	1.96	2.02	1.89	1.98	.601	.564
ARG	4.38	4.60	4.46	4.44	4.52	4.60	1.716	.221
ADA	.14	.18	.11	.11	.15	.11	1.441	.275
BALA	.05	.05	.02	.04	.04	.05	.171	.845
Παρατηρήσεις στην ομάδα				3	5	7		

Πίνακας 9.6. Τα αποτελέσματα της K-means ομαδοποίησης



Γράφημα 9.4. Πολλαπλό διάγραμμα σημείων για τις μεταβλητές που χρησιμοποιήσαμε

Βλέπουμε λοιπόν πως η πρώτη ομάδα (+) έχουν μεγάλες τιμές για τη μεταβλητή HYP/BALA και PRO/BALA και B/A. Η ομάδα 3 ( ) είναι για όλες τις μεταβλητές σε μια μεσαία κατάσταση ενώ η ομάδα 2 (°) έχει σε όλες τις μεταβλητές τις μικρότερες τιμές.

9.5 Ιεραρχική ομαδοποίηση

Στην ιεραρχική ομαδοποίηση, ο αριθμός των ομάδων δεν είναι γνωστός από πριν. Οι μέθοδοι λειτουργούν ιεραρχικά με την έννοια ότι ξεκινούν χρησιμοποιώντας κάθε παρατήρηση σαν μια ομάδα και σε κάθε βήμα ενώνουν σε ομάδες τις παρατηρήσεις που βρίσκονται πιο κοντά. Στην πραγματικότητα οι ιεραρχικοί αλγόριθμοι δουλεύουν είτε προς τα εμπρός είτε προς τα πίσω. Δηλαδή:

- Κάποιοι αλγόριθμοι ξεκινούν με όλες τις παρατηρήσεις σε μια ομάδα. Η παρατήρηση που βρίσκεται πιο μακριά από τις υπόλοιπες (αυτό μπορεί να οριστεί με διάφορους τρόπους) φεύγει από τη μεγάλη ομάδα και σχηματίζει μια καινούρια ομάδα μόνη της. Στη συνέχεια βρίσκουμε τη δεύτερη πιο απομακρυσμένη και τη διώχνουμε, αυτή μπορεί είτε να σχηματίσει μια ομάδα μόνη της ή να πάει στην ομάδα που είχαμε στείλει την προηγούμενη κι έτσι προχωράμε μέχρι να μετακινήσουμε όλες τις παρατηρήσεις. Αυτοί οι αλγόριθμοι συνήθως ονομάζονται divisive.

- Πιο διαδεδομένοι είναι οι αντίστροφοι αλγόριθμοι γνωστοί ως agglomerative. Οι αλγόριθμοι ξεκινούν με κάθε παρατήρηση ως μια ομάδα και ενώνουν στη συνέχεια ομάδες που είναι πιο κοντινές. Θα δούμε αυτούς τους αλγορίθμους στη συνέχεια με μεγαλύτερη λεπτομέρεια

Οι ιεραρχικές μέθοδοι, επειδή σε κάθε βήμα χρησιμοποιούν έναν πίνακα αποστάσεων (δηλαδή τις αποστάσεις όλων των παρατηρήσεων από τις υπόλοιπες) χρειάζονται πολύ χρόνο και χώρο στον υπολογιστή και για αυτό είναι ασύμφορες για μεγάλα σετ δεδομένων.

9.5.1 Ο αλγόριθμος

Θα μπορούσαμε να περιγράψουμε τον αλγόριθμο για τις agglomerative μεθόδους ως εξής. Για να διευκολύνουμε την περιγραφή θα χρησιμοποιούμε τη λέξη ομάδα αν και πρέπει κανείς να έχει στο νου του πως πολλές ομάδες αποτελούνται από μια μόνο παρατήρηση και επομένως σε αυτή την περίπτωση η ομάδα ταυτίζεται με την παρατήρηση. Επομένως έχουμε

- **Βήμα 1.** Δημιούργησε τον πίνακα αποστάσεων για όλες τις ομάδες
- **Βήμα 2.** Βρες τη μικρότερη απόσταση και ένωσε τις δύο παρατηρήσεις με τη μικρότερη απόσταση. Δηλαδή δημιουργούμε μια ομάδα με τις παρατηρήσεις που είναι πιο κοντά. Αν η μικρότερη απόσταση αφορά μια ήδη δημιουργηθείσα ομάδα και μια παρατήρηση απλά βάζουμε αυτή την παρατήρηση σε αυτή την ομάδα ή αν αφορά 2 ομάδες που ήδη υπάρχουν τις ενώνουμε.
- **Βήμα 3.** Αν δεν έχουν όλες οι παρατηρήσεις μπει σε μια ομάδα πήγαινε στο βήμα 1 αλλιώς σταμάτα.

Υπάρχουν μερικά σημαντικά σημεία όπου ο ερευνητής πρέπει να αποφασίσει

Κατ' αρχάς πρέπει να αποφασιστεί το είδος της απόστασης που θα χρησιμοποιηθεί. Μιλήσαμε σε προηγούμενη ενότητα για την επιλογή της κατάλληλης απόστασης συνεπώς θεωρούμε πως ο ερευνητής έχει διαλέξει την απόσταση με την οποία θα ομαδοποιήσει τα δεδομένα

9.5.2 Επιλογή μεθόδου

Ένα άλλο σημαντικό σημείο για τον αλγόριθμό είναι πως θα υπολογίσουμε την απόσταση της ομάδας που φτιάξανε (είτε από συγχώνευση άλλων ομάδων είτε από συγχώνευση παρατηρήσεων). Υπάρχουν πολλές μέθοδοι μερικές από τις οποίες είναι οι

- **Nearest Neighbour**: Η μέθοδος του κοντινότερου γείτονα(nearest neighbor or single linkage) υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως τη μικρότερη απόσταση από μια παρατήρηση μέσα στην μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Η μέθοδος έχει κάποιες χρήσιμες μαθηματικές ιδιότητες αλλά παράγει ομάδες που δεν είναι συμπαγείς και συνήθως δημιουργεί μερικές πολύ μεγάλες ομάδες και κάποιες πάλι πολύ μικρές.

- **Furthest neighbour**: Σε αυτή τη περίπτωση συμβαίνει το αντίθετο, δηλαδή, η μέθοδος του μακρύτερου γείτονα (furthest neighbor or complete linkage) υπολογίζει την απόσταση ανάμεσα σε δύο ομάδες ως τη μεγαλύτερη απόσταση από μια παρατήρηση μέσα στην μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Οι ομάδες που δημιουργούνται είναι συνήθως συμπαγείς αλλά αποτυγχάνει να δημιουργήσει κάποιες μικρές μα πολύ συμπαγείς ομάδες

- **Average between groups**: Σε αυτή την περίπτωση η απόσταση είναι ο μέσος της απόσταση ανάμεσα σε όλες τις αποστάσεις της μιας ομάδας με τα στοιχεία της άλλης. Αν για παράδειγμα η μια ομάδα περιλαμβάνει της παρατηρήσεις {1,2} και η άλλη τις παρατηρήσεις {3,4,5} τότε η απόσταση είναι ο μέσος των αποστάσεων d(1,3), d(1,4), d(1,5), d(2,3) , d(2,4), d(2,5).

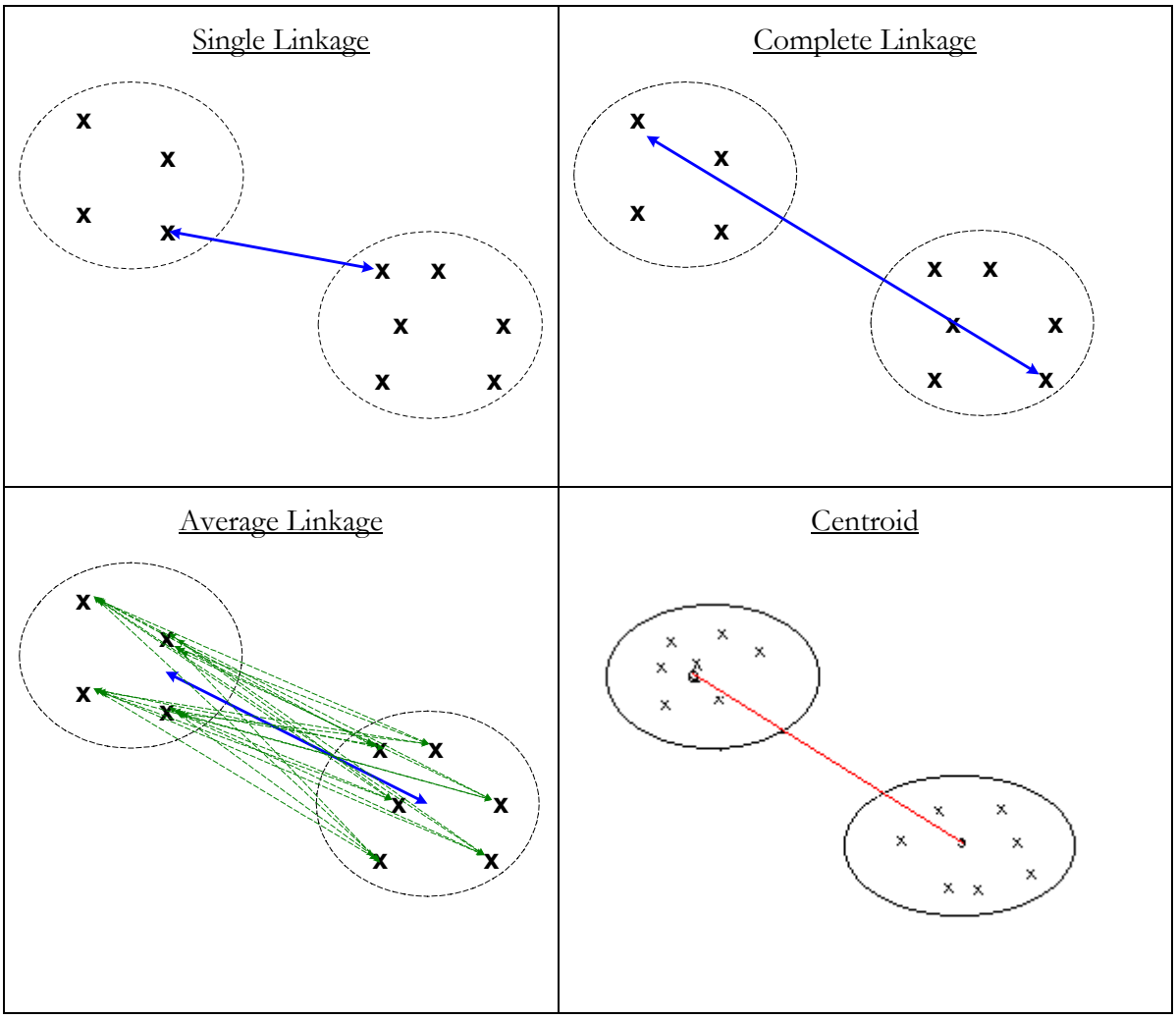
- **Average within groups**: Στην περίπτωση αυτή η απόσταση είναι ο μέσος όλων των αποστάσεων που προκύπτουν όταν ενώσουμε τις δύο ομάδες. Δηλαδή στην περίπτωση των ομάδων που είχαμε πριν η νέα απόσταση θα είναι ο μέσος των αποστάσεων d(1,2), d(1,3), d(1,4), d(1,5), d(2,3) , d(2,4), d(2,5), d(3,4), d(3,5), d(4,5)

- **Centroid**: Η απόσταση υπολογίζεται ως η απόσταση των κέντρων των ομάδων. Η μέθοδος αυτή έχει μερικές καλές ιδιότητες και παράγει συνήθως ομάδες συμπαγείς και ελλειπτικές

- **Ward method**: Η μέθοδος του Ward διαφέρει από τις υπόλοιπες και είναι σχεδιασμένη να ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες. Για κάθε παρατήρηση μπορούμε να υπολογίσουμε την απόσταση της (συνήθως ευκλείδεια) από το κέντρο της ομάδας. Αν αθροίσουμε για όλες τις ομάδες έχουμε μια τιμή που είναι το συνολικό άθροισμα. Αρχικά αυτό το άθροισμα είναι 0, αφού κάθε παρατήρηση είναι και μια ομάδα άρα η απόσταση από το κέντρο της είναι 0. Σε κάθε βήμα ενώνουμε τις ομάδες οι οποίες αν ενωθούν οδηγούν στη μικρότερη αύξηση του συνολικού αθροίσματος αποστάσεων. Η

μέθοδος έχει μερικές πολύ καλές ιδιότητες και συνήθως δημιουργεί ομάδες με παρόμοιο αριθμό παρατηρήσεων. Για αυτό και πολύ συχνά χρησιμοποιείται στην πράξη.

Στο γράφημα 9.5 μπορεί κανείς να δει πως λειτουργούν και υπολογίζουν την απόσταση διάφορες μέθοδοι.



Γράφημα 9.5. Διαγραμματική απεικόνιση μεθόδων υπολογισμού αποστάσεων μεταξύ ομάδων.

Θα πρέπει κανείς να παρατηρήσει πως για όλες τις μεθόδους χρειαζόμαστε απλά έναν πίνακα αποστάσεων με βάση τον οποίο υπολογίζουμε σε κάθε βήμα τις καινούριες αποστάσεις. Αυτό δεν ισχύει για τη centroid μέθοδο όπου χρειαζόμαστε το κέντρο της ομάδας. Αν τα στοιχεία μας δεν είναι συνεχή το κέντρο δεν μπορεί να είναι απλά οι μέσοι των μεταβλητών και σε αυτή την περίπτωση χρησιμοποιούμε την κορυφή ή τη διάμεσο. Θα πρέπει να τονιστεί πως για μερικές μορφές δεδομένων ούτε αυτό είναι δυνατό και σε αυτή την περίπτωση καλό είναι να αποφεύγουμε τη μέθοδο αυτή.

Παρόμοια προβλήματα σχετικά με τον υπολογισμό των κέντρων υπάρχουν και στη μέθοδο του Ward

Συγκρίνοντας τις μεθόδους μεταξύ τους θα πρέπει να γνωρίζουμε πως από πειράματα προσομοίωσης οι μέθοδοι με την καλύτερη επίδοση είναι η μέθοδος του Ward και η average linkage. Η μέθοδος του κοντινότερου γείτονα είναι αυτή με τη χειρότερη επίδοση. Παρόλα αυτά σε πολλά προβλήματα δεν είναι ξεκάθαρο ποια μέθοδος είναι προτιμότερη. Αυτό που θα πρέπει να έχει πάντα 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 έχει κατά νου πως κάθε μέθοδος δουλεύει καλύτερα με συγκεκριμένη μορφή δεδομένων

Για παράδειγμα μέθοδοι που βασίζονται σε τετραγωνικές αποστάσεις, όπως ο αλγόριθμος K-means και η μέθοδος του Ward τείνουν να βρίσκουν ομάδες με περίπου ίδια διακύμανση. Γενικά οι περισσότερες μέθοδοι αποτυγχάνουν να βρουν ομάδες με περίεργα σχήματα. Σε αυτή την περίπτωση η μέθοδος του κοντινότερου γείτονα μπορεί αν είναι πιο αποδοτική.

Τέλος ένα ακόμα σημείο είναι πως για μερικές από τις μεθόδους δεν είναι απαραίτητο να υπάρχει μια αύξηση σε κάθε βήμα της απόστασης, δηλαδή η απόσταση των ομάδων που συγχωνεύεται δεν πρέπει απαραίτητα να είναι αύξουσα.

9.5.3 Παράδειγμα και σύγκριση των μεθόδων

Για να δούμε πως δουλεύουν οι μέθοδοι που αναφέραμε ας χρησιμοποιήσουμε τα δεδομένα που υπάρχουν στον πίνακα 9.7. Αφορούν 5 χώρες και τα αντίστοιχα ποσοστά γεννήσεων, θανάτων και βρεφικής θνησιμότητας, αποτελούν Δε μέρος ενός μεγαλύτερου σετ δεδομένων. Για λόγους παρουσίασης θα χρησιμοποιήσουμε μόνο τις 5 παρατηρήσεις.

Γεννήσεις ανά 1000 κατοίκους	Θάνατοι ανά 1000 κατοίκους	Θάνατοι βρεφών σε 1000 γεννήσεις	Χώρα
24.7	5.7	30.8	Αλβανία
12.5	11.9	14.4	Βουλγαρία
11.6	13.4	14.8	Ουγγαρία
14.3	10.2	16	Πολωνία
13.6	10.7	26.9	Ρουμανία

Πίνακας 9.7

Χρησιμοποιώντας λοιπόν αυτές τις 3 μεταβλητές και ευκλείδεια απόσταση καταλήγουμε στον πίνακα αποστάσεων που είναι ο εξής:

	Αλβανία	Βουλγαρία	Ουγγαρία	Πολωνία	Ρουμανία
Αλβανία	0.0000				
Βουλγαρία	21.356	0.0000			
Ουγγαρία	22.066	1.794	0.0000		
Πολωνία	18.640	2.948	4.355	0.0000	
Ρουμανία	12.784	12.6066	12.558	10.934	0.0000

Πίνακας 9.8. Ο πίνακας αποστάσεων για τις παρατηρήσεις

Βλέπουμε λοιπόν πως οι 2 πιο κοντινές παρατηρήσεις είναι οι Βουλγαρία και η Ουγγαρία, άρα αυτές θα ενωθούν και θα αποτελέσουν μια ομάδα. Το ζητούμενο είναι πως θα υπολογίσουμε την απόσταση κάθε παρατήρησης από τις υπόλοιπες με την ομάδα αυτή.

Σύμφωνα με τη μέθοδο του κοντινότερου γείτονα η απόσταση θα είναι η μικρότερη από τις αποστάσεις των στοιχείων της ομάδας με κάθε παρατήρηση. Έτσι για την Αλβανία έχουμε πως

d(Αλβανία, Βουλγαρία) = 21.356 και

d(Αλβανία, Ουγγαρία) = 22.066

επομένως η απόσταση της Αλβανίας από αυτή την ομάδα θα είναι

d(Αλβανία, { Βουλγαρία, Ουγγαρία}) = 21.356.

Με τον ίδιο τρόπο βρίσκουμε και τα υπόλοιπα στοιχεία του πίνακα και επομένως μετά από ένα βήμα ο πίνακας αποστάσεων είναι ο

	Αλβανία	{Βουλγαρία, Ουγγαρία }	Πολωνία	Ρουμανία
Αλβανία	0.0000			
{Βουλγαρία, Ουγγαρία }	21.356	0.0000		
Πολωνία	18.640	2.948	0.0000	
Ρουμανία	12.784	12.5579	10.934	0.0000

Πίνακας 9.9. Ο πίνακας αποστάσεων για τις παρατηρήσεις με τη χρήση της μεθόδου του κοντινότερου γείτονα

Στην περίπτωση του μακρύτερου γείτονα θα βρίσκαμε πως

d(Αλβανία, { Βουλγαρία, Ουγγαρία}) = 22.066

και ο πίνακας θα γινόταν

	Αλβανία	{Βουλγαρία, Ουγγαρία }	Πολωνία	Ρουμανία
Αλβανία	0.0000			
{Βουλγαρία, Ουγγαρία }	22.066	0.0000		
Πολωνία	18.640	4.3555	0.0000	
Ρουμανία	12.784	12.6056	10.934	0.0000

Πίνακας 9.10. Ο πίνακας αποστάσεων για τις παρατηρήσεις με τη χρήση της μεθόδου του μακρύτερου γείτονα

Παρατηρείστε πως οι πίνακες πια διαφέρουν κι επομένως είναι πολύ πιθανό να οδηγήσουν σε διαφορετική ομαδοποίηση.

Αντίστοιχα για την περίπτωση τόσο των μεθόδων Average between groups και Average within groups οι καινούριες αποστάσεις θα είναι οι ίδιες καθώς η μια ομάδα περιέχει μια παρατήρηση και θα είναι

d(Αλβανία, { Βουλγαρία, Ουγγαρία}) =

(d(Αλβανία, Βουλγαρία) + d(Αλβανία, Ουγγαρία)) /2 =

(21.356 + 22.066)/2 = 21.712

Ο πίνακας θα γίνει τώρα

	Αλβανία	{Βουλγαρία, Ουγγαρία }	Πολωνία	Ρουμανία
Αλβανία	0.0000			
{Βουλγαρία, Ουγγαρία }	21.712	0.0000		
Πολωνία	18.640	3.652	0.0000	
Ρουμανία	12.784	12.582	10.934	0.0000

Πίνακας 9.11. Ο πίνακας αποστάσεων για τις παρατηρήσεις με τη χρήση των μεθόδων Average between groups και Average within groups

Για τη μέθοδο centroid πρέπει να υπολογίσουμε το μέσο της ομάδας που μόλις κατασκευάσαμε. Αυτός θα είναι μια παρατήρηση με τιμές

Γεννήσεις ανά 1000 κατοίκους	Θάνατοι ανά 1000 κατοίκους	Θάνατοι βρεφών σε 1000 γεννήσεις
12.05	12.65	14.6

Και επομένως ο πίνακας αποστάσεων θα γίνει

	Αλβανία	{Βουλγαρία, Ουγγαρία }	Πολωνία	Ρουμανία
Αλβανία	0.0000			
{Βουλγαρία, Ουγγαρία }	21.697	0.0000		
Πολωνία	18.640	3.609	0.0000	
Ρουμανία	12.784	12.549	10.934	0.0000

Πίνακας 9.12. Ο πίνακας αποστάσεων για τις παρατηρήσεις με τη χρήση της μεθόδου centroid

Η μέθοδος του Ward απαιτεί πιο πολύπλοκους υπολογισμούς και για αυτό δεν θα δώσουμε λεπτομέρειες. Επαναλαμβάνοντας τη διαδικασία από κάθε πίνακα καταλήγουμε σε

μια ομαδοποίηση. Θα περιγράψουμε τα βήματα της μεθόδου single linkage που είναι πιο απλή αλλά προφανώς θα μπορούσε κανείς να δει με ίδιο τρόπο τα βήματα και των υπολοίπων μεθόδων.

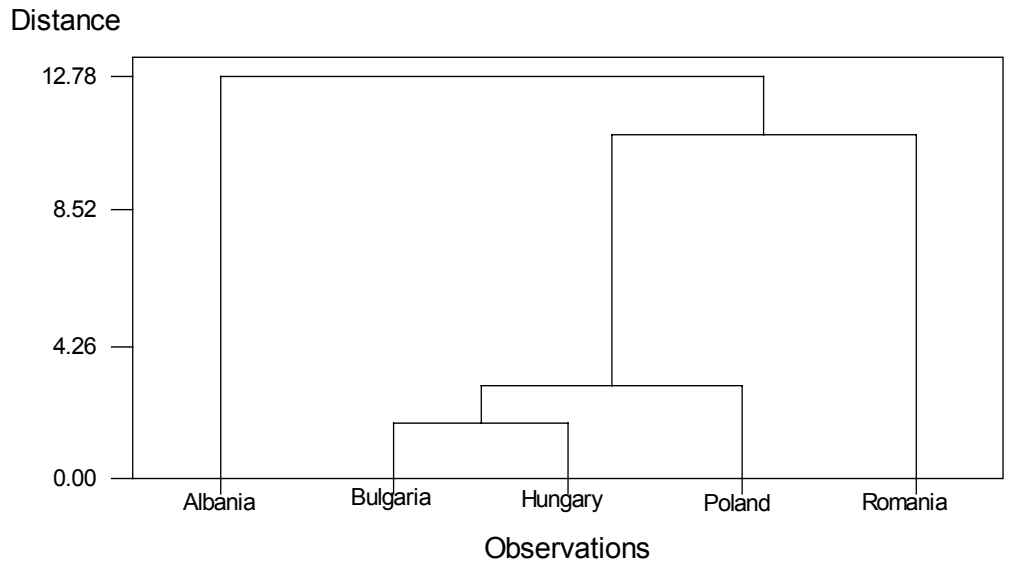
Με βάση τις αποστάσεις του πίνακα 9.9 θα συγχωνεύσουμε στην ομάδα {Βουλγαρία, Ουγγαρία} τη Πολωνία και θα πάρουμε

	Αλβανία	{Βουλγαρία, Ουγγαρία, Πολωνία}	Ρουμανία
Αλβανία	0.0000		
{Βουλγαρία, Ουγγαρία, Πολωνία}	18.640	0.0000	
Ρουμανία	12.784	10.934	0.0000

Επομένως στο επόμενο βήμα θα ενώσουμε την υπάρχουσα ομάδα με τη Ρουμανία και θα προκύψει ο παρακάτω πίνακας

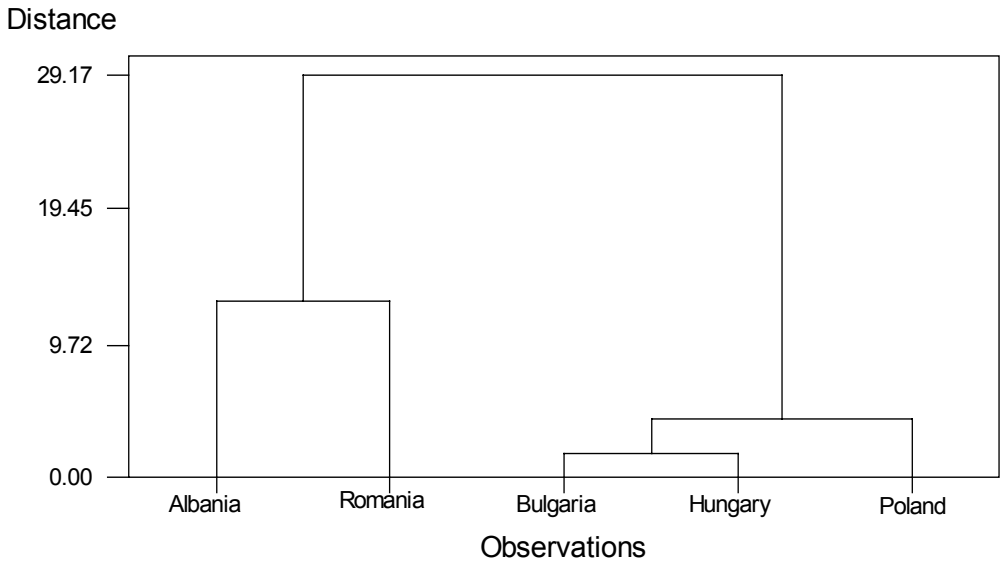
	Αλβανία	{Βουλγαρία, Ουγγαρία, Πολωνία, Ρουμανία }
Αλβανία	0.0000	
{Βουλγαρία, Ουγγαρία, Πολωνία, Ρουμανία }	12.784	0.0000

Μπορούμε να αναπαραστήσουμε όλη την ιστορία της ιεραρχικής ομαδοποίησης με ένα δενδρόγραμμα ως εξής



Γράφημα 9.6. Δενδρόγραμμα για τα δεδομένα μας με τη χρήση της μεθόδου του κοντινότερου γείτονα

Το δενδρόγραμμα ενώνει τις παρατηρήσεις με μια γραμμή. Αυτό επαναλαμβάνεται σε κάθε βήμα με αποτέλεσμα στο τέλος όλες οι παρατηρήσεις να είναι ενωμένες με κάποιο μονοπάτι. Συνήθως στον έναν άξονα έχουμε τις παρατηρήσεις και στον άλλο την τιμή της απόστασης με την οποία ενώσαμε τις παρατηρήσεις (ομάδες) ώστε να έχουμε μια ένδειξη πως προχώρησε η διαδικασία. Όπως θα δούμε αργότερα αυτό είναι χρήσιμο και για την εύρεση του βέλτιστου αριθμού ομάδων.



Γράφημα 9.7. Δενδρόγραμμα για τα δεδομένα μας με τη χρήση της μεθόδου του Ward

Για το συγκεκριμένο παράδειγμα και άλλες μέθοδοι καταλήγουν στο ίδιο δενδρόγραμμα (εκτός ίσως από αλλαγές στο επίπεδο της απόστασης όπου γίνονται οι συγχωνεύσεις ομάδων) εκτός από τη μέθοδο του Ward. Για τη μέθοδο αυτή βρίσκουμε το δενδρόγραμμα του γραφήματος 9.7. Σύμφωνα λοιπόν με αυτή τη μέθοδο βρίσκουμε διαφορετικές ομάδες. Τώρα πια η Αλβανία και η Ρουμανία σχηματίζουν μια ομάδα κάτι που πριν δεν είχε συμβεί.

Το δενδρόγραμμα αποτελεί ένα πολύτιμο οπτικό εργαλείο για την ιεραρχική ομαδοποίηση καθώς αφενός περιέχει ολόκληρη την ιστορία της ομαδοποίησης, βοηθάει στην επιλογή της λύσης που τελικά θα κρατήσουμε αλλά εμφανίζει και τη δυναμική της μεθόδου. Για παράδειγμα αυτό που είδαμε στην περίπτωση του κοντινότερου γείτονα είναι η χαρακτηριστική περίπτωση αυτής της μεθόδου όπου συνήθως οι παρατηρήσεις ενσωματώνονται μια-μια σε μια μεγάλη ομάδα.

9.5.4 Χαρακτηριστικά του αλγόριθμου

Ένα από τα μειονεκτήματα της ιεραρχικής ομαδοποίησης είναι ότι δεν συμφέρει από άποψη υπολογιστικού φόρτου για μεγάλα σετ δεδομένων. Δεδομένου πως πρέπει κανείς να σχηματίσει έναν πίνακα αποστάσεων, ακόμα και αν η απόσταση είναι συμμετρική χρειάζεται κανείς να υπολογίσει και να αποθηκεύσει $n(n-1)/2$ αποστάσεις. Ακόμα και για 100 παρατηρήσεις ($n = 100$) αυτό σημαίνει πως αρχικά θα αποθηκευτούν 4950 αποστάσεις. Σε κάθε βήμα αυτός ο πίνακας θα πρέπει να ανανεώνεται. Επομένως θα πρέπει συνέχεια να διαβάζουμε και να γράφουμε στη μνήμη του υπολογιστή και άρα το υπολογιστικό κόστος είναι πολύ μεγάλο

Επίσης η μέθοδος έχει το μεγάλο μειονέκτημα, πως ομάδες που φτιάχνονται σε αρχικά βήματα δεν μπορούν να χωρίσουν και επομένως οι παρατηρήσεις που ενώνονται σε αρχικά βήματα μένουν μαζί για πάντα. Γενικά η δυναμική του αλγόριθμου εξαρτάται από τον τρόπο που υπολογίζουμε την απόσταση ανάμεσα σε ομάδες. Πολύ συχνά ο αλγόριθμος καταφέρνει να δημιουργεί μερικές ομάδες με πολλές παρατηρήσεις και αφήνει κάποιες παρατηρήσεις να είναι μόνες τους μια ομάδα.

Ένα επίσης σημαντικό πρόβλημα είναι πως ουσιαστικά η μέθοδος μας εφοδιάζει με μια ποικιλία λύσεων, μια για κάθε διαφορετικό αριθμό ομάδων. Συνεπώς χρειαζόμαστε ένα κριτήριο για να διαλέξουμε τη λύση που θα κρατήσουμε, με άλλα λόγια πόσες ομάδες θα αποτελούν την τελική λύση.

9.5.5 Εφαρμογή

Ας δούμε όμως τη μέθοδο στην πράξη. Τα δεδομένα είναι τα ίδια με αυτά που είχαμε πριν και αφορούν τα βιβλία και τις μετρήσεις πάνω σε αυτά Θα χρησιμοποιήσουμε τις 4 μεταβλητές που είδαμε προηγουμένως πως έχουν κάποια καλή διακριτική ικανότητα.

Μεταβλητή	Μέση Τιμή	Τυπική Απόκλιση
B/A	.527	.032
ARG/ORN	25.681	7.988
LYS/ADA	17.262	5.843
PRO/BALA	374.937	202.797

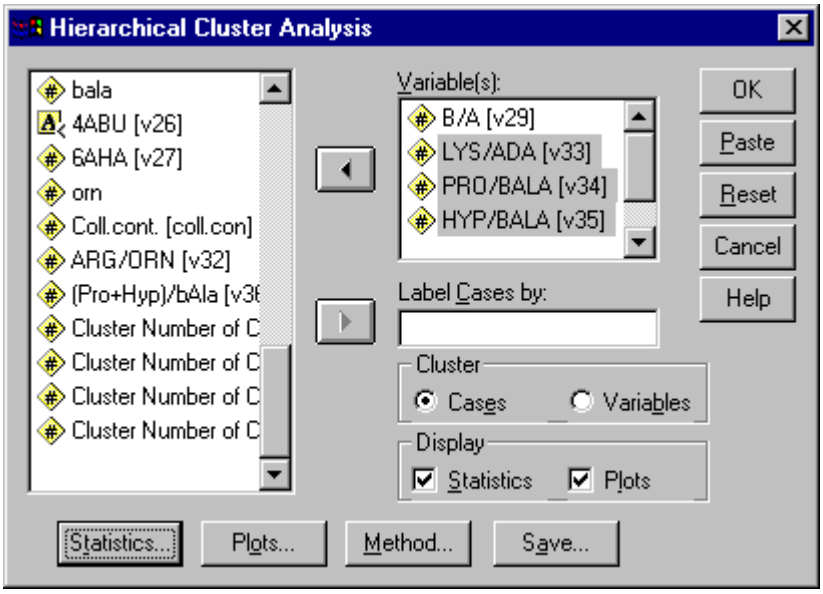
Επειδή όμως οι μεταβλητές έχουν πολύ διαφορετικές τυπικές αποκλίσεις είναι χρήσιμο να τυποποιήσουμε τις μεταβλητές ώστε να έχουν την ίδια τυπική απόκλιση. Αν δεν μετασχηματίσουμε τα δεδομένα η μεταβλητή με τη μεγαλύτερη τυπική απόκλιση

(PRO/BALA) θα έχει πολύ μεγαλύτερο βάρος. Παρατηρείστε πόσο μεγαλύτερη είναι η τυπική της απόκλιση. Στον πίνακα 9.13 μπορεί να δει κάποιος τις ευκλείδειες αποστάσεις (χρησιμοποιήσαμε για την ανάλυση την Ευκλείδεια Απόσταση). Οι παρατηρήσεις 7 και 9 είναι οι παρατηρήσεις με τη μικρότερη απόσταση.

Για να εκτελέσουμε την ιεραρχική ομαδοποίηση στο SPSS διαλέγουμε

Analyse > Classify > Hierarchical Clustering

κι εμφανίζεται το κεντρικό παράθυρο που βλέπεται στην εικόνα 9.4.



Εικόνα 9.4. Το βασικό παράθυρο για ιεραρχική ομαδοποίηση

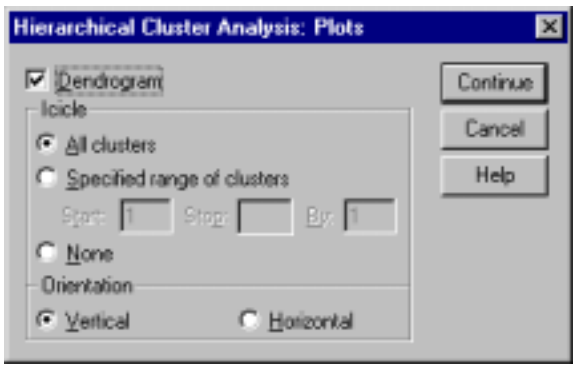
	Squared Euclidean Distance													
Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Case 2	25.25													
Case 3	7.39	8.65												
Case 4	1.65	17.75	2.88											
Case 5	5.75	10.60	.41	1.48										
Case 6	20.29	.79	7.84	15.04	9.66									
Case 7	10.78	3.76	1.96	5.43	2.09	3.32								
Case 8	14.67	1.50	4.36	9.56	5.32	.74	.94							
Case 9	9.92	4.02	1.92	5.02	2.03	3.25	.04	.91						
Case 10	17.66	8.37	14.89	16.59	15.28	4.54	8.22	5.06	7.54					
Case 11	20.32	3.28	11.31	15.41	11.56	2.13	4.17	2.09	4.03	2.41				
Case 12	23.79	.30	8.52	17.43	10.79	.30	4.12	1.37	4.22	7.02	3.46			
Case 13	13.47	7.31	.95	6.62	1.91	8.15	2.71	5.03	3.01	18.94	12.77	7.80		
Case 14	41.18	3.35	19.79	33.49	23.81	4.08	13.45	7.85	13.67	13.04	9.22	2.70	17.50	
Case 15	8.94	5.14	2.17	4.28	1.89	4.19	.21	1.45	**.11**	7.60	4.17	5.45	3.65	15.79

This is a dissimilarity matrix

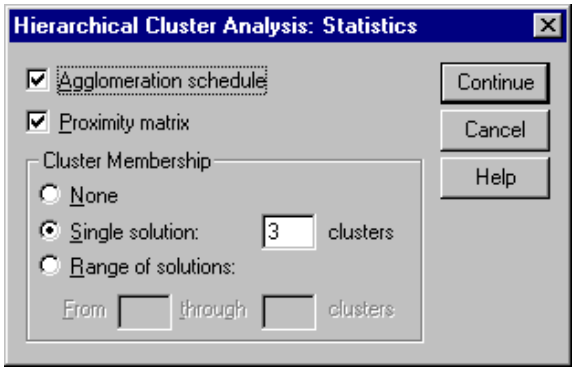
Πίνακας 9.13. Οι αποστάσεις όλων των παρατηρήσεων μεταξύ τους. Επειδή ο πίνακας είναι συμμετρικός μόνο τα στοιχεία της κάτω διαγωνίου δίνονται

Σε αυτό το παράθυρο πρέπει να επιλέξουμε τις μεταβλητές που θα χρησιμοποιήσουμε στην ανάλυση. Από αυτό το παράθυρο βλέπουμε ότι μας δίνονται μια πλειάδα επιλογών. Κατά αρχάς το SPSS μας επιτρέπει να κάνουμε και ομαδοποίηση ως προς μεταβλητές (αντί δηλαδή να ομαδοποιήσουμε τις παρατηρήσεις μας να ομαδοποιήσουμε τις μεταβλητές. Κάτι τέτοιο είναι επικίνδυνο και ανόητο μερικές φορές αφού η διαδικασία είναι χωρίς νόημα. Χρειάζεται μεγάλη προσοχή αν δοκιμάσετε ποτέ κάτι τέτοιο και γενικά πρέπει να το αποφεύγετε, καθώς υπάρχουν άλλες μέθοδοι στη στατιστική που μπορούν να σας ομαδοποιήσουν κατά κάποια έννοια τις μεταβλητές σας). Αυτά που πρέπει να διαλέξουμε είναι ποια απόσταση θα χρησιμοποιήσουμε, τι γραφήματα θα φτιάξουμε, με ποιόν τρόπο θα υπολογίζουμε τις αποστάσεις ανάμεσα σε ομάδες, ποιες λύσεις θέλουμε να σώσουμε για περαιτέρω επεξεργασία και διάφορα άλλα που θα δούμε τώρα αμέσως.

Το δενδρόγραμμα και το γράφημα Icicle είναι 2 γραφήματα που μπορούν να μας δώσουν γραφικά τη σειρά με την οποία οι παρατηρήσεις ενώνονται για να δημιουργήσουν ομάδες. Από το παράθυρο που βλέπουμε στην εικόνα 9.5 μπορούμε να επιλέξουμε αυτά τα γραφήματα. Επειδή αυτά θα περιγράψουν όλη την διαδικασία, αν ο αριθμός των παρατηρήσεων είναι πολύ μεγάλος, τα γραφήματα δεν θα είναι ιδιαίτερα ευκολοδιάβαστα. Για αυτό, όπως μπορείτε και να δείτε, μπορούμε να επιλέξουμε το εύρος του αριθμού των ομάδων για τις οποίες θα εμφανιστεί το γράφημα. Επίσης μπορούμε να καθορίσουμε αν το γράφημα θα εμφανιστεί οριζόντια ή κάθετα.



Εικόνα 9.5. Οι επιλογές για τα γραφήματα που θα πάρουμε όταν κάνουμε ιεραρχική ομαδοποίηση

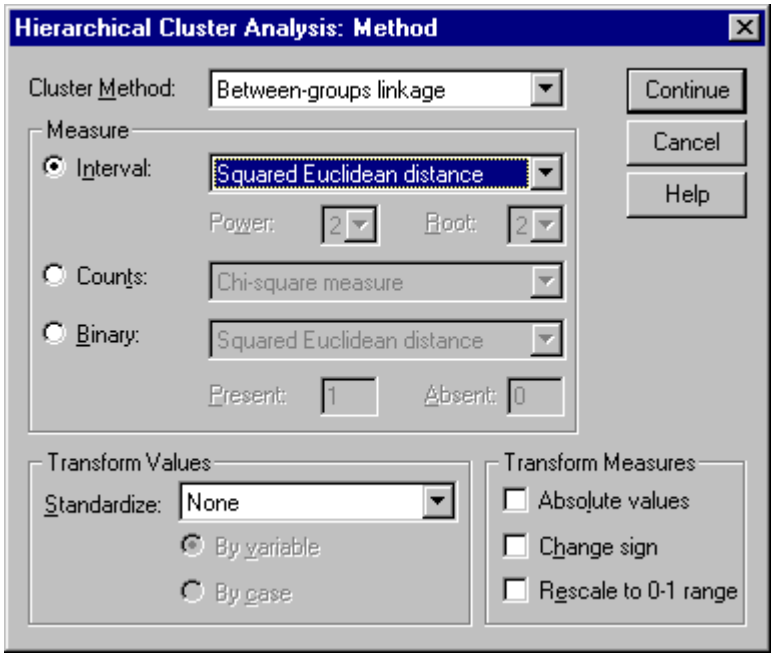


Εικόνα 9.6. Οι επιλογές για τα στατιστικά που θα πάρουμε όταν κάνουμε ιεραρχική ομαδοποίηση

Ομοίως στην εικόνα 9.6 βλέπουμε το παράθυρο της επιλογής Statistics. Οι δυνατές επιλογές έχουν να κάνουν με τις πληροφορίες που θα εμφανιστούν. Έτσι με την επιλογή Proximity Matrix εμφανίζουμε τον πίνακα αποστάσεων όλων των παρατηρήσεων, ενώ με την επιλογή Agglomerative Schedule εμφανίζονται κάποιες ποσότητες που όπως θα δούμε είναι χρήσιμες για να βρούμε τον αριθμό των ομάδων που θα κρατήσουμε. Επίσης στο κάτω μέρος του παραθύρου μπορούμε να επιλέξουμε να δούμε (και όχι να σώσουμε σε μεταβλητή) σε ποια ομάδα ανήκει κάθε παρατήρηση τόσο για συγκεκριμένο αριθμό ομάδων όσο και για διάφορα πλήθη ομάδων.

Γενικά μια καλή τακτική όταν κάνουμε ιεραρχική ομαδοποίηση είναι να αποφεύγουμε να ζητάμε στατιστικά που έχουν να κάνουν με τη διαδικασία καθώς αν ο αριθμός των παρατηρήσεων είναι μεγάλος (πχ πάνω από 100) τα αποτελέσματα δεν θα είναι καθόλου εύκολο να εμφανιστούν στην οθόνη πολύ δε περισσότερο να τα διαβάσουμε και να τα ερμηνεύσουμε. Επομένως καλό είναι να τρέξουμε μια φορά την ανάλυση με όσο γίνεται μικρότερο δυνατό αριθμό αποτελεσμάτων και μετά αφού πια έχουμε αποκτήσει μια εικόνα να ζητήσουμε κάποια άλλα αποτελέσματα..

Και ερχόμαστε τώρα στο πιο σημαντικό παράθυρο. Από την επιλογή Method εμφανίζεται το παράθυρο της εικόνας 9.7. Σε αυτό θα πρέπει να καθορίσουμε τη μέθοδο με την οποία θα υπολογίσουμε την απόσταση ανάμεσα σε 2 ομάδες, καθώς και την απόσταση που θα χρησιμοποιήσουμε. Από θεωρητικής πλευράς περιγράψαμε και προτείναμε τις καλύτερες επιλογές. Παρατηρείστε πως τα μέτρα αποστάσεων είναι ομαδοποιημένα έτσι ώστε ανάλογα με τον τύπο των δεδομένων ο χρήστης να μπορεί να επιλέξει το κατάλληλο μέτρο. Παρατηρείστε επίσης πως στο κάτω μέρος εμφανίζεται ένα πλήθος μετασχηματισμών των δεδομένων ώστε να μπορεί κάποιος να μεγαλώσει τις δυνατές επιλογές. Χρησιμοποιώντας μετασχηματισμούς μέσα από αυτό το παράθυρο ο χρήστης γλιτώνει την ανάγκη να δημιουργήσει καινούριες μεταβλητές με άλλες διαδικασίες.



Εικόνα 9.7. Επιλογή μεθόδου που θα χρησιμοποιήσουμε για την ιεραρχική ομαδοποίηση

Τέλος και πάλι μπορούμε να δημιουργήσουμε μεταβλητές που να μας δείχνουν, για τη συγκεκριμένη λύση με το συγκεκριμένο αριθμό ομάδων, που ανήκει κάθε παρατήρηση από το παράθυρο της εικόνας 9.8. Τώρα όμως ο αριθμός των ομάδων διαφέρει και άρα οι επιλογές μας είναι περισσότερες. Έτσι αν επιθυμούμε μπορούμε να δημιουργήσουμε μεταβλητές για πολλές δυνατές λύσεις ανάλογα με τον αριθμό των ομάδων

Εικόνα 9.8. Δημιουργία νέων μεταβλητών

Επιλέξαμε λοιπόν τις 4 μεταβλητές που είδαμε στην εφαρμογή της K-means ομαδοποίησης ότι παρέχουν ικανότητα ομαδοποίησης. Ο πίνακας 9.14 που ακολουθεί μας δείχνει τη σειρά με την οποία έγινε η ομαδοποίηση για δεδομένη απόσταση και μέθοδο (Απόσταση = Τετραγωνική Ευκλείδεια Απόσταση, Μέθοδος = Ward).

	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
Stage	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	7	9	.021	0	0	2
2	7	15	.122	1	0	10
3	6	12	.271	0	0	5
4	3	5	.475	0	0	8
5	2	6	.789	0	3	6
6	2	8	1.575	5	0	11
7	1	4	2.398	0	0	13
8	3	13	3.285	4	0	10
9	10	11	4.491	0	0	12
10	3	7	7.459	8	2	13
11	2	14	10.806	6	0	12
12	2	10	16.944	11	9	14
13	1	3	25.526	7	10	14
14	1	2	56.000	13	12	0

Πίνακας 9.14. Agglomeration Schedule. Η σειρά με την οποία έγιναν οι ομαδοποιήσεις

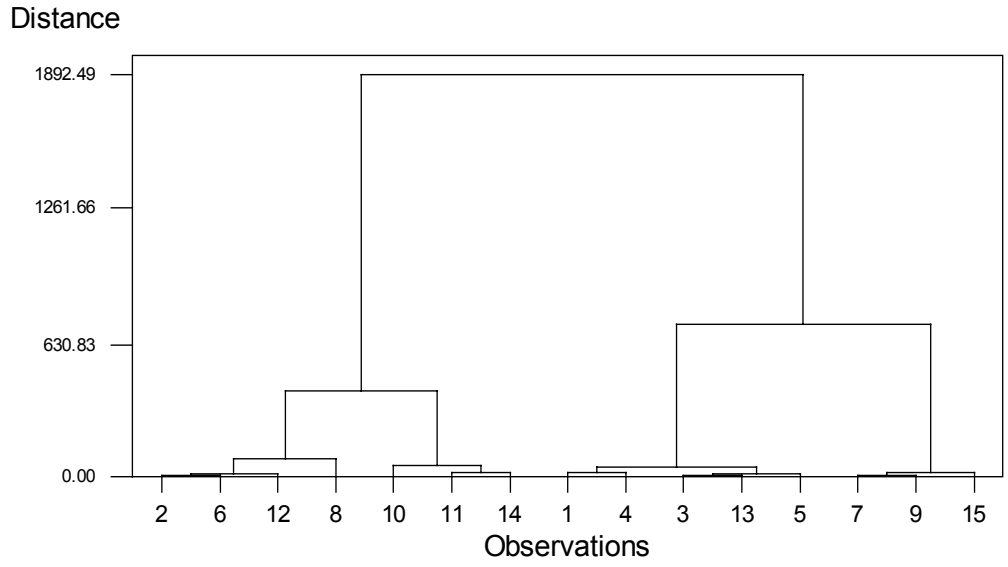
Επειδή λοιπόν η μικρότερη απόσταση ήταν ανάμεσα στις παρατηρήσεις 7 και 9 αυτές δημιούργησαν στο πρώτο στάδιο μια ομάδα ενώ οι υπόλοιπες 13 παρατηρήσεις ήταν κάθε μια από μια ομάδα. Άρα έχουμε μετά το πρώτο στάδιο 14 ομάδες. Κάτω από τη στήλη Coefficients βλέπουμε την τιμή της απόστασης ανάμεσα στις 2 παρατηρήσεις που ενώθηκαν. Επίσης βλέπουμε για κάθε μια παρατήρηση σε ποιο στάδιο αυτή ξαναχρησιμοποιείται για να δημιουργήσει άλλη ομάδα. Δηλαδή η παρατήρηση 7 ξαναχρησιμοποιείται στο στάδιο 2. Φυσικά τώρα πια δεν είναι μια μόνη της παρατήρηση αλλά μαζί με την 9 έχουν δημιουργήσει μια ομάδα και αυτή η ομάδα θα ξαναφανεί στο στάδιο 2. Όντως στο στάδιο 2 η παρατήρηση 15 ενώνεται με τις 7 και 9 κι έχουμε μια ομάδα με 3 παρατηρήσεις και τις υπόλοιπες 12 μόνες τους.

Στο 3ο στάδιο ενώνονται οι παρατηρήσεις 6 και 12, ενώ στο 4ο στάδιο οι παρατηρήσεις 3 και 5. Επομένως μετά το 4ο στάδιο έχουμε τις ομάδες {7,9,15}, {6,12} και {3,5} ενώ οι υπόλοιπες παρατηρήσεις είναι μόνες τους μια ομάδα. Προχωρώντας λοιπόν μέχρι το τέλος καταλήγουμε με μια ομάδα που περιέχει όλες τις παρατηρήσεις. Ο παρακάτω πίνακας περιέχει τις ομάδες για διάφορα πλήθη ομάδων (2,3 και 4) που προκύπτουν από την ιεραρχική μέθοδο. Βλέπουμε πως παρατηρήσεις που μπήκαν μαζί στην ίδια ομάδα σε αρχικό στάδιο μένουν μαζί για πάντα.

Η στήλη Coefficients είναι χρήσιμη καθώς η τιμή της μεγαλώνει σε κάθε στάδιο. (αν είχαμε χρησιμοποιήσει ομοιότητα αντί για απόσταση θα μειωνόταν). Επομένως ένα κριτήριο για το πότε θα σταματάμε είναι όταν η τιμή στη στήλη αυτή γίνει ξαφνικά πολύ μεγάλη. Για παράδειγμα μετά από τις 3 ομάδες βλέπουμε τιμή να πηδά από το 25 στο 56 περίπου. Αυτό σημαίνει ότι πια οι παρατηρήσεις (ομάδες) είναι πια αρκετά μακριά η μια από την άλλη και άρα περαιτέρω ομαδοποίηση δεν είναι λογική.

	1η ομάδα	2η ομάδα	3η ομάδα	4η ομάδα
2 ομάδες	1,3,4,5,7,9,13,15	2,6,8,10,11,12,14	-	-
3 ομάδες	1, 4,	2,6,8,10,11,12,14	3, 5, 7, 9, 13,15	-
4 ομάδες	1, 4,	2,6,8,12,14	3, 5, 7, 9, 13,15	10,11

Το δενδρόγραμμα είναι ένα γράφημα που μας δείχνει ακριβώς αυτό που περιγράψαμε πριν με λόγια, τη σειρά με την οποία ενώνονται οι παρατηρήσεις. Βλέπετε πως οι παρατηρήσεις ενώνονται με γραμμές ανάλογα με το στάδιο στο οποίο μπαίνουν στην ίδια ομάδα



Γράφημα 9.8. Δενδρόγραμμα για ιεραρχική ταξινόμηση των δεδομένων

Το Icicle γράφημα μας περιγράφει ακριβώς την ίδια διαδικασία. Αλλά με πολύ χειρότερα γραφικά και για αυτό όταν υπάρχει η δυνατότητα να πάρουμε το δενδρόγραμμα η επιλογή του μπορεί να αποφευχθεί.

Τελειώνοντας αυτή την παρουσίαση των μεθόδων ομαδοποίησης , στον πίνακα 9.15 μπορεί να δει κανείς τις ομάδες όπως αυτές δημιουργήθηκαν με τις 2 μεθόδους επιλέγοντας διάφορα μέτρα αποστάσεις. Παρατηρείστε πόσο διαφορετικά είναι τα αποτελέσματα

	K-means	Ιεραρχική μέθοδος			
		Μέθοδος Ward			Nearest Neighbour
Παρατήρηση		Ευκλείδεια απόσταση με μετασχηματισμό	Απόσταση Block	Ευκλείδιεα χωρίς μετασχηματισμό	Ευκλείδεια απόσταση με μετασχηματισμό
1	1	1	1	1	1
2	2	2	2	2	1
3	1	3	1	1	1
4	1	1	1	1	1
5	1	3	1	1	1
6	2	2	2	2	1
7	3	3	3	3	1
8	3	2	2	2	1
9	3	3	3	3	1
10	2	2	2	2	2
11	2	2	2	2	1
12	2	2	2	2	1
13	1	3	1	1	1
14	2	2	2	2	3
15	3	3	3	3	1

Πίνακας 9.15. Η ομαδοποίησης των παρατηρήσεων μας σε 3 ομάδες χρησιμοποιώντας διάφορες μεθόδους. Παρατηρείστε ότι υπάρχουν διαφορές από μέθοδο σε μέθοδο

9.6 Ανάλυση σε ομάδες με τη χρήση πιθανοθεωρητικού μοντέλου

Οι μέθοδοι που περιγράφηκαν προηγούμενα δεν στηρίζονται σε κανένα πιθανοθεωρητικό μοντέλο και στην πραγματικότητα προσφέρουν πολύ λίγα στοιχεία στατιστικής συμπερασματολογίας. Οι προσεγγίσεις τους είναι κυρίως μαθηματικές και σε κανένα σημείο δεν λαμβάνεται υπόψη η μεταβλητότητα που ίσως έχει σοβαρό ρόλο στα αποτελέσματα.

Θα πρέπει να συνηγορήσει κανείς ότι σε αρκετές εφαρμογές αυτή η έλλειψη κάποιου μοντέλου είναι επιθυμητή, ουσιαστικά αφήνουμε τα δεδομένα να μιλήσουν χωρίς να τα προσαρμόζουμε σε κάποιο ιδεατό και πιθανότατα λάθος μοντέλο. Από την άλλη αυτή η έλλειψη στατιστικού υποβάθρου μας εμποδίζει από το να κάνουμε στατιστική συμπερασματολογία.

Στα πλαίσια αυτά έχει αναπτυχθεί μια ολότελα διαφορετική μεθοδολογία η οποία δεν έχει να κάνει με την έννοια της απόστασης αλλά χρησιμοποιεί συγκεκριμένα στατιστικά μοντέλα .

Έστω πως ένας πληθυσμός αποτελείται από k υποπληθυσμούς (ομάδες), καθένας από τους οποίους είναι ομοιογενής και επομένως όλα τα μέλη του υποπληθυσμού ακολουθούν μια συγκεκριμένη κατανομή, έστω f_j όπου ο δείκτης δείχνει πως μιλάμε για την κατανομή

του *j* υποπληθυσμού. Για να διευκολύνουμε την παρουσίαση ας υποθέσουμε πως όλοι οι υποπληθυσμοί ακολουθούν την ίδια κατανομή αλλά με διαφορετικές παραμέτρους. Π.χ. ας υποθέσουμε πως όλοι οι υποπληθυσμοί ακολουθούν μια κανονική κατανομή αλλά με διαφορετική μέση τιμή έστω μ_j, *j=1,…k*.. Επομένως συμβολίζουμε την κατανομή του *j* υποπληθυσμού ως $f(x \mid \theta_j)$.

Αν πάρουμε τυχαία ένα άτομο από τον πληθυσμό αυτό και δεν γνωρίζουμε από ποιο υποπληθυσμό προέρχεται τότε από το θεώρημα ολικής πιθανότητας η κατανομή του θα είναι

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

όπου $0 < p_j < 1, \sum_{j=1}^{k} p_j = 1$ δηλώνει την πιθανότητα ένα τυχαίο άτομο να ανήκει στον υποπληθυσμό *j*. Παρατηρείστε πως κανείς μπορεί να γενικεύσει την ιδέα υποθέτοντας πως υπάρχουν άπειροι υποπληθυσμοί και να αντικαταστήσει το άθροισμα με ολοκλήρωμα.

Ας περιοριστούμε στη συζήτηση μίξεων κανονικών κατανομών και ας υποθέσουμε για αρχή πως έχουμε *k=2*, δηλαδή μόλις δύο υποπληθυσμούς. Αν με

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

συμβολίσουμε τη συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής τότε ένα μείγμα από δύο κανονικές θα έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x \mid \mu_1, \mu_2, \sigma^2, p_1) = p_1 f(x \mid \mu_1, \sigma^2) + (1-p_1) f(x \mid \mu_2, \sigma^2) =$$
$$\frac{p_1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2} \right) + \frac{(1-p_1)}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2} \right)$$

Το παραπάνω μοντέλο λοιπόν περιγράφει την περίπτωση 2 ομάδων με μονοδιάστατη συνεχή δεδομένα, δηλαδή μια μόνο μεταβλητή. Στην πιο ρεαλιστική περίπτωση η $f(x \mid \theta_j)$ είναι μια πολυμεταβλητή κανονική κατανομή, οπότε ο αριθμός των μεταβλητών είναι μεγάλος. Με βάση αυτά που είδαμε σε προηγούμενο κεφάλαιο θα ισχύει

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

και επομένως με βάση αυτά που ξέρουμε από την εκτιμητική, οι παράμετροι του μοντέλου μας που είναι οι πιθανότητες $p_j, j = 1,...,k$ τα διανύσματα των μέσων μ_j και οι αντίστοιχοι πίνακες διακύμανσης $\mathbf{\Sigma}_j$ θα πρέπει να εκτιμηθούν από τα δεδομένα. Η πιθανοφάνεια θα πάρει την όχι εύχρηστη μορφή

$$L(\Theta) = \prod_{i=1}^{n} \ln\left(\sum_{j=1}^{k} p_j f(\mathbf{x}_i \mid \mu_j, \Sigma_j) \right)$$

Είναι σαφές ότι η εκτίμηση των παραμέτρων δεν είναι εύκολο να γίνει, αυτός ίσως είναι και ο λόγος που αυτά τα μοντέλα μόνο τα τελευταία χρόνια βρίσκουν ολοένα και αυξανόμενη εφαρμογή.

Παρόλα αυτά η εκτίμηση των παραμέτρων μπορεί να γίνει σχετικά εύκολα με τη χρήση του αλγορίθμου EM. Ο αλγόριθμος EM είναι ένας επαναληπτικός αλγόριθμος για την εκτίμηση με τη μέθοδο μεγίστης πιθανοφάνειας, κατάλληλος για τι περιπτώσεις που τα δεδομένα περιέχουν missing τιμές, ή (και αυτό είναι το πιο ενδιαφέρον) μπορούμε να αναπαραστήσουμε τα δεδομένα σαν να περιέχουν missing τιμές

Δεν θα δώσουμε περισσότερες λεπτομέρειες για τον αλγόριθμο αυτό. Θα πρέπει όμως να σημειώσουμε τα ακόλουθα:

- Το μεγαλύτερο μέρος της μεθοδολογίας έχει αναπτυχθεί για πολυμεταβλητές κανονικές κατανομές κι επομένως η μέθοδος είναι κατάλληλη για ομαδοποίηση πολυμεταβλητών συνεχών δεδομένων
- Αντίστοιχα έχει αναπτυχθεί η ανάλογη μεθοδολογία για μίξεις πολυωνυμικών κατανομών που επιτρέπουν την ομαδοποίηση κατηγορικών δεδομένων
- Η μεθοδολογία επιτρέπει τη στατιστική συμπερασματολογία. Για παράδειγμα μπορούμε να διαλέξουμε τον αριθμό των ομάδων χρησιμοποιώντας την πιθανοφάνεια των μοντέλων με διαφορετικό αριθμό ομάδων.
- Επίσης μπορεί κανείς να δει πως αφού υπάρχει ένα μοντέλο τότε η κατάταξη παρατηρήσεων σε ομάδες γίνεται με βάση πιθανοθεωρητικά κριτήρια. Συγκεκριμένα για κάθε παρατήρηση μπορούμε να υπολογίσουμε την εκ των υστέρων πιθανότητα η παρατήρηση να ανήκει σε κάθε ομάδα ως

$$w_{ij} = \frac{p_j f(\mathbf{x}_i \mid \mu_j, \Sigma_j)}{\sum_{j=1}^{k} p_j f(\mathbf{x}_i \mid \mu_j, \Sigma_j)}$$

Επομένως κατατάσσουμε κάθε παρατήρηση στην ομάδα με τη μεγαλύτερη posterior πιθανότητα. Είναι πολύ ενδιαφέρον πως η παραπάνω σχέση επιτρέπει την πιθανοθεωρητική κατάταξη των παρατηρήσεων δηλαδή μια παρατήρηση δεν ανήκει σε κάποιο συγκεκριμένη ομάδα αλλά σε κάθε ομάδα με κάποια πιθανότητα. Για παράδειγμα αν θέλουμε να κατατάξουμε αναποφάσιστους ψηφοφόρους σε από ένα προεκλογικό γκάλοπ, αντί να κατατάξουμε τον καθένα σε κάποιο κόμμα με βάση των εκ των υστέρων πιθανότητα να ανήκει εκεί τον κατατάσσουμε στα κόμματα με κάποια πιθανότητα στο καθένα. Αυτή η μορφή ομαδοποίησης όπου κάθε παρατήρηση δεν

ανήκει αναγκαστικά σε μια ομάδα αλλά ανήκει σε κάθε ομάδα με κάποια πιθανότητα συχνά αναφέρεται ως fuzzy clustering.

- Είναι πολύ ενδιαφέρον να πούμε πως η μέθοδος K-means αποτελεί ειδική περίπτωση του παραπάνω μοντέλου βασισμένο στην πολυμεταβλητή κανονική κατανομή, όπου όλοι οι πίνακες διακύμανσης είναι σφαιρικοί! Αυτό από μόνο του ίσως είναι ικανό να καταλάβετε την υπεραπλούστευση που περιέχει η μέθοδος K-means.

9.7 Άλλοι αλγόριθμοι

Οι αλγόριθμοι που είδαμε προηγουμένως (K-means, hierarchical, model based clustering) αν και αποτελούν τους πιο διαδεδομένους για την ομαδοποίηση δεδομένων δεν είναι παρά ένα μικρό μέρος από αυτούς που έχουν προταθεί στη βιβλιογραφία. Θα πρέπει να τονιστεί πως επειδή η ανάλυση σε ομάδες αφορά ένα μεγάλο εύρος ερευνητικών αντικειμένων πολλές μέθοδοι έχουν αναπτυχθεί και είναι γνωστές σε συγκεκριμένες γνωστικές περιοχές. Ας μην ξεχνάμε και τους περιορισμούς που θέτονται εξαιτίας της μη ύπαρξης εύχρηστων και κυρίως διαδεδομένων προγραμμάτων για να υλοποιούν πιο πολύπλοκους αλγορίθμους. Για λόγους πληρότητας αλλά και για να πάρει κανείς μια ιδέα θα παρουσιάσουμε εδώ κάποιες ιδέες αποφεύγοντας πάντως να δώσουμε πολλές λεπτομέρειες.

Ένας εναλλακτικός αλγόριθμος διαμέρισης που μοιάζει με τον αλγόριθμο K-means είναι ο ακόλουθος. Ξεκινάμε από μια αρχική διαμέριση. Στη συνέχεια υπολογίζουμε τη μεταβολή του κριτηρίου που χρησιμοποιούμε (για παράδειγμα αυτό θα μπορούσε να είναι η συνολική within cluster απόσταση) που προκύπτει αν μετακινήσουμε μια παρατήρηση από την ομάδα που είναι σε μια άλλη ομάδα. Από όλες τις δυνατές μετακινήσεις, κάνουμε τη μετακίνηση που βελτιώνει το κριτήριο μας περισσότερο. Αυτό γίνεται επαναληπτικά μέχρι να μην μπορούμε να βελτιώσουμε περαιτέρω το κριτήριο. Υπάρχουν αλγόριθμοι που βρίσκουν με γρήγορο τρόπο ποια παρατήρηση πρέπει να μετακινηθεί.

Ένας αλγόριθμος που μοιάζει με την ιεραρχική ταξινόμηση είναι ο αλγόριθμος minimum spanning tree. Ο αλγόριθμος αυτός αν και δεν είναι ακριβώς ιεραρχικός χρησιμοποιεί την έννοια του δενδρογράμματος ως εξής. Αν ενώσουμε όλες τις παρατηρήσεις μεταξύ τους ψάχνουμε να βρούμε το βέλτιστο δένδρο το οποίο συνδέει όλες τις παρατηρήσεις και ελαχιστοποιεί το άθροισμα τετραγώνων των αποστάσεων, δηλαδή κάθε κλαδί του δένδρου έχει κόστος ίσο με την απόσταση των δύο παρατηρήσεων που ενώνει και ψάχνουμε να βρούμε το δένδρο με τη μικρότερη απόσταση.

Για να αποφευχθεί ο μεγάλος υπολογιστικός φόρτος στον αλγόριθμο K-means έχει προταθεί ο αλγόριθμος Clustering of Large Applications (γνωστός ως CLARA). Ο αλγόριθμος αυτός διαλέγει τυχαία υποδείγματα από το αρχικό δείγμα. Στη συνέχεια τρέχουμε τον αλγόριθμο K-means σε αυτά τα δεδομένα και κατατάσσουμε τις υπόλοιπες παρατηρήσεις σύμφωνα με την απόσταση τους από τα κέντρα που έχουμε βρει.

Επαναλαμβάνουμε τη διαδικασία 5 ή περισσότερες φορές και επιλέγουμε την ομαδοποίηση που βελτιστοποιεί κάποιο κριτήριο. Ο αλγόριθμος προφανώς προσπαθεί να περιορίσει τον υπολογιστικό φόρτο του να δουλεύουμε με όλα τα δεδομένα άλλα έχει τον κίνδυνο πως μπορεί να αποτύχει να ομαδοποιήσει όλα τα δεδομένα με τον καλύτερο τρόπο. Αν πάλι πάρουμε περισσότερα υποδείγματα τότε το υπολογιστικό κόστος ακυρώνει την επιλογή της προσέγγισης αυτής.

9.8 Κριτήρια επιλογής αριθμού ομάδων

Βασισμένα στην ανάλυση διακύμανσης

Ας υποθέσουμε πως έχουμε πολυμεταβλητά συνεχή δεδομένα. Έχοντας κατατάξει τις παρατηρήσεις σε ομάδες, ουσιαστικά έχουμε δεδομένα που μοιάζουν με αυτά στην MANOVA. Συνεπώς με την ίδια ακριβώς τεχνική μπορούμε να διαμερίσουμε το συνολικό άθροισμα τετραγώνων σε δύο μέρη αυτό που δείχνει τις αποκλίσεις μέσα στις ομάδες και σε αυτό που δείχνει τις αποκλίσεις ανάμεσα στις ομάδες.

Συγκεκριμένα για συνεχή δεδομένα μπορεί κανείς να αναλύσει τις συνολικές τετραγωνικές αποκλίσεις σε δύο μέρη, αυτές μέσα στις ομάδες και αυτές ανάμεσα στις ομάδες, όπως φαίνεται στον πίνακα 9.16. Δεδομένου πως

Πηγή	SSP πίνακας
Μεταξύ γκρουπ (between)	$B = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})'$
ΜΕΣΑ ΣΤΑ ΓΚΡΟΥΠ (within)	$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)'$
Συνολική	$T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{\bullet\bullet})(x_{ij} - \bar{x}_{\bullet\bullet})'$

Πίνακας 9.16.

Για μια πετυχημένη ομαδοποίηση θέλουμε ο πίνακας B να είναι όσο γίνεται μεγαλύτερος ενώ ο πίνακας W όσο γίνεται μικρότερα αφού θα αντιστοιχεί σε όμοια στοιχεία μέσα στην ομάδα. Συνεπώς μπορούμε να φτιάξουμε κριτήρια βασισμένα σε αυτές τις ποσότητες για να διαλέξουμε την καλύτερη ομαδοποίηση. Αυτά τα κριτήρια μπορεί να βασίζονται στο ίχνος του πίνακα W ή στην ορίζουσα του πίνακα W. Επειδή όμως δεν λαμβάνουμε υπόψη μας και τον αριθμό των ομάδων τέτοια κριτήρια θα προτιμούν μεγάλο αριθμό ομάδων. Ένα κριτήριο που λαμβάνει υπόψη του και τον αριθμό των ομάδων είναι το

$$c = \frac{tr(B)}{g-1} \Big/ \frac{tr(W)}{n-g}$$

το οποίο στη MANOVA αφορούσε το κριτήριο Λ του Wilks. Διαλέγουμε την ομαδοποίηση που έχει τη μεγαλύτερη τιμή. Αυτό γιατί αν φτιάξουμε ομάδες με παρατηρήσεις όμοιες μέσα στην ομάδα αλλά με μεγάλες διαφορές από ομάδα σε ομάδα τότε περιμένουμε πως τα στοιχεία του πίνακα B θα είναι μεγάλα και του W μικρά και άρα το κριτήριο θα 'χει μεγάλη τιμή.

Το παραπάνω κριτήριο αφορά συνεχή δεδομένα και επομένως έχει περιορισμένη χρησιμότητα.

Βασισμένα στις αποστάσεις

Μια άλλη σειρά κριτηρίων δεν βασίζεται στα αρχικά δεδομένα αλλά στις αποστάσεις και επομένως μπορούν να χρησιμοποιηθούν για πολλές εφαρμογές. Συγκεκριμένα αν συμβολίσουμε την *i* παρατήρηση ως x_i (προφανώς είναι ένα διάνυσμα αποτελούμενο από πολλές τιμές, κι επίσης συμβολίσουμε με $\overline{x}$ και $\overline{x}_k$ το συνολικό μέσο και το μέσο της *k* ομάδας αντίστοιχα, τότε για κάθε ομάδα μπορούμε να υπολογίσουμε την ποσότητα

$$W_k = \sum_{i \in G} d(x_i, \overline{x}_k),$$

όπου το άθροισμα εκτείνεται για όλες τις παρατηρήσεις στην ομάδα. στην ουσία μετράμε το άθροισμα των αποστάσεων από το κέντρο της ομάδας. Συνήθως οι υπολογισμοί αφορούν την ευκλείδεια απόσταση. Επίσης η έννοια του κέντρου δεν είναι απαραίτητα η μέση τιμή και αυτό εξαρτάται όπως πολλές φορές έχουμε πει από τη μορφή των δεδομένων. Αν συμβολίσουμε με

$$P_k = \sum_{i=1}^{k} W_i$$

το άθροισμα όλων των αποκλίσεων των παρατηρήσεων από το κέντρο τους, θα θέλαμε για καλή ομαδοποίηση αυτή η ποσότητα να είναι όσο γίνεται πιο μικρή. Ο δείκτης δείχνει ότι αφορά τη λύση με *k* ομάδες.

Τέλος ορίζουμε ως

$$T = \sum_{i=1}^{n} d(x_i, \overline{x})$$

το άθροισμα των αποστάσεων κάθε παρατήρησης από το συνολικό κέντρο.

Ο αναγνώστης θα μπορούσε να παρατηρήσει πως ουσιαστικά χρησιμοποιούμε την ιδέα της ανάλυσης διακύμανσης και για αυτό τα αποτελέσματα είναι ακριβή στην περίπτωση της ευκλείδειας απόστασης.

Τα μέτρα που χρησιμοποιούμε είναι τα εξής

$$R^2 = 1 - \frac{P_k}{T},$$

$$F = \frac{\left(\frac{T - P_k}{k-1}\right)}{\left(\frac{P_k}{n-k}\right)}$$

το οποίο και καλείται pseudoF καθώς μοιάζει πολύ με το F test της απλής ανάλυσης διακύμανσης και

$$t^2 = \frac{W_r - W_k - W_m}{\frac{W_k + W_m}{n_k + n_m - 2}}$$

όπου *r* είναι η ομάδα που προκύπτει αν ενώσουμε τις ομάδες *k* και *m*. Το t^2 είναι ένα μέτρο που μας δείχνει αν είναι χρήσιμο να ενώσουμε δύο ομάδες σε μια καινούρια, ενώ τα πρώτα δύο κριτήρια αφορούν την επιλογή της καλύτερης ομαδοποίησης.

Ένα άλλο κριτήριο καθορίζει πως αν ο λόγος

$$\left(\frac{P_k}{P_{k+1}} - 1\right)(n - k - 1) > 10$$

τότε πρέπει να προχωρήσουμε προσθέτοντας την καινούρια ομάδα στις ήδη υπάρχουσες

Ένα άλλο κριτήριο αφορά την τιμή της απόστασης σε κάθε βήμα της ιεραρχικής ομαδοποίησης και συγκεκριμένα τη μικρότερη απόσταση με βάση την οποία συγχωνεύουμε δύο ομάδες. Αν συμβολίσουμε με z_j την απόσταση στο *j* βήμα τότε σταματάμε τη διαδικασία αν

$$z_{j+1} > \bar{z} + ks_z$$

όπου $\bar{z} = \sum_{i=1}^{n} z_i / n$ και $s_z = \sqrt{\sum_{i=1}^{n} (z_i - \bar{z})^2 / n}$ είναι η μέση τιμή και η τυπική απόκλιση αντίστοιχα των τιμών αυτών και *k* είναι κάποια κριτική τιμή, συνήθως ίση με 1.96. Επίσης μπορεί κανείς να χρησιμοποιήσει ένα οπτικό κριτήριο κάνοντας το γράφημα της απόστασης σε κάθε βήμα. Το γράφημα αυτό μοιάζει με το scree plot της παραγοντικής ανάλυσης.

Στην περίπτωση που χρησιμοποιούμε model based ομαδοποίηση τότε μπορούμε να βασίσουμε την επιλογή του αριθμού των ομάδων σε κριτήρια βασισμένα στην πιθανοφάνεια.

Τελειώνοντας αυτή την παρουσίαση είναι προφανές ότι για να βρούμε την καλύτερη ομαδοποίηση πρέπει να επαναλάβουμε πολλές φορές τη διαδικασία και να συγκρίνουμε τα αποτελέσματα. Οι δύο βασικές μέθοδοι που είδαμε συνήθως χρησιμοποιούνται συμπληρωματικά. Με την ιεραρχική μέθοδο παίρνουμε ουσιαστικά όλες τις λύσεις και άρα μπορούμε να βρούμε το βέλτιστο αριθμό ομάδων τις οποίες μορφοποιούμε στη συνέχεια χρησιμοποιώντας τον αλγόριθμο K-means.

Τέλος θα πρέπει να αναφέρουμε πως στην περίπτωση που έχουμε fuzzy clustering όπως για παράδειγμα στην περίπτωση model based clustering τότε μπορεί κανείς να χρησιμοποιήσει διάφορα άλλα μέτρα ώστε να επιλέξει τον αριθμό των ομάδων. Για παράδειγμα αν συμβολίσουμε με $w_{ij},\ i=1,...,n, j=1,...,k$ την πιθανότητα να ανήκει η i παρατήρηση στη j ομάδα τότε ένα κριτήριο βασισμένο στην έννοια της εντροπίας είναι το

$$I(k) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\ln(w_{ij})}{n\ln(1/k)}$$

χρησιμοποιώντας βέβαια ότι $0\log 0 = 0$. Το κριτήριο παίρνει τιμές στο διάστημα [0,1]. Στην περίπτωση που κάθε παρατήρηση έχει καταταχθεί σε μια ομάδα αποκλειστικά τότε το κριτήριο έχει την τιμή 1. Τιμές κοντά στο 0 δείχνουν πως η ομαδοποίηση είτε απέτυχε είτε δεν υπάρχει πραγματικά ομαδοποίηση των δεδομένων. Αν και το κριτήριο δίνει μια γενική εικόνα σχετικά με το πόσο καλή ήταν η ομαδοποίηση έχει μερικά μειονεκτήματα. Για παράδειγμα δεν χρησιμοποιεί τα δεδομένα αλλά μόνο τις πιθανότητες. Επίσης δίνει μια γενική εικόνα και όχι μια ειδική εικόνα σχετικά με το αν όλα πήγαν καλά. Δηλαδή αν υπάρχουν κάποιες σχεδόν άδειες ομάδες μπορεί το κριτήριο να είναι υψηλό παρόλα αυτά η ομαδοποίηση δεν είναι τόσο καλή. Στην πράξη τιμές κοντά στο 0.80 υποδεικνύουν πολύ καλή ομαδοποίηση. Με το κριτήριο αυτό διαλέγουμε τον αριθμό των ομάδων που έχει υψηλότερη τιμή το κριτήριο.

9.9 Διάφορα άλλα θέματα

9.9.1 Μεγάλα σετ δεδομένων

Η ευρεία χρήση των υπολογιστών καθώς και η αφθονία δεδομένων (πχ ένα σούπερ μάρκετ έχει πια σε ημερήσια βάση καταγεγραμμένες όλες τις συναλλαγές, για όλα τα υποκαταστήματα του, αυτό σημαίνει πάνω από 100.000 εγγραφές κάθε μέρα) η προσέγγιση του προβλήματος της ομαδοποίησης παρατηρήσεων έχει αλλάξει. Το βάρος πια πέφτει στην εύρεση αλγορίθμων που να μπορούν να χρησιμοποιήσουν την τεράστια πληροφορία σε σχετικά μικρό χρόνο. Αυτό έχει ως αποτέλεσμα τη δημιουργία αλγορίθμων με βασικό κριτήριο τον υπολογιστικό φόρτο ακόμα και με κόστος στην ακρίβεια των αποτελεσμάτων.

Για αυτό το σκοπό έχουν αναπτυχθεί κάποιοι αλγόριθμοι ομαδοποίησης που είναι στην ουσία τροποποιήσεις των υπαρχόντων με σκοπό την υπολογιστική βελτιστοποίηση τους.

Θα δώσουμε ένα τέτοιο παράδειγμα σε σχέση με τον αλγόριθμο K-means. Αν κάποιες παρατηρήσεις ανήκουν σε μια σειρά από επαναλήψεις στην ίδια ομάδα τότε αυτές δεν χρησιμοποιούνται στις υπόλοιπες επαναλήψεις για να κερδηθεί χρόνος. Σε αυτή την περίπτωση υπάρχει ο κίνδυνος ότι θα μπορούσαν δυνητικά να επηρεάσουν την όλη διαδικασία αφού ουσιαστικά τις αγνοούμε! Θα πρέπει μάλιστα να τονιστεί ότι το σφάλμα από τέτοιες προσεγγίσεις δεν έχει μετρηθεί παρά εμπειρικά και επομένως έχουμε αλγορίθμους σε καθαρά εμπειρική βάση, για τους οποίους γνωρίζουμε ότι δουλεύουν ικανοποιητικά αλλά ουσιαστικά αγνοούμε τα χαρακτηριστικά τους.

9.9.2 Ενδιαφέροντα σημεία

Οι περισσότερες μέθοδοι είναι σχεδιασμένες να δουλεύουν με δεδομένα τα οποία είναι περίπου σφαιρικά και να βρίσκουν ομάδες με περίπου τον ίδιο πίνακα διακύμανσης. Αντίθετα αν οι ομάδες έχουν περίεργα σχήματα τότε είναι πιθανό οι μέθοδοι να μην αναγνωρίσουν αυτές τις ομάδες και η ανάλυση να οδηγηθεί σε λανθασμένα αποτελέσματα. Ένας τρόπος για να το αποφύγουμε αυτό είναι να χρησιμοποιήσουμε κάποια τεχνική ώστε να μετασχηματίσουμε τα δεδομένα ώστε να είναι περίπου σφαιρικά κι επομένως οι μέθοδοι να έχουν πια μεγαλύτερη επιτυχία. Συνήθως αυτό γίνεται με μια τεχνική που ονομάζεται ανάλυση κανονικών συσχετίσεων και η οποία είναι η πολυμεταβλητή αντίστοιχη της ανάλυσης σε κύριες συνιστώσες.

Πολλές φορές η ανάλυση σε ομάδες χρησιμοποιείται συμπληρωματικά με άλλες πολυμεταβλητές τεχνικές. Για παράδειγμα αν ο αριθμός των μεταβλητών που σκοπεύουμε να χρησιμοποιήσουμε είναι μεγάλος μπορεί κανείς να χρησιμοποιήσει ανάλυση σε κύριες συνιστώσες ώστε να συμπυκνώσει την πληροφορία σε λιγότερες μεταβλητές και να χρησιμοποιήσει τις κύριες συνιστώσες για να κάνει ομαδοποίηση. Το κέρδος είναι αφενός η μείωση των διαστάσεων του προβλήματος αφετέρου μια καλύτερη ερμηνεία των ομάδων μέσα από τις κύριες συνιστώσες. Το κόστος μπορεί να είναι η ανάγκη για τη χρήση πολλών κυρίων συνιστωσών και το χάσιμο πληροφορίας κατά το μετασχηματισμό. Επίσης μικρές ομάδες που συνήθως αντιστοιχούν σε μικρές κύριες συνιστώσες μπορεί να παραληφθούν τελείως και συνεπώς να μην βρεθούν κατά την ομαδοποίηση

Αντίστοιχες τεχνικές μπορούν να χρησιμοποιηθούν και για κατηγορικά δεδομένα Οι τεχνικές αυτές είναι ιδιαίτερα δεδομένες στη γαλλική σχολή της στατιστικής. Χωρίς να μπούμε σε λεπτομέρειες οι τεχνικές της ανάλυσης αντιστοιχιών και της πολλαπλής ανάλυσης αντιστοιχιών (Correspondence analysis και Multiple Correspondence analysis αντίστοιχα) έχουν σαν σκοπό κάτι παρεμφερές με την ανάλυση σε κύριες συνιστώσες αλλά για κατηγορικά δεδομένα. Συγκεκριμένα οι μέθοδοι καταλήγουν σε κάποιες καινούριες μεταβλητές οι οποίες ονομάζονται άξονες και στην ουσία αποτελούν τις συντεταγμένες αν τα αρχικά δεδομένα προβληθούν σε ένα χώρο μικρότερων διαστάσεων. Επομένως όπως και

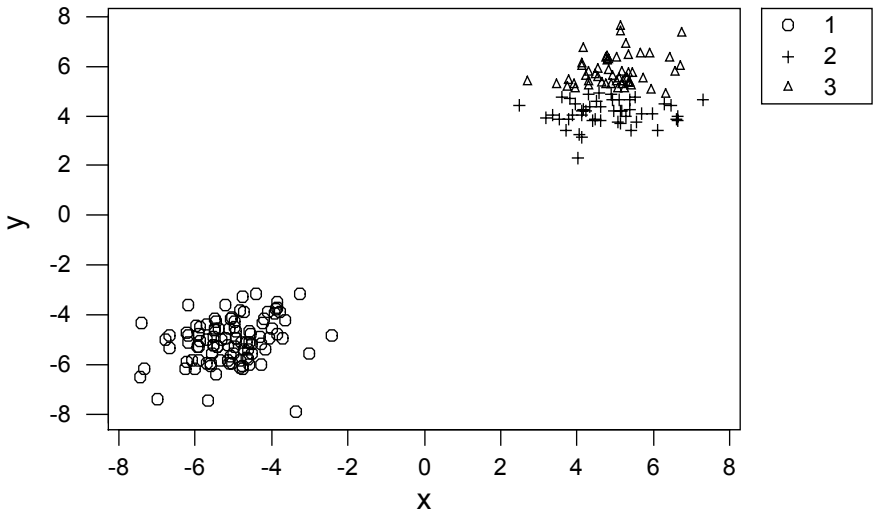
στη ανάλυση σε κύριες συνιστώσες μειώνουμε τις διαστάσεις του προβλήματος αλλά τώρα υπάρχει κι ένα καινούριο χρήσιμο στοιχείο. Οι αρχικές κατηγορικές μεταβλητές έχουν μετασχηματιστεί σε συνεχείς. Θα σταματήσουμε εδώ τη σύντομη περιγραφή αυτών των μεθόδων οι οποίες είναι αρκετά αμφισβητήσιμες από μεγάλο μέρος στατιστικών. Αυτό που είναι ενδιαφέρον είναι πως από κατηγορικά δεδομένα έχουμε καταλήξει σε συνεχή δεδομένα τα οποία κατά κάποιον τρόπο περιέχουν μεγάλο μέρος της πληροφορίας. Επομένως, δεδομένου πως είναι πολύ πιο εύκολο να κάνουμε ομαδοποίηση σε συνεχή δεδομένα μπορούμε να προχωρήσουμε σε ομαδοποίηση χρησιμοποιώντας τους άξονες που έχουμε πάρει.

9.9.3 Επιτυχία της μεθόδου

Όπως τονίσαμε πολλές φορές μέχρι τώρα η ανάλυση σε ομάδες δεν είναι τόσο απλή όσο φαίνεται. Μπορεί ο υπολογιστής να μας δώσει εύκολα κάποια ομαδοποίηση αλλά πρέπει να είμαστε σε θέση να δούμε αν αυτή η ομαδοποίηση έχει έννοια και κυρίως αν μπορούμε να την εμπιστευτούμε. Θα πρέπει να έχουμε πάντα στο μυαλό μας πως η λύση είναι λάθος είτε επειδή προσαρμόσαμε λάθος αριθμό ομάδων (πχ στην πραγματικότητα οι ομάδες είναι 3 κι εμείς κατασκευάσαμε 4) είτε επειδή υπάρχει τυχαία μεταβλητότητα κι επομένως θα θέλαμε να ξέρουμε κατά πόσο η λύση που βρήκαμε μπορεί να την ανάγουμε και σε άλλα σχετικά δεδομένα ή απλά περιγράφει τα δεδομένα που έχουμε στα χέρια μας. Επίσης παρατηρείστε πως υπάρχουν και άλλοι παράγοντες που καθορίζουν αυτό που βρήκαμε και αυτοί έχουν να κάνουν με τη μέθοδο και τα χαρακτηριστικά (παράμετροι) της που χρησιμοποιήθηκαν

Ένας πρώτος τρόπος ελέγχου είναι μια οπτική περιγραφή των δεδομένων. Κατά πόσο παρατηρήσεις που είναι κοντά μεταξύ τους οπτικά, έχουν καταταχθεί σε ίδιες ομάδες, αν οι ομάδες είναι καλά χωρισμένες μεταξύ τους κλπ

Για παράδειγμα δείτε το επόμενο γράφημα. Από τα 2 νέφη σημείων μπορεί κανείς να δει πως υπάρχουν 2 ομάδες, όμως επειδή ομαδοποιήσαμε τα σημεία σε 3 ομάδες, μια εκ των δύο ομάδων έχει χωριστεί στα 2. Παρόλα αυτά οι ομάδες θα είναι αρκετά συμπαγείς και αν δεν κοιτάξουμε την εικόνα μπορεί και να μην αντιληφθούμε το πρόβλημα



Γράφημα 9.9.

Επειδή όμως σε μεγαλύτερες διαστάσεις δεν είναι τόσο απλό να δούμε τα δεδομένα, μπορούμε να καταφύγουμε και σε κάποια μέτρα.

Προκειμένου λοιπόν να επιβεβαιώσουμε τα αποτελέσματα συνήθως καταφεύγουμε σε διάφορες τεχνικές όπως Cross validation και resampling τεχνικές. Ουσιαστικά δηλαδή επιλέγουμε κάποιο μικρότερο δείγμα από τα δεδομένα και προσπαθούμε να δούμε αν θα καταλήξουμε σε όμοια ομαδοποίηση σε αυτό το μικρότερο δείγμα,. Αν η ομαδοποίηση είναι καλή περιμένει κανείς πως και από το μικρότερο δείγμα θα βρει παρόμοια αποτελέσματα και επομένως έτσι επικυρώνουμε τα αποτελέσματα που έχουμε πάρει. Υπάρχουν πολλοί δείκτες για να ελέγξουμε αν δύο ομαδοποιήσεις είναι ίδιες. Συνήθως καταλήγουμε σε έναν πίνακα συνάφειας όπου όμως τα δεδομένα είναι εξαρτημένα αφού αφορούν τις ίδιες παρατηρήσεις. Δεν θα επεκταθούμε περισσότερο σε αυτές τις τεχνικές.

10 ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

10.1 Εισαγωγή

Η βασική ιδέα της διαχωριστικής ανάλυσης είναι να κατατάξει παρατηρήσεις (συνήθως πολυδιάστατες) σε γνωστούς πληθυσμούς με γνωστές κατανομές για κάθε πληθυσμό. Η διαχωριστική (ή διακριτική ανάλυση για άλλους συγγραφείς, discriminant analysis στα αγγλικά) αποτελεί μια μέθοδο με πλήθος εφαρμογών σε πολλές επιστήμες.

Ας υποθέσουμε ότι έχουμε Κ πληθυσμούς (ομάδες) $\Pi_1, \Pi_2, \ldots, \Pi_K$ με $K \geq 2$. Τότε για κάθε πληθυσμό Π_k έχουμε και μία κατανομή $f_k(\mathbf{x})$. Σκοπός της διαχωριστικής συνάρτησης είναι να «διαχωρίσει» ή να κατανείμει κάθε παρατήρηση στους Κ *γνωστούς* πληθυσμούς – ομάδες. Προφανώς ψάχνουμε για ένα διαχωριστικό κανόνα που μπορεί να κατατάξει σωστά όσο τον δυνατόν περισσότερες παρατηρήσεις.

Οι εφαρμογές της μεθόδου είναι πάρα πολλές. Είναι επίσης σημαντικό να αναφέρουμε πώς σε άλλες επιστήμες η μέθοδος αναφέρεται και με άλλες ονομασίες, όπως για παράδειγμα αναγνώριση προτύπων (pattern recognition) στην επιστήμη της πληροφορικής. Μερικά παραδείγματα εφαρμογών της μεθόδου είναι τα εξής::

- Στην Ιατρική συνήθως το ενδιαφέρον είναι να διαγνώσουμε την ασθένεια κάποιου ασθενή με βάση κάποια συμπτώματα που αυτός έχει. Δεδομένου πως για κάθε αρρώστια είναι γνωστά τα συμπτώματα της, θέλουμε να κατασκευάσουμε έναν κανόνα ο οποίος λαμβάνοντας υπόψη τα συμπτώματα αλλά και τη γνώση μας για τα συμπτώματα ενός συνόλου ασθενειών να κάνει διάγνωση για τον καινούριο ασθενή.

- Στα χρηματοοικονομικά οι τράπεζες ενδιαφέρονται να εντοπίσουν 'καλούς' και 'κακούς' πελάτες πριν τη χορήγηση δανείου η πιστωτικής κάρτας (credit scoring). Ως 'καλούς' και 'κακούς' μπορούμε να θεωρήσουμε αυτούς που πληρώνουν κανονικά τις δόσεις τους και αυτούς που δεν πληρώνουν αντίστοιχα. Συνεπώς με τη χρήση ιστορικών στοιχειών σχετικά με άτομα που έλαβαν δάνειο από την τράπεζα η τράπεζα μπορεί να σχηματίσει κανόνες ώστε να κατατάξει έναν καινούριο πελάτη σε

μια από τις δύο κατηγορίες και, πιθανότατα, να αρνηθεί τη χορήγηση δανείου, είτε να χορηγήσει το δάνειο με όρους σύμφωνους με το επίπεδο κινδύνου (risk) που έχει διαγνώσει για τον νέο πελάτη.

- Στο χώρο του marketing όπου ζητείται ο διαχωρισμός επιτυχημένος και αποτυχημένων αγορών ή διαφημιστικών εκστρατειών. Στην πρώτη περίπτωση μια εταιρεία αποφασίζει αν θα μπει σε μια αγορά ή όχι ενώ στη δεύτερη περίπτωση ποια διαφημιστική εκστρατεία ταιριάζει καλύτερα στην κάθε περίπτωση.

- Μια τελευταία εφαρμογή της διαχωριστικής ανάλυσης προέρχεται από το χώρο της ασφάλισης όπου μια εταιρεία πρέπει να αποφασίσει αν θα ασφαλίσει ή όχι ένα κίνδυνο (insurance risk management) χρησιμοποιώντας υπάρχοντα στοιχεία και δημιουργώντας αντίστοιχους κανόνες.

- Στις προεκλογικές καμπάνιες και γκάλοπ, συνήθως υπάρχει ένα έντονο πρόβλημα με τους αναποφάσιστους και γενικά αυτούς που δεν δηλώνουν καθαρά την προτίμηση τους. Σε αυτή την περίπτωση η διαχωριστική ανάλυση μπορεί να δημιουργήσει κανόνες ώστε ο αναποφάσιστος να κατατάσσεται σε κάποια ομάδα ψήφου.

- Πολλές φορές οι εικόνες από δορυφόρους δεν είναι άμεσα εκμεταλλεύσιμες και χρειάζονται επεξεργασία. Με βάση κάποια προηγούμενα δεδομένα μπορεί κανείς να κατατάξει το είδος της βλάστησης με βάση την εικόνα από το δορυφόρο και τη χρήση της διαχωριστικής ανάλυσης για την κατασκευή κανόνων κατάταξης.

- Στις κοινωνικές επιστήμες υπάρχει έντονο το ενδιαφέρον να κατατάξουμε ομάδες πληθυσμού σε συγκεκριμένες κοινωνικές ομάδες με βάση μια σειρά από χαρακτηριστικά που έχουν, όπως προβλήματα, οικονομικοκοινωνικά χαρακτηριστικά κλπ. Τέτοιες αποφάσεις μπορούν να χρησιμοποιηθούν για τη δημιουργία συγκεκριμένης κοινωνικής πολιτικής για παράδειγμα.

Τα παραδείγματα προφανώς δεν εξαντλούνται σε αυτά που μόλις αναφέρθηκαν αλλά δείχνουν την ποικιλία εφαρμογών της μεθόδου. Είναι ενδιαφέρον να παρατηρήσει κανείς πως η κατάταξη γίνεται είτε σε δύο (π.χ. παράδειγμα τράπεζας) είτε σε περισσότερες ομάδες (π.χ. παράδειγμα ιατρικής διάγνωσης).

Τέλος, να υπογραμμίσουμε ενώ η διαχωριστική ανάλυση μοιάζει με την ανάλυση σε ομάδες, που είδαμε στο προηγούμενο κεφάλαιο, έχει σημαντικές διαφορές από αυτή. Η πρώτη και ποιο σημαντική είναι ότι στη διαχωριστική ανάλυση οι ομάδες είναι γνωστές ενώ στην ανάλυση σε ομάδες δεν είναι και σκοπός μας είναι να βρούμε αυτές τις ομάδες. Για το λόγο αυτό ο στόχος είναι διαφορετικός. Στη διαχωριστική ανάλυση κύριο μέλημα μας είναι η κατασκευή ενός κανόνα που θα μας βοηθήσει να λάβουμε αποφάσεις στο μέλλον ενώ στην ανάλυση κατά συστάδες ο κύριος στόχος μας είναι να δημιουργήσουμε ομοειδής ομάδες με σκοπό την κατανόηση των ήδη υπαρχόντων στοιχείων και τη μείωση της διασποράς σε επιμέρους ομάδες.

Πριν προχωρήσουμε στη θεωρία της μεθόδου ας δούμε ένα απλό παράδειγμα για να πάρουμε μια πρώτη γεύση του τι προσπαθούμε να κάνουμε.

Ας υποθέσουμε πως έχουμε δύο πληθυσμούς. Ο πρώτος πληθυσμός ακολουθεί $N(0,1)$ ενώ ο δεύτερος $N(3,1.5)$. Και έστω ότι έχουμε μια καινούρια παρατήρηση με τιμή 4 και θέλουμε να την κατατάξουμε σε έναν από τους δύο πληθυσμούς. Ένας απλός τρόπος είναι να υπολογίσουμε την πιθανοφάνεια της παρατήρησης κάτω από κάθε μοντέλο και αυτό με τη μεγαλύτερη πιθανοφάνεια. Δηλαδή, επειδή $f(4 \mid \mu = 0, \sigma^2 = 1) = 0.000134$ ενώ $f(4 \mid \mu = 3, \sigma^2 = 1.5) = 0.212965$ θα κατατάξουμε την παρατήρηση στο δεύτερο πληθυσμό.

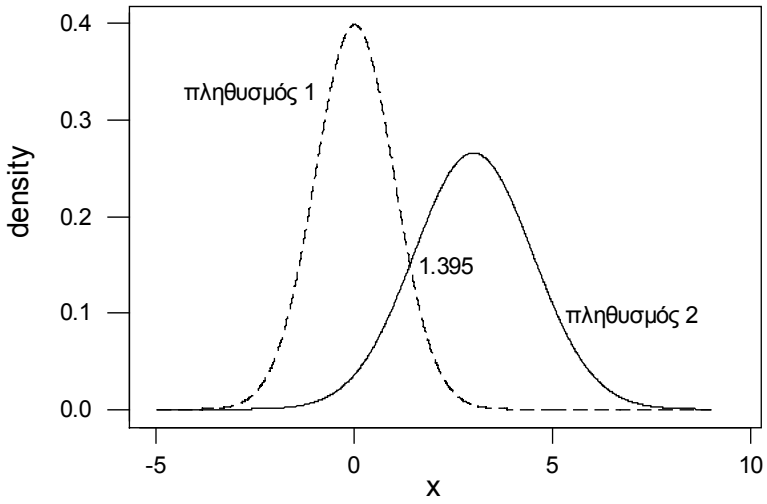
Στην πράξη η κατάσταση είναι πιο σύνθετη καθώς συνήθως δεν γνωρίζουμε τις παραμέτρους για κάθε πληθυσμό κι επομένως πρέπει να τις εκτιμήσουμε από ένα δείγμα, αλλά και θέλουμε έναν γενικό κανόνα κατάταξης καινούριων παρατηρήσεων. Από το γράφημα 10.1 που ακολουθεί μπορεί κανείς να κατασκευάσει έναν τέτοιο κανόνα. Μέχρι την τιμή 1.395 ο πρώτος πληθυσμός έχει μεγαλύτερη πιθανοφάνεια ενώ για τιμές μεγαλύτερες ο δεύτερος έχει μεγαλύτερη πιθανοφάνεια.

Προφανώς η παραπάνω ιδέα μπορεί να γενικευτεί

- Για περισσότερους από δύο πληθυσμούς
- Για μη συνεχή δεδομένα και για την ακρίβεια για κάθε μορφής δεδομένα
- Για περισσότερες από μια μεταβλητές με τη χρήση πολυδιάστατων κατανομών

Στην πράξη υπάρχουν και πολλά άλλα θέματα που πρέπει να αντιμετωπιστούν. Για παράδειγμα, η μορφή του κανόνα κατάταξης όταν έχουμε περισσότερες από μια μεταβλητές συνήθως είναι μια γραμμική συνάρτηση. Μπορεί κανείς να επεκτείνει την ιδέα ώστε να φτιαχτούν κανόνες με μη γραμμικές συναρτήσεις. Επίσης μπορεί κανείς αντί να χρησιμοποιήσει κάποια παραμετρική μορφή για την κατανομή του πληθυσμού να την εκτιμήσει με κάποια μέθοδο, όπως πχ με τη μέθοδο των kernels. Θα πρέπει να τονιστεί ότι η μεγάλη πρόοδος που έχει συντελεστεί τα τελευταία χρόνια οφείλεται στη χρήση υπολογιστών που επιτρέπει τη χρήση πολυπλοκότερων υποθέσεων

Όλα τα παραπάνω συνηγορούν στην άποψη πως η διαχωριστική ανάλυση είναι μια μέθοδος αρκετά πολύπλοκη στην κατασκευή της. Από την άλλη μεριά οι εφαρμογές της μοιράζονται σε μια ποικιλία επιστημών και γνωστικών αντικειμένων και για αυτό είναι πολύ χρήσιμη και διαδεδομένη στην επιστημονική κοινότητα. Θα παρουσιάσουμε τη μέθοδο ξεκινώντας από την απλή μορφή της και θα αναφέρουμε λίγα πράγματα για τις σύγχρονες επεκτάσεις της.



Γράφημα 10.1 Η πιθανοφάνεια για τους δύο πληθυσμούς του παραδείγματος. Οι πυκνότητες τέμνονται στο σημείο x=1395

10.2 Ο Βασικός Κανόνας Διαχωρισμού Δυο Ομάδων

Το βασικό στοιχείο της διαχωριστικής ανάλυσης είναι η δημιουργία κανόνων απόφασης σχετικά με την κατάταξη παρατηρήσεων σε διάφορους πληθυσμούς. Επομένως, στην ουσία έχουμε να αντιμετωπίσουμε ένα πρόβλημα θεωρίας αποφάσεων. Έτσι, όταν μπορούμε να ποσοτικοποιήσουμε τις απώλειες λόγω λανθασμένης κατάταξης μπορούμε να γράψουμε το αναμενόμενο κόστος ταξινόμησης μιας παρατήρησης που προέρχεται από την k ομάδα (ECM: expected cost of misclassification) δίδεται ως εξής:

$$ECM_k = \pi_k \sum_{l=1}^{K} C(l \mid k) P(l \mid k)$$

όπου C(l|k) είναι το κόστος να κατατάξουμε την παρατήρηση στη l ομάδα ενώ ανήκει στην k, αν k=l τότε το κόστος είναι μηδενικό, P(l|k) είναι η πιθανότητα να κατατάξουμε την παρατήρηση στη l ομάδα ενώ ανήκει στην k, και π_k είναι η εκ των προτέρων πιθανότητα (prior probability) να ανήκει μια παρατήρηση στον k πληθυσμό (ομάδα). Επίσης $f_k(\mathbf{x})$ είναι η συνάρτηση πιθανότητας (ή συνάρτηση πυκνότητας πιθανότητας) να παρατηρηθούν οι τιμές (χαρακτηριστικά) του διανύσματος **x** όταν βρισκόμαστε στην k ομάδα.

Το συνολικό κόστος είναι ίσο με το άθροισμα των επιμέρους ECM_k. Φυσικά επιλέγουμε να κατατάξουμε την παρατήρηση στην ομάδα με το μικρότερο αναμενόμενο κόστος λανθασμένης κατάταξης το οποίο είναι ισοδύναμο με ελαχιστοποίηση του συνολικού κόστους λανθασμένης κατάταξης.

Όταν έχουμε δύο ομάδες (K=2) τότε:

$ECM_1 = \pi_1 [\, C(1|1)\, P(1|1) + C(2|1)\, P(2|1) \,] = \pi_1\, C(2|1)\, P(2|1)$

$ECM_2 = \pi_2 [\, C(1|2)\, P(1|2) + C(2|2)\, P(2|2) \,] = \pi_2\, C(1|2)\, P(1|2)$

εφόσον $C(1|1)=C(2|2)=0$. Άρα ο βασικός κανόνα διαχωρισμού γίνεται:

επιλέγω να κατατάξω την παρατήρηση μου στην 1η ομάδα αν $ECM_1 \le ECM_2$ αλλιώς την κατατάσσω στη 2η ομάδα.

Μπορεί να δειχτεί πως ο παραπάνω κανόνας μπορεί να γραφτεί με τη χρήση των πυκνοτήτων κάθε πληθυσμού ως:

- *Αν* $\dfrac{f_1(\mathbf{x}_i)}{f_2(\mathbf{x}_i)} \ge \dfrac{\pi_2}{\pi_1} \times \dfrac{C(1|2)}{C(2|1)}$ *τότε κατατάσσουμε την i παρατήρηση στην 1η ομάδα*

- *Διαφορετικά κατατάσσουμε την i παρατήρηση στη 2η ομάδα.*

Όπου $\mathbf{x}_i$ είναι το διάνυσμα με τα χαρακτηριστικά (μεταβλητές) της i παρατήρησης, $C(1|2)$ είναι το κόστος που προέρχεται από την λανθασμένη καταχώρηση μιας παρατήρησης στην 1η ομάδα (ενώ πραγματικά ανήκει στη 2η) και $C(2|1)$ είναι το κόστος που προέρχεται από την λανθασμένη καταχώρηση μιας παρατήρησης στην 2η ομάδα (ενώ πραγματικά ανήκει στη 1η).

Ουσιαστικά μπορεί να δει κάποιος πως αν αγνοήσουμε το κόστος, αν δηλαδή θεωρήσουμε ίδιο κόστος για κάθε τύπο εσφαλμένης κατάταξης, τότε ο κανόνας κατατάσσει ανάλογα με την εκ των υστέρων πιθανότητα η παρατήρηση να ανήκει στην κάθε ομάδα. Δηλαδή η πιθανοφάνεια κάθε ομάδας σταθμίζεται με την εκ των προτέρων πιθανότητα η παρατήρηση να ανήκει σε κάθε ομάδα. Στην πράξη αυτό σημαίνει πως χρησιμοποιούμε την πληροφορία που έχουμε σχετικά με τη συχνότητα κάθε ομάδας.

Παρατηρείστε πως αν το κόστος είναι το ίδιο για τους δύο διαφορετικούς τρόπους λανθασμένης κατάταξης και οι εκ των προτέρων πιθανότητες ίδιες, τότε το κριτήριο είναι το ίδιο με αυτό που είδαμε στην προηγούμενη ενότητα όπου κατατάξαμε την παρατήρηση στον πληθυσμό με τη μεγαλύτερη πιθανοφάνεια.

Παράδειγμα:

Έστω μια ασθένεια με ποσοστό στον πληθυσμό 2%. Το κόστος να μην εντοπίσουμε σωστά τον ασθενή είναι 10 φορές μεγαλύτερο από το κόστος να μην εντοπίσουμε σωστά έναν υγιή. Στην ουσία αυτό σημαίνει ότι προτιμάμε να στείλουμε άδικα κάποιον για περαιτέρω θεραπεία παρά να μην εντοπίσουμε κάποιον ασθενή και να τον αφήσουμε χωρίς θεραπεία. Τότε 1ος πληθυσμός είναι οι υγιείς και 2ος οι ασθενείς, $C(1|2)$ = C(Υγιής | Ασθενής) = 10 C(Ασθενής | Υγιής)= 10 $C(2|1)$ οπότε ο διαχωριστικός κανόνας γίνεται:

Αν $\dfrac{f_1(\mathbf{x}_i)}{f_2(\mathbf{x}_i)} \ge \dfrac{0.02}{0.98} \times 10 = 0.204$ τότε θεωρούμε το i άτομο ως Υγιή (1η ομάδα) διαφορετικά θεωρούμε το i άτομο ως ασθενή (2η ομάδα) και του χορηγούμε περαιτέρω θεραπεία.

10.3 Διαχωρισμός Δυο Ομάδων με τη Χρήση της Κανονικής Κατανομής

Σε αυτή την ενότητα θα δούμε βήμα – βήμα πως θα φτάσουμε στη δημιουργία μιας διαχωριστικής συνάρτησης υποθέτοντας κανονικότητα των πληθυσμών. Η πιο συχνή επιλογή για την κατανομή τις κατανομές $f_k(\mathbf{x})$, $k=1,...,K$, των δεδομένων μέσα στην k ομάδα είναι η πολυμεταβλητή κανονική κατανομή. Επίσης για ευκολία υποθέτουμε ίσους πίνακες συνδιακύμανσης $\mathbf{\Sigma}$ αν και αργότερα θα δούμε πως αυτή η υπόθεση δεν είναι απαραίτητη. Έτσι κάθε ομάδα – πληθυσμός διαφέρει μόνο ως προς τις μέσες τιμές $\boldsymbol{\mu}_k$. Έτσι για την παρατήρηση $\mathbf{x}_i$ που ανήκει στον k πληθυσμό (δηλαδή $\mathbf{x}_i | k \sim N_p(\boldsymbol{\mu}_k, \mathbf{\Sigma})$) έχουμε

$$f_k(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \mathbf{\Sigma}) = (2\pi)^{-p/2} \mid \mathbf{\Sigma} \mid^{-1/2} e^{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_k)' \mathbf{\Sigma}^{-1} (\mathbf{x}_i-\boldsymbol{\mu}_k)}$$

που είναι η πυκνότητα μιας πολυμεταβλητής κανονικής κατανομής.

Θυμηθείτε ότι η ποσότητα $(\mathbf{x}_i - \boldsymbol{\mu}_k)' \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)$ ορίζει ένα μέτρο απόστασης i παρατήρησης από το μέσο της k ομάδας (απόσταση Mahalanobis).

Παίρνοντας λογαρίθμους στον διαχωριστικό κανόνα της προηγούμενης παραγράφου έχουμε

$$\ln \frac{f_1(\mathbf{x}_i \mid \boldsymbol{\mu}_1, \mathbf{\Sigma})}{f_2(\mathbf{x}_i \mid \boldsymbol{\mu}_2, \mathbf{\Sigma})} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Sigma}^{-1} \mathbf{x}_i - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

Θέτοντας $\mathbf{L}=\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ έχουμε ότι αν

$$\mathbf{L}'\mathbf{x}_i - \frac{1}{2}\mathbf{L}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq k_0 = \ln\left(\frac{\pi_2}{\pi_1}\frac{C(1 \mid 2)}{C(2 \mid 1)}\right)$$

κατατάσσουμε την παρατήρηση στην $1^η$ ομάδα αλλιώς την κατατάσσουμε στην $2^η$ ομάδα. Ο παραπάνω κανόνας ταυτίζεται και με το διαχωρισμό του Fisher και η συνάρτηση

$$U(\mathbf{x})=\mathbf{L}'\mathbf{x}-0.5\mathbf{L}'(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$$

λέγεται και γραμμική διαχωριστική συνάρτηση του Fisher. Έστω τώρα ότι έχουμε μια καινούρια παρατήρηση με χαρακτηριστικά $\mathbf{x}$. Σε αυτή την περίπτωση αν $k_0=0$ ο διαχωριστικός κανόνας γίνεται:

αν $U(\mathbf{x})\geq 0$ τότε κατατάσσουμε στην $1^η$ ομάδα αλλιώς στην $2^η$.

Επιπλέον αν πάρουμε τη μέση τιμή της $U(\mathbf{x})$ τότε έχουμε

$$E[U(\mathbf{x})]= \mathbf{L}'E(\mathbf{x})-0.5\mathbf{L}'(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2),$$

όπου $E(\mathbf{x})$ είναι η μέση τιμή των χαρακτηριστικών η οποία είναι ίση με $\boldsymbol{\mu}_k$ εάν η παρατήρηση $\mathbf{x}$ προέρχεται πραγματικά από την k ομάδα.

Αν λοιπόν η παρατήρηση $\mathbf{x}$ προέρχεται από την $1^η$ ομάδα τότε

$$E[U(\mathbf{x})]=\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=\frac{1}{2}a$$

ενώ αν η παρατήρηση **x** προέρχεται από την $2^{η}$ ομάδα τότε

$$E[U(\mathbf{x})]=-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=-\frac{1}{2}a,$$

όπου $\alpha=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ είναι η απόσταση Mahalanobis μεταξύ των μέσων των δύο ομάδων. Με τον ίδιο τρόπο μπορούμε να βρούμε ότι η διακύμανση του $U(\mathbf{x})$ δίδεται ως εξής

$$V[U(\mathbf{x})]= V(\mathbf{L}'\mathbf{x}-0.5\mathbf{L}'(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)) = V(\mathbf{L}'\mathbf{x}) = \mathbf{L}'V(\mathbf{x})\,\mathbf{L}=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\Sigma}\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) =$$

$$= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) =\alpha$$

Άρα από τα παραπάνω έχουμε ότι

$U(\mathbf{x})\sim N(\alpha/2,\alpha)$ αν η παρατήρηση **x** προέρχεται από την $1^{η}$ ομάδα

$U(\mathbf{x})\sim N(-\alpha/2,\alpha)$ αν η παρατήρηση **x** προέρχεται από τη $2^{η}$ ομάδα

οι μέσες τιμές των σκορ των διαχωριστικών συναρτήσεων ονομάζονται και κεντροειδή (centroid) καθώς είναι τα διανύσματα των μέσων και ουσιαστικά καθορίζουν το κέντρο της ομάδας.

Η πιθανότητα να κατατάξουμε λάθος μια παρατήρηση στη $1^{η}$ ομάδα δίδεται ως εξής:

P(Κατάταξη στην $1^{η}$ ομάδα | ανήκει πραγματικά στη $2^{η}$ ομάδα) =

$$= P(U(\mathbf{x}) \geq k_0 \mid U(\mathbf{x})\sim N(-\alpha/2,\alpha)) =$$

$$= P\left(\frac{U(\mathbf{x})+a/2}{\sqrt{a}} \geq \frac{k_0+a/2}{\sqrt{a}}\right) = P\left(Z \geq \frac{k_0+a/2}{\sqrt{a}}\right) = 1-P\left(Z < \frac{k_0+a/2}{\sqrt{a}}\right) =$$

$$= 1-\Phi\left(\frac{k_0+a/2}{\sqrt{a}}\right)$$

όπου $\Phi(x)$ η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής. Όμοια, η πιθανότητα να κατατάξουμε λάθος μια παρατήρηση στη $2^{η}$ ομάδα δίδεται ως εξής:

P(Κατάταξη στην $2^{η}$ ομάδα | ανήκει πραγματικά στη $1^{η}$ ομάδα) =

$$= P(U(\mathbf{x}) < k_0 \mid U(\mathbf{x})\sim N(\alpha/2,\alpha)) =$$

$$= P\left(\frac{U(\mathbf{x})-a/2}{\sqrt{a}} < \frac{k_0-a/2}{\sqrt{a}}\right) = P\left(Z < \frac{k_0-a/2}{\sqrt{a}}\right) = \Phi\left(\frac{k_0-a/2}{\sqrt{a}}\right).$$

Άρα η συνολική πιθανότητα λάθους δίδεται:

P(λανθασμένης κατάταξης) =

=P(λανθασμένης κατάταξης | ανήκει στη $1^η$ ομάδα) P(ανήκει στη $1^η$ ομάδα)+

+P(λανθασμένης κατάταξης | ανήκει στη $2^η$ ομάδα) P(ανήκει στη $2^η$ ομάδα)=

$= \pi_1 \Phi(\{k_0-\alpha/2\}/\sqrt{a})+ \pi_2 [1-\Phi(\{k_0+\alpha/2\}/\sqrt{a})]$

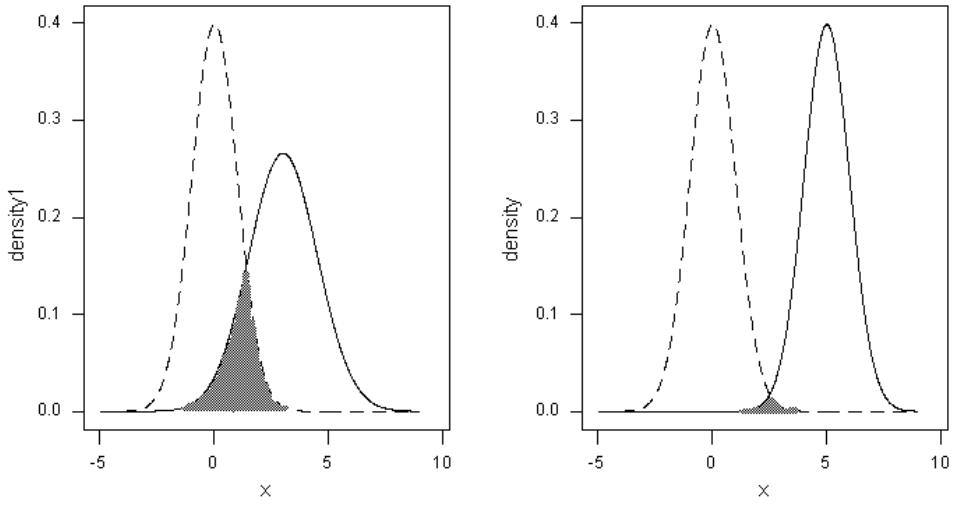
και για $k_0=0$ η πιθανότητας λανθασμένης κατάταξης γίνεται ίση με

$$\pi_1\Phi\left(-\sqrt{a}/2\right)+\pi_2\left[1-\Phi\left(\sqrt{a}/2\right)\right]$$

Επειδή όμως λόγω της συμμετρίας της κανονικής κατανομής ισχύει πως $\Phi\left(-\sqrt{a}/2\right)=\left[1-\Phi\left(\sqrt{a}/2\right)\right]$, προκύπτει πως η πιθανότητα λανθασμένης κατάταξης λανθασμένης κατάταξης είναι ίση με $\Phi\left(-\sqrt{a}/2\right)$. Άρα όσο πιο μεγάλο είναι το α (δηλαδή πιο μακριά τα δύο κεντροειδή) τόσο πιο καλός (επιτυχημένος) ο διαχωρισμός.

Αυτό είναι κάτι που περιμέναμε. Αν οι δύο πληθυσμοί είναι κοντά είναι λογικό να είναι πιο δύσκολο να τους ξεχωρίσουμε και άρα είναι πιο πιθανό να κάνουμε λάθος.

Στο γράφημα 10.2 μπορούμε να δούμε ένα απλό παράδειγμα με μονοδιάστατες κατανομές. Σε αυτή την περίπτωση η απόσταση είναι απλά η διαφορά των μέσων. Η πιθανότητα σφάλματος είναι το γραμμοσκιασμένο εμβαδόν σε κάθε περίπτωση. Όταν οι κατανομές είναι πιο μακριά η μια από την άλλη η πιθανότητα αυτή μικραίνει.



Γράφημα 10.2. Πιθανότητα λανθασμένης κατάταξης για δύο παραδείγματα με μονοδιάστατες κατανομές .

Ας υπενθυμίσουμε πως σε όλα τα παραπάνω έχουμε υποθέσει κανονικότητα των δεδομένων και ισότητα των πινάκων συνδιακυμάνσεων. Οι παραπάνω υποθέσεις πολλές

φορές δεν είναι ρεαλιστικές. Η ισότητα των πινάκων διακυμάνσεων μπορεί να ξεπεραστεί εφαρμόζοντας παρόμοια διαδικασία αλλά καταλήγοντας σε πιο σύνθετη μορφή διαχωριστικής συνάρτησης.

Μέχρι εδώ έχουμε υποθέσει γνωστούς πληθυσμούς και γνωστές κατανομές δηλαδή ότι γνωρίζουμε εκ των προτέρων ότι ο ένας πληθυσμός είναι κανονικός με μέση τιμή $\boldsymbol{\mu}_1$και πίνακα συνδιακυμάνσεων $\boldsymbol{\Sigma}$ (δηλαδή $N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$) ενώ ο δεύτερος πληθυσμός είναι κανονικός με μέση τιμή $\boldsymbol{\mu}_2$και πίνακα συνδιακυμάνσεων $\boldsymbol{\Sigma}$ (δηλαδή $N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$). Στην πράξη τα $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2$ και $\boldsymbol{\Sigma}$ είναι άγνωστα και τα εκτιμούμε από τις δειγματικές μέσες τιμές $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ και τη συνδυασμένη (pooled) εκτίμηση του $\boldsymbol{\Sigma}$, $\mathbf{S}_p$ η οποία δίδεται ως $\mathbf{S}_p=\omega_1\mathbf{S}_1+\omega_2\mathbf{S}_2$ με $\mathbf{S}_k$τη εκτίμηση του πίνακα διακυμάνσεων της k ομάδας, $\omega_k=(n_k-1)/(n_1+n_2-2)$ και n_k το μέγεθος του δείγματος της k ομάδας, για k=1,2. Στην περίπτωση περισσότερων από δύο ομάδες ο τύπος της σταθμισμένης διακύμανσης διαμορφώνεται αντίστοιχα ως σταθμικός μέσος των επιμέρους διακυμάνσεων.

Αν δεν υπάρχει πληροφορία σχετικά με τις εκ των προτέρων πιθανότητες, αυτές μπορούν να εκτιμηθούν από τις σχετικές συχνότητες κάθε ομάδας, δηλαδή εκτιμούμε το π_k από το λόγο n_k/n. Θα πρέπει να τονιστεί εδώ πως αυτό υπονοεί τυχαία δειγματοληψία, κάτι που δυστυχώς δεν είναι πάντα αληθές. Πολλές φορές σε προβλήματα κατάταξης ο ερευνητής παίρνει ξεχωριστά τα δείγματα από κάθε πληθυσμό. Για παράδειγμα μπορεί ο τρόπος δειγματοληψίας να απαιτεί τυχαίο δείγμα ίσου μεγέθους από κάθε πληθυσμό. Αυτό δεν αποτελεί ένδειξη σχετικά με την εκ των προτέρων πιθανότητα κάθε πληθυσμού.

Άρα μέχρι τώρα χρησιμοποιήσαμε n παρατηρήσεις για να βρούμε ένα γραμμικό μετασχηματισμό των δεδομένων $U(\mathbf{x})$ ο οποίος μεγιστοποιεί την πιθανοφάνεια και κάνει μέγιστο το t τεστ για την σύγκριση των δύο ομάδων. Στην πράξη όταν προσέλθει ένα καινούριο άτομο / παρατήρηση υπολογίζουμε το σκορ $U(\mathbf{x})$, το συγκρίνουμε με την κρίσιμη τιμή k_0 και το κατατάσσουμε ανάλογα.

Προβλήματα μπορεί να εμφανιστούν λόγω της μη κανονικότητας των πληθυσμών. Άλλες όμως κατανομές δυσχεραίνουν πολύ τους υπολογισμούς. Επίσης ο προσδιορισμός του k_0 φαίνεται να έχει προβλήματα.

Σε αυτό το σημείο θα πρέπει να τονιστεί το εξής. Ο συντελεστής $\mathbf{L}$ δεν είναι μοναδική λύση για την κατασκευή του ίδιου διαχωριστικού κανόνα. Έτσι αν θέσουμε $\mathbf{L}^*=c\mathbf{L}$ όπου c μια σταθερή θετική ποσότητα τότε ο παραπάνω κανόνας γίνεται ισοδύναμα: *ταξινόμηση στην 1η ομάδα αν* $\mathbf{L}^{*'}\mathbf{x}_{(i)}-\frac{1}{2}\mathbf{L}^{*'}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\geq ck_0$ *αλλιώς ταξινόμηση στη 2η* , δηλαδή απλά αλλάζει το κρίσιμο σημείο, τα κεντροειδή (μέσες τιμές) και οι διακυμάνσεις της κάθε ομάδας. Για αυτό το λόγο, προκειμένου να προκύπτει μια μοναδική λύση θέτουμε κάποιον περιορισμό, μια τυποποίηση δηλαδή. Συνήθης τυποποίηση δίδεται για $c=1/\sqrt{\mathbf{L}'\mathbf{L}}$.

10.4 Η Λογική της Διαχωριστικής Συνάρτησης του Fisher

Ο διαχωριστικός κανόνας του Fisher βασίζεται στην μετατροπή των χαρακτηριστικών **x** σε μονοδιάστατα *σκορ* μέσω μιας συνάρτησης η οποία λέγεται διαχωριστική συνάρτηση (discriminant function). Τα *σκορ* των δύο ομάδων θα πρέπει να είναι όσο το δυνατόν πιο απομακρυσμένα έτσι ώστε να μπορούμε εύκολα με βάση αυτά τα *σκορ* να κάνουμε διαχωρισμό και ταξινόμηση των δύο ομάδων. Έτσι λοιπόν ο Fisher πρότεινε τη χρήση γραμμικών συνδυασμών για τη δημιουργία αυτών των *σκορ* χωρίς να γίνει κάποια υπόθεση για την κατανομή των ομάδων. Η γραμμικότητα υιοθετήθηκε για λόγους ευκολίας. Παρόλα αυτά υπέθεσε ισότητα των πινάκων συνδιακύμανσης αφού χρησιμοποίησε τη συνδυασμένη κοινή (pooled) εκτίμηση $\mathbf{S}_p$.

Έστω λοιπόν ότι τα *σκορ* δίνονται ως U_1 για την 1η ομάδα και ως U_2 για τη 2η ομάδα. Τότε ένα μέτρο του πόσο κοντά είναι τα *σκορ* των δύο ομάδων δίνεται από την απόσταση των μέσων τιμών ($\overline{U}_1$- $\overline{U}_2$). Ο Fisher μέτρηση αποστάσεις σε τυπικές αποκλίσεις και κατά απόλυτες τιμές, δηλαδή πήρε σαν μέτρο απόστασης των δύο ομάδων την ποσότητα: $D = \frac{|\overline{U}_1 - \overline{U}_2|}{S_U}$ με $S_U = \frac{\sum_{i\in G_1}(U_i - \overline{U}_1)^2 + \sum_{i\in G_2}(U_i - \overline{U}_2)^2}{n_1 + n_2 - 2}$ όπου $i\in G_i$ σημαίνει ότι λαμβάνουμε υπόψη τις παρατηρήσεις που ανήκουν στην i ομάδα. Σκοπός είναι να μεγιστοποιήσουμε την ποσότητα D ή αντίστοιχα την απόσταση D^2 καθώς αυτό σημαίνει πως τα *σκορ* των δύο ομάδων θα είναι όσο γίνεται πιο διαφορετικά μεταξύ τους

Έστω ο γραμμικός συνδυασμός $\mathbf{L}'\mathbf{x}$ τότε πρέπει να μεγιστοποιήσουμε την ποσότητα

$$D^2 = \frac{[\mathbf{L}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2}{\mathbf{L}'\mathbf{S}_p\mathbf{L}} .$$

Από την ανισότητα Cauchy – Schwarz έχουμε ότι για κάθε p×1 διανύσματα **a** και **b** ισχύει ότι $(\mathbf{a}'\mathbf{b})^2 \le (\mathbf{a}'\mathbf{a})\ (\mathbf{b}'\mathbf{b})$. Εφόσον ο πίνακας συνδιακυμάνσεων είναι θετικά ορισμένος μπορούμε να θέσουμε $\mathbf{a}=\mathbf{S}_p^{1/2}\mathbf{L}$ και $\mathbf{b}=\mathbf{S}_p^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ τότε έχουμε

$$[\mathbf{L}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2 \le (\mathbf{L}'\mathbf{S}_p^{1/2}\mathbf{S}_p^{1/2}\mathbf{L})\ [(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1/2}\mathbf{S}_p^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)] \Leftrightarrow$$

$$[\mathbf{L}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2 \le (\mathbf{L}'\mathbf{S}_p\mathbf{L})\ [(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)] \Leftrightarrow$$

$$D^2 = \frac{[\mathbf{L}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2}{\mathbf{L}'\mathbf{S}_p\mathbf{L}} \le (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Άρα για $\mathbf{L} = c\ \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, όπου c>0, έχουμε $D^2=(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ δηλαδή τη μέγιστη απόσταση μεταξύ των μέσων και τον καλύτερο δυνατόν διαχωρισμό. Το c είναι μια σταθερά και συνήθως παίρνουμε c=1. Ο διαχωριστικός κανόνας ολοκληρώνεται ορίζοντας την κρίσιμη τιμή η οποία δεν είναι άλλη από την μέση τιμή των $\overline{U}_1$ και $\overline{U}_2$ δηλαδή η ποσότητα

$m= (\overline{U}_1 + \overline{U}_2)/2 = \mathbf{L}'(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$

η οποία ισαπέχει από τα $\overline{U}_1$ και $\overline{U}_2$. Έτσι ο διαχωριστικό κανόνας γίνεται:

αν $\boldsymbol{L}'\boldsymbol{x} \geq m$ *(*$\boldsymbol{L}'\boldsymbol{x}-m \geq 0$*) τότε κατατάσσουμε στην 1η ομάδα αλλιώς στην 2η.*

Η ποσότητα $\boldsymbol{L}'\boldsymbol{x}-m$ είναι όμως ίση με τη διαχωριστική συνάρτηση που βρήκαμε χρησιμοποιώντας τη θεωρία αποφάσεων υποθέτοντας κανονικές κατανομές $k_0=0$. Συνεπώς και οι δύο διαχωριστικοί κανόνες που προέκυψαν με διαφορετική λογική ταυτίζονται.

10.5 Γενίκευση Διαχωριστικής Ανάλυσης σε Κ ομάδες

Μέχρι τώρα παρουσιάσαμε την προσέγγιση υποθέτοντας την ύπαρξη μόνο δύο ομάδων. Όπως είπαμε ο διαχωριστικός κανόνας ελαχιστοποίησης του κόστους λανθασμένης ταξινόμησης όταν έχουμε δύο ομάδες είναι:

κατατάσσουμε την i παρατήρηση στην 1η ομάδα αν $\frac{f_1(\mathbf{x}_i)}{f_2(\mathbf{x}_i)} \geq \frac{\pi_2}{\pi_1} \times \frac{C(1|2)}{C(2|1)}$ *αλλιώς κατατάσσουμε την i παρατήρηση στη 2η ομάδα.*

Για να γενικεύσουμε τη μέθοδο σε διαχωρισμό Κ ομάδων πρέπει να υπολογίσουμε τα σκορ:

$$W_k = -\sum_{l=1}^{K} \pi_l C(k|l) f_l(\mathbf{x})$$

τα οποία αντιστοιχούν σε ελαχιστοποίηση του συνολικού κόστους λανθασμένης κατάταξης, και καταχωρούμε την παρατήρηση μας στην ομάδα με το μεγαλύτερο σκορ. Στην περίπτωση των κανονικών κατανομών και όταν έχουμε ίσα κόστη μπορούμε εναλλακτικά να υπολογίσουμε τα σκορ $W_k=\mathbf{L}_k'\mathbf{x}-\frac{1}{2}\mathbf{L}_k'\bar{\mathbf{x}}_k+\ln(\pi_k)$ για k=1,2,…,K, όπου Κ ο αριθμός των ομάδων, $\bar{\mathbf{x}}_k$ ο δειγματικός μέσος της k ομάδας, $\mathbf{L}_k=\mathbf{S}_p^{-1}\bar{\mathbf{x}}_k$ και $\mathbf{S}_p$ είναι ο κοινός συνδυασμένος εκτιμητής του πίνακα διακύμανσης-συνδιακύμανσης που δίδεται ως

$$\mathbf{S}_p=\omega_1\mathbf{S}_1+\omega_2\mathbf{S}_2+\ldots+\omega_K\mathbf{S}_K$$

με $\mathbf{S}_k$ την εκτίμηση του πίνακα διακυμάνσεων της k ομάδας, $\omega_k=(n_k-1)/(n-K)$ και n_k ο αριθμός των παρατηρήσεων στην k ομάδα. Οι γραμμικές συναρτήσεις W_k λέγονται και γραμμικές διαχωριστικές συναρτήσεις του Fisher και οι τιμές που τελικά παίρνουν λέγονται σκορ των διαχωριστικών συναρτήσεων του Fisher. Τα κεντροειδή τους είναι οι αντίστοιχες μέσες τιμές ενώ οι κανονικοποιημένες διαχωριστικές συναρτήσεις είναι μειωμένες κατά μια (δηλαδή K-1) και είναι ανάλογες των διαφορών $Z_k= W_k-W_K$ για k=1,2,…,K-1 (δηλαδή σε κάθε περίπτωση συγκρίνουν τη κάθε ομάδα με κάποια βασική ομάδα η οποία συνήθως είναι η τελευταία ή η πρώτη).

10.6 Γενίκευση Διαχωριστικής Ανάλυσης του Fisher σε Κ ομάδες

Ο Fisher εναλλακτικά πρότεινε μια επέκταση της μεθόδου του για τα διαχωρισμό Κ ομάδων. Έτσι λοιπόν προτείνει τη χρήση Κ-1 γραμμικών συνδυασμών της μορφής $\mathbf{L}_k'\mathbf{x}$ με $\mathbf{L}_k$ να είναι τα διανύσματα του πίνακα $\mathbf{\Delta} = (n-K)\ \mathbf{S}_p^{-1}\mathbf{W}$ (υπό τον περιορισμό ότι $\mathbf{L}'\mathbf{S}_p\mathbf{L}=1$) με σειρά που αντιστοιχεί στο μέγεθος των ιδιοτιμών. Δηλαδή $\mathbf{L}_1$ είναι το ιδιοδιάνυσμα που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή, $\mathbf{L}_2$ είναι το ιδιοδιάνυσμα που αντιστοιχεί στην 2η μεγαλύτερη ιδιοτιμή κ.ο.κ. Όπου $\mathbf{W} = \sum_{k=1}^{K} n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'$ είναι ένα μέτρο της διακύμανσης των μέσων τιμών των Κ ομάδων. Σημείωση ότι $\mathbf{L}_1$ μεγιστοποιεί την ποσότητα $D^2 = (n-K)\ \mathbf{L}'\mathbf{W}\mathbf{L} / \mathbf{L}'\mathbf{S}_p\mathbf{L}$

η οποία είναι ένα μέτρο της απόστασης μεταξύ των μέσων δηλαδή ένα μέτρο διαχωρισμού των ομάδων κατά αντιστοιχία με την απόσταση που είχαμε όταν Κ=2.

Η ερμηνεία των παραπάνω διαχωριστικών συναρτήσεων είναι ότι η 1η διαχωριστική συνάρτηση μεγιστοποιεί τις διαφορές των μέσων σε μια διάσταση. Η 2η διαχωριστική συνάρτηση μεγιστοποιεί την απόσταση των μέσων σε μια κατεύθυνση ορθογώνια στην 1η, η 3η μας δείχνει την απόσταση σε μια 3η διάσταση ανεξάρτητη των άλλων 2 κ.ο.κ. Μπορούμε να περιγράψουμε τις διαχωριστικές συναρτήσεις σαν παράγοντες (f*actors*) που διαχωρίζουν βέλτιστα τα κεντροειδή (μέσες τιμές) σε σχέση με τη διασπορά μέσα σε κάθε ομάδα.

Ο διαχωριστικός κανόνας αν κρατήσουμε r διαχωριστικές συναρτήσεις γίνεται:

Ταξινομούμε την παρατήρηση x στην k ομάδα αν $\sum_{l=1}^{r}[\mathbf{L}_l(\mathbf{x}-\bar{\mathbf{x}}_k)]^2 \le \sum_{k=l}^{r}[\mathbf{L}_l(\mathbf{x}-\bar{\mathbf{x}}_i)]^2$ *για όλα τα i διαφορετικά του k.*

10.7 Άλλες Προσεγγίσεις για το Διαχωρισμό Ομάδων

Η διαχωριστική ανάλυση δεν είναι η μοναδική μέθοδος που προσπαθεί να κατατάξει τις παρατηρήσεις σε ομάδες. Υπάρχουν πολλές άλλες μέθοδοι που κάνουν, ή τουλάχιστον σκοπό έχουν να κάνουν, το ίδιο. Πολλές από αυτές τις μεθόδους, δεδομένου πως αναπτύχθηκαν με βάση άλλες επιστήμες, δεν διαθέτουν σημαντικό στατιστικό υπόβαθρο. Παρόλα αυτά εμφανίζουν πολύ καλά εμπειρικά αποτελέσματα.

Άλλες προσεγγίσεις που μπορούμε να χρησιμοποιήσουμε εναλλακτικά είναι

- η λογιστική παλινδρόμηση (logistic regression),
- τα δένδρα παλινδρόμησης και ταξινόμησης (CART: classification and regression trees)
- τα νευρωνικά δίκτυα (neural networks)

- Διάφοροι αλγόριθμοι μάθησης που συνδυάζουν στοιχεία από όλα τα παραπάνω

Η λογιστική παλινδρόμηση στην ουσία είναι γενίκευση της απλής γραμμικής παλινδρόμησης για την περίπτωση όπου η εξαρτημένη μεταβλητή Υ είναι δίτιμη (δηλαδή 0:αποτυχία, 1:επιτυχία). Σε αυτή την περίπτωση έχουμε

$Y_i \sim$ Binomial (p_i , N_i) με $\ln[p_i/(1- p_i)] = \beta_0+\beta_1X_{1i}+\ldots+\beta_pX_{pi}$

όταν τα δεδομένα δίνονται ως αριθμός επιτυχιών Y_i σε σύνολο N_i πειραμάτων ή

$Y_i \sim$ Bernoulli(p_i) με $\ln[p_i/(1- p_i)] = \beta_0+\beta_1X_{1i}+\ldots+\beta_pX_{pi}$

όταν η Y_i υποδεικνύει σε ποια ομάδα ανήκει η i παρατήρηση. Από τα παραπάνω μπορούμε να υπολογίσουμε την πιθανότητα για κάθε παρατήρηση να ανήκει στην $1^η$ ή στη $2^η$ ομάδα η οποία είναι δίδεται ως εξής

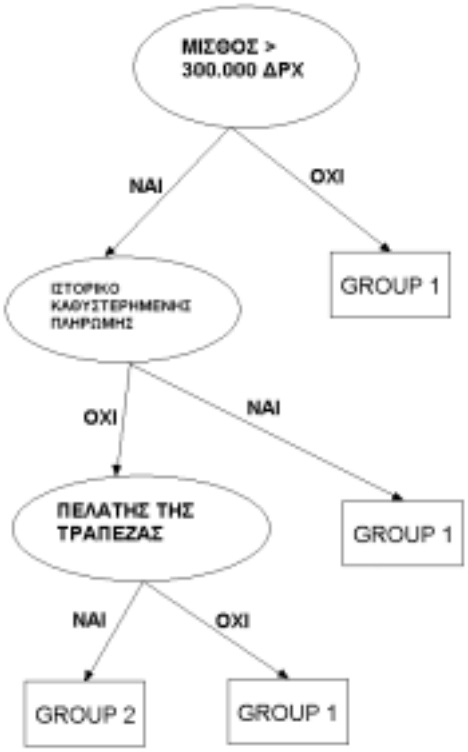
$$p_i = \frac{e^{\beta_0+\beta_1X_{1i}+\ldots+\beta_pX_{pi}}}{1+e^{\beta_0+\beta_1X_{1i}+\ldots+\beta_pX_{pi}}}.$$

Αν πάρουμε ίσα κόστη και εκ-των-προτέρων πιθανότητες τότε κατατάσσουμε στην $2^η$ ομάδα ($Y=1$) αν $p_i \geq 0$ αλλιώς στη $1^η$ ομάδα($Y=0$). Η παραπάνω προσέγγιση μπορεί να επεκταθεί σε Κ ομάδες μέσω της πολυωνυμικής λογιστικής παλινδρόμησης (multinomial logistic regression). Η σχέση λογιστικής παλινδρόμησης και διαχωριστικής ανάλυσης είναι πολύ μεγάλη. Ειδικά για την περίπτωση με 2 ομάδες τα αποτελέσματα είναι αρκετά όμοια (εξαρτάται βέβαια και από τις υποθέσεις που έχουν γίνει για τον πληθυσμό). Τα μοντέλα λογιστικής παλινδρόμησης έχουν το πλεονέκτημα πως αυτόματα υπολογίζουν τις πιθανότητες κάθε ομάδας κάτι που μόνο έμμεσα μπορεί να γίνει με τη διαχωριστική ανάλυση. Από την άλλη μεριά η διαχωριστική ανάλυση στηρίζεται σε ρεαλιστικότερες μεθόδους και είναι υπολογιστικά απλούστερη.

Τα δένδρα παλινδρόμησης και κατάταξης συνδέονται περισσότερο με την ανάλυση σε ομάδες παρά με τη διαχωριστική ανάλυση. Η μέθοδος ξεκινάει με όλες τις παρατηρήσεις σε μια ομάδα και «σπάει» το δείγμα σε ομάδες ανάλογα με τα χαρακτηριστικά τους όπως για παράδειγμα Ηλικία>45. Η διαδικασία συνεχίζεται μέχρι ένας κανόνας παύσης ικανοποιηθεί. Η μέθοδος αναπτύχθηκε κυρίως από τους Breiman et al. (1984) και είναι διαθέσιμη στο στατιστικό πακέτο Splus.

Παράδειγμα:

Έστω ότι μια τράπεζα έχει στη διάθεση της τις μεταβλητές: μισθό, ιστορικό κακής πληρωμής και το αν είναι πελάτης ή όχι και θέλει να εξετάσει σε ποιους μελλοντικούς πελάτες θα πρέπει να δώσει δάνειο. Τότε ένα δένδρο παλινδρόμησης και ταξινόμησης θα είναι δυνατό να δίνεται ως εξής:



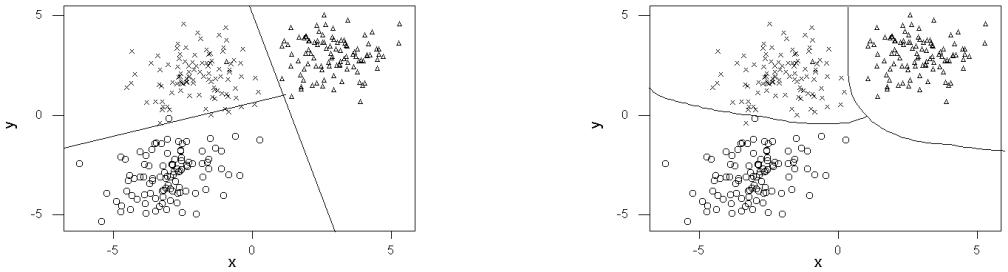
Γράφημα 10.3 Ένα δένδρο παλινδρόμησης και ταξινόμησης

Από το παραπάνω δένδρο βλέπουμε ότι έχουμε 2 ομάδες. Η τράπεζα μπορεί να αποφασίσει τελικά να δίνει δάνειο αν ο μισθός είναι τουλάχιστον 300.000 δρχ, αν δεν έχει ιστορικό καθυστερημένης πληρωμής και αν είναι πελάτης. Συνήθως στα δέντρα ταξινόμησης χρειάζεται να διακριτοποίησουμε τις συνεχείς μεταβλητές ώστε να δουλέψουν καλύτερα οι αλγόριθμοι και αυτό μπορεί να οδηγήσει σε χάσιμο πληροφορίας

Τέλος τα νευρωνικά δίκτυα μια εντατική υπολογιστικά προσέγγιση η οποία μετατρέπει εισερχόμενη πληροφορία σε επιθυμητή εξερχόμενη πληροφορία. Η επεξεργασία της πληροφορίας βασίζεται σε συνδυασμένα δίκτυα μικρών επεξεργαστικών ομάδων οι οποίοι λέγονται νευρώνες (neuron) ή κόμβοι (nodes). Τα νευρωνικά δίκτυα αποτελούν μια απλοποιημένη εφαρμογή του τρόπου λειτουργίας του ανθρώπινου μυαλού. Τρία είναι τα βασικά συστατικά ενός νευρωνικού δικτύου: οι κόμβοι, ο τρόπος σύνδεσης τους και ο αλγόριθμος με τον οποίο βρίσκουμε τις τιμές των παραμέτρων του δικτύου. Τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν για διαχωρισμό ομάδων με επιδόσεις ανάλογες της λογιστικής παλινδρόμησης και της διαχωριστικής ανάλυσης.

Επίσης να τονιστεί πως η διαχωριστική ανάλυση όπως παρουσιάστηκε προηγουμένως περιορίστηκε σε πολυμεταβλητές κανονικές κατανομές πληθυσμών και γραμμικές διαχωριστικές συναρτήσεις. Προφανώς μπορεί κανείς να χρησιμοποιήσει πιο γενικά μοντέλα με κόστος τη μεγαλύτερη πολυπλοκότητα των κανόνων που θα σχηματίσει.

Στο γράφημα 10.4 μπορεί κανείς να δει ένα παράδειγμα μη γραμμικής διαχωριστικής ανάλυσης. Αριστερά παρουσιάζεται μια γραμμική διαχωριστική ανάλυση. Στην ουσία οι διαχωριστικές συναρτήσεις που είδαμε πριν χωρίζουν το επίπεδο σε διάφορα μέρη. Στο παράδειγμα έχουμε χρησιμοποιήσει 2 μεταβλητές για να είναι δυνατή η απεικόνιση τους. Η ανάλυση ουσιαστικά χωρίζει το επίπεδο σε 3 μέρη, τις 3 ομάδες. Στο δεξί μέρος βλέπουμε τετραγωνικές διαχωριστικές συναρτήσεις για αυτό πια το επίπεδο χωρίζεται από καμπύλες. Αυτό αποτελεί ένα απλό παράδειγμα μη γραμμικής διαχωριστικής ανάλυσης



Γράφημα 10.4 Γραμμική και μη διαχωριστική ανάλυση σε προσομοιωμένα δεδομένα.

10.8 Άλλα θέματα

Περιγράψαμε συνοπτικά μερικά αρχικά θέματα σχετικά με τη διαχωριστική ανάλυση. Σε αυτή την ενότητα θα προσπαθήσουμε να παρουσιάσουμε εν συντομία μερικά θέματα που σχετίζονται με τη διαχωριστική ανάλυση

10.8.1 Καλή προσαρμογή του μοντέλου

Έχοντας βρει τις διαχωριστικές συναρτήσεις και πριν τις χρησιμοποιήσουμε για την κατάταξη νέων παρατηρήσεων είναι χρήσιμο να αξιολογήσουμε κατά πόσο τα αποτελέσματα είναι ικανά για τη σωστή κατάταξη νέων παρατηρήσεων.

Ένας απλό τρόπος είναι να χρησιμοποιήσουμε τα αποτελέσματα για να κατατάξουμε τις παρατηρήσεις του δείγματος και να βρούμε το ποσοστό των σωστών κατατάξεων. Στην περίπτωση της τέλειας διαμέρισης τότε περιμένουμε το ποσοστό των σωστών προβλέψεων να είναι 1. Όμως αυτή η προσέγγιση έχει πρόβλημα καθώς χρησιμοποιούμε τις ίδιες παρατηρήσεις να φτιάξουμε το μοντέλο και για να δούμε την ικανότητα του. Με αυτό τον τρόπο συνήθως υπερεκτιμάμε την ικανότητα του μοντέλου. Εναλλακτικά μπορούμε να χρησιμοποιήσουμε resampling τεχνικές για να εκτιμήσουμε το ποσοστό επιτυχίας της ανάλυσης. Για παράδειγμα με τη χρήση cross-validation χωρίζουμε

το δείγμα σε δύο κομμάτια, το ένα είναι το δείγμα μάθησης (training set) και το άλλο το δείγμα επικύρωσης (test set). Έτσι βρίσκουμε το μοντέλο χρησιμοποιώντας τις παρατηρήσεις της πρώτης ομάδας και στη συνέχεια προβλέπουμε τις τιμές των μελών της δεύτερης. Μπορούμε να επαναλάβουμε τη διαδικασία πολλές φορές. Επίσης το ποσοστό των παρατηρήσεων που θα πάνε σε κάθε δείγμα εξαρτάται από την εφαρμογή, το μέγεθος του δείγματος και άλλα χαρακτηριστικά.

Εναλλακτικά μπορεί κανείς να πάρει bootstrap δείγματα από τα δεδομένα και να χρησιμοποιήσει για πρόβλεψη τις παρατηρήσεις που δεν είναι στο δείγμα.

10.8.2 Μη παραμετρική διαχωριστική ανάλυση

Μέχρι τώρα είδαμε την περίπτωση όπου κάθε πληθυσμός ακολουθούσε μια πολυμεταβλητή κανονική κατανομή. Στη μη παραμετρική διαχωριστική ανάλυση δεν χρειάζεται να υποθέσουμε τίποτα για την κατανομή του πληθυσμού αφού και την κατανομή θα την εκτιμήσουμε από τα δεδομένα. Θα πρέπει να τονιστεί πως και στην περίπτωση της παραμετρικής διαχωριστικής ανάλυσης με τη χρήση πολυμεταβλητών κανονικών κατανομών οι παράμετροι εκτιμούνται από το δείγμα. Ως μη παραμετρική ανάλυση εννοούμε πως δεν έχουμε κάποια παραμετρική μορφή για την κατανομή του πληθυσμού αλλά την εκτιμάμε μη παραμετρικά με τη χρήση της μεθόδου των kernels.

Έστω πως τα δεδομένα μας είναι μονοδιάστατα. Η εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας με τη χρήση του kernel $K(x)$ δίνεται ως

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\,K\left(\frac{x-x_i}{h}\right)$$

όπου $K(x)$ είναι μια συνάρτηση (συνήθως συνάρτηση πυκνότητας πιθανότητας) για την οποία ισχύει $K(x)\geq 0$ για κάθε x, $\int_{-\infty}^{+\infty}K(x)dx=1$, $\int_{-\infty}^{+\infty}xK(x)dx=0$ και $0<\int_{-\infty}^{+\infty}x^2K(x)dx<\infty$. Χωρίς να είναι απαραίτητο, είναι όμως χρήσιμο, είναι συμμετρική ως προς το 0. Το h είναι μια παράμετρος λείανσης (smoothing parameter) που ονομάζεται και παράθυρο που ουσιαστικά καθορίζει πόσο απότομη ή όχι θα είναι η εκτιμήτρια αλλά και από ποιες παρατηρήσεις θα ληφθεί πληροφορία για την εκτίμηση. Συνήθεις μορφές kernels αποτελούν η τυποποιημένη κανονική κατανομή και η συνάρτηση του Epaneshnikov που έχει κάποιες πολύ καλές ιδιότητες.

Δεν θα πούμε περισσότερα για την μέθοδο αυτή απλά θα επεκτείνουμε την ιδέα σε πολυμεταβλητά δεδομένα. Έτσι μια πολυμεταβλητή κατανομή εκτιμάται ως

$$\hat{f}(x_1,x_{2,},...,x_d)=\hat{f}(\mathbf{x})=\frac{1}{n|H|}\sum_{i=1}^{n}K_d\left[\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)\right],$$

όπου τώρα $\mathbf{x}_i$ είναι οι παρατηρήσεις μας που είναι διανύσματα και $\mathbf{H}$ είναι ο πίνακας που περιέχει τα παράθυρα προς όλες τις διαστάσεις (μεταβλητές) που χρησιμοποιούμε. Όπως καταλαβαίνετε ο παραπάνω τύπος είναι σε μια πολύ γενική μορφή. Αυτά που πρέπει να καθορίσουμε είναι ο πίνακας $\mathbf{H}$ και η μορφή του kernel**.** Συνήθως ως kernel χρησιμοποιούμε πολυμεταβλητές κανονικές κατανομές με μέσο το μηδενικό διάνυσμα και μοναδιαίο πίνακα διακύμανσης, ενώ το παράθυρο είναι συνήθως ένας διαγώνιος πίνακας.

Έχοντας λοιπόν εκτιμήσει τις κατανομές των πληθυσμών μπορούμε να χρησιμοποιήσουμε ως κανόνες κατάταξης βασισμένα στις εκ των υστέρων πιθανότητες. Αν δηλαδή $\hat{f}_k(\mathbf{x})$ είναι μια εκτίμηση της πυκνότητας για τον k πληθυσμό τότε κατατάσσουμε

$\pi_k \hat{f}_k(\mathbf{x}) \geq \pi_j \hat{f}_j(\mathbf{x})$ για κάθε j.

10.8.3 Σχέση με την ανάλυση κατά συστάδες

Στο προηγούμενο κεφάλαιο μιλήσαμε για την ανάλυση σε συστάδες. Αν χρησιμοποιήσουμε το πιθανοθεωρητικό μοντέλο (model based clusteping) τότε κάθε παρατήρηση έχει συνάρτηση πυκνότητας.

Αν πάρουμε τυχαία ένα άτομο από τον πληθυσμό αυτό και δεν γνωρίζουμε από ποιο υποπληθυσμό προέρχεται τότε από το θεώρημα ολικής πιθανότητας η κατανομή του θα είναι

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

όπου $0 < p_j < 1, \sum_{j=1}^{k} p_j = 1$ δηλώνει την πιθανότητα ένα τυχαίο άτομο να ανήκει στον υποπληθυσμό *j* και $f(x \mid \theta_j)$ η κατανομή του *j* πληθυσμού.

Σκοπός της ανάλυσης είναι η εκτίμηση των παραμέτρων αλλά και η εκτίμηση της μη παρατηρήσιμης μεταβλητής z_{ij} που παίρνει την τιμή 1 αν η i παρατήρηση ανήκει στον j πληθυσμό και 0 αν όχι. Στη διαχωριστική ανάλυση, οι τιμές των z_{ij} είναι γνωστές και δεν χρειάζεται να τις εκτιμήσουμε. Για την ακρίβεια ο τρόπος εκτίμησης είναι ο ίδιος και για τις 2 μεθόδους.

Είναι πολύ ενδιαφέρουσα η περίπτωση όπου για μερικές παρατηρήσεις γνωρίζουμε σε ποια ομάδα ανήκουν ενώ σε άλλες όχι επειδή οι τιμές δεν έχουν παρατηρηθεί. Δηλαδή ουσιαστικά χρειαζόμαστε διαχωριστική ανάλυση για να φτιάξουμε έναν κανόνα και να προβλέψουμε την ομάδα για τις άλλες παρατηρήσεις. Χρησιμοποιώντας όμως όλες τις παρατηρήσεις μπορούμε να εκτιμήσουμε τις ομάδες που ανήκει κάθε παρατήρηση συνδυάζοντας ουσιαστικά στοιχεία από την ανάλυση κατά συστάδες και τη διαχωριστική ανάλυση.

10.9 Διαχωριστική ανάλυση με το SPSS

10.9.1 Ένα Απλό Παράδειγμα

Για να εξηγήσουμε την εφαρμογή της διαχωριστικής ανάλυσης μέσω του SPSS θα χρησιμοποιήσουμε το ακόλουθο απλό παράδειγμα:

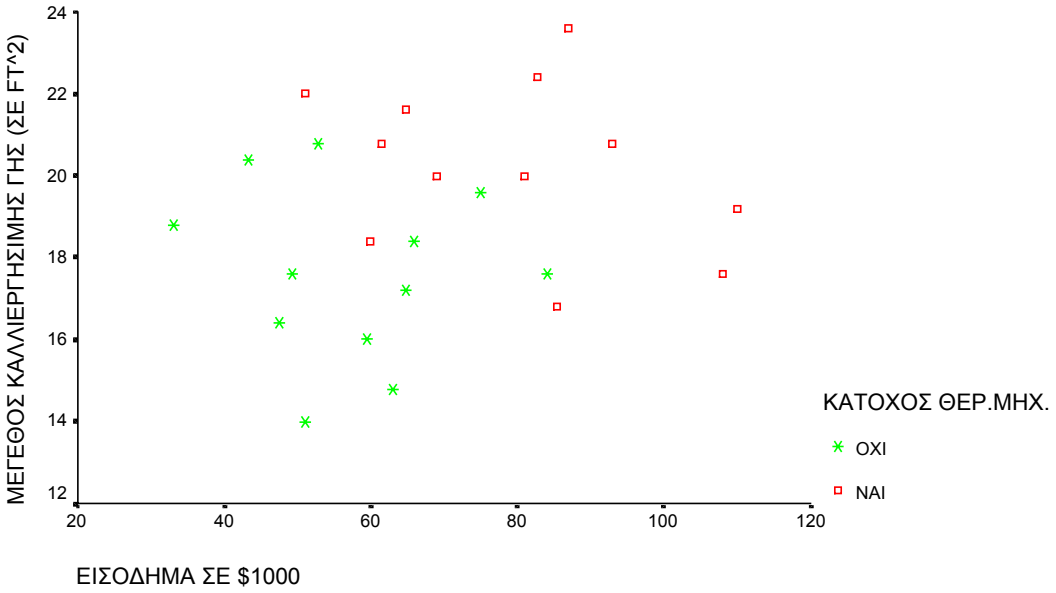
Έστω μια εταιρία κατασκευής θεριστικών μηχανών η οποία θέλει να προσδιορίσει τις προοπτικές των πωλήσεων της. Για το λόγο αυτό ενδιαφέρεται να ταξινομήσει τον αγροτικό πληθυσμό σε μελλοντικό ή όχι κάτοχο θεριστικής μηχανής με βάση το εισόδημα του και την έκταση της καλλιεργήσιμης γης που κατέχει. Έτσι παίρνει τυχαία ένα δείγμα από 12 κατόχους και 12 μη κατόχους αγρότες ώστε προχωρήσει στη δημιουργία ενός διαχωριστικού κανόνα κατάταξης των υποψήφιων πελατών της. Το δείγμα δίδεται στον ακόλουθο πίνακα 10.1:

Κάτοχοι Μηχανής		Μη κάτοχοι Μηχανής	
Εισόδημα ($1000)	Μέγεθος γης (1000 ft^2)	Εισόδημα ($1000)	Μέγεθος γης (1000 ft^2)
60.0	18.4	75.0	19.6
85.5	16.8	52.8	20.8
64.8	21.6	64.8	17.2
61.5	20.8	43.2	20.4
87.0	23.6	84.0	17.6
110.1	19.2	49.2	17.6
108.0	17.6	59.4	16.0
82.8	22.4	66.0	18.4
69.0	20.0	47.4	16.4
93.0	20.8	33.0	18.8
51.0	22.0	51.0	14.0
81.0	20.0	63.0	14.8

Πίνακας 10.1. Τα δεδομένα του παραδείγματος

Στο αρχείο μας ονομάζουμε τις μεταβλητές μας ως INCOME (εισόδημα), LOTSIZE (μέγεθος καλλιεργήσιμης γης) και ως GROUP (την ομάδα των αγροτών).

Σαν πρώτο βήμα είναι χρήσιμο να κάνουμε ένα διάγραμμα σημείων μεταξύ των δύο μεταβλητών μας.

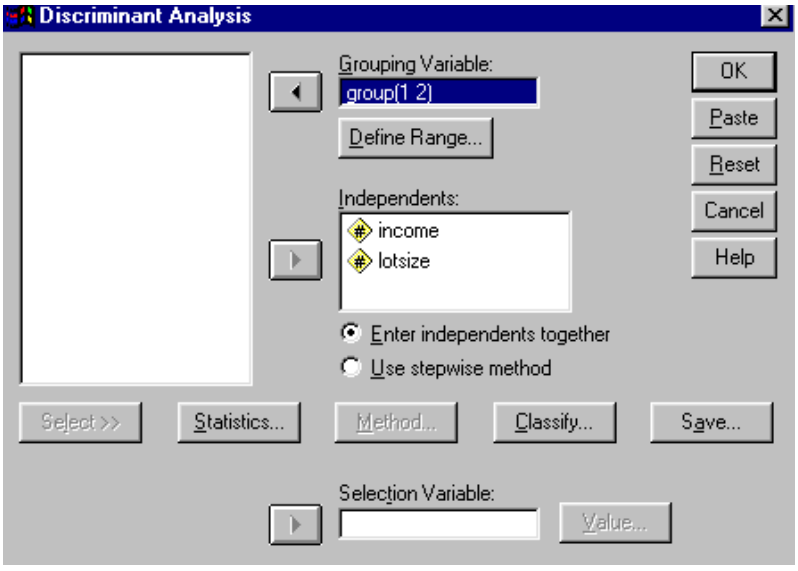


Γράφημα 10.5 Διάγραμμα σημείων για τα δεδομένα του παραδείγματος

Από το διάγραμμα βλέπουμε έναν σαφή διαχωρισμό των δύο ομάδων. Αγρότες με μεγάλο εισόδημα και μεγάλες εκτάσεις γης κατέχουν θεριστικές μηχανές. Θα προχωρήσουμε τώρα στην κατασκευή του διαχωριστικού κανόνα.

10.9.2 Ανάλυση του Παραδείγματος με το SPSS

Το βασικό μενού της Διαχωριστικής ανάλυσης στο SPSS μπορεί να βρεθεί στο μονοπάτι **Analyze>Classify>Discriminant** το οποίο έχει την ακόλουθη μορφή (εικόνα 10.1).



Εικόνα 10.1 Το βασικό menu για τη διαχωριστική ανάλυση στο SPSS

Τα κύρια ζητούμενα είναι:

Grouping Variable: Εδώ ζητείται η μεταβλητή που καθορίζει τις ομάδες. Στο παράδειγμα μας τοποθετούμε τη μεταβλητή GROUP η οποία μας υποδεικνύει εάν ο αγρότης είναι κάτοχος η όχι θεριστικής μηχανής. Επιπλέον πρέπει να ορίσουμε το εύρος των ομάδων στην επιλογή "Define Range". Εδώ ορίζουμε σαν ελάχιστη τιμή (minimum) το 1 και μέγιστη (maximum) το 2.

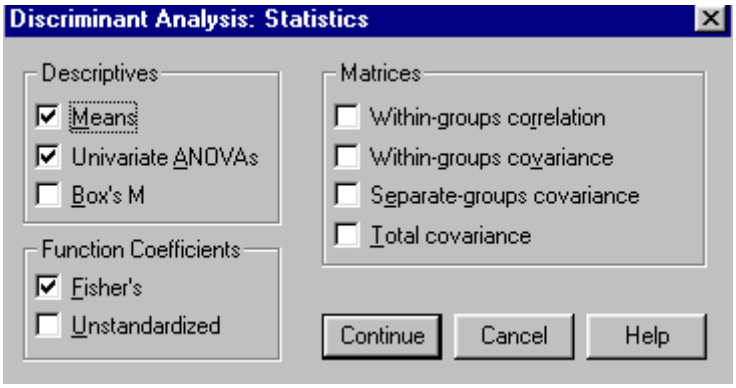
Independents: Σε αυτή τη θέση τοποθετούμε τις ανεξάρτητες μεταβλητές μας με βάση τις οποίες θα γίνει η ταξινόμηση της κάθε παρατήρησης. Οι μεταβλητές πρέπει να είναι ποσοτικές. Στο παράδειγμα μας οι μεταβλητές αυτές είναι το εισόδημα και το μέγεθος της καλλιεργήσιμης γης (INCOME και LOTSIZE).

Οι υπόλοιπες επιλογές περιλαμβάνουν τη χρήση όλων των ανεξάρτητων μεταβλητών (**Enter independents together**) ή εναλλακτικά τη χρήση κλιμακωτών μεθόδων επιλογής των ανεξάρτητων μεταβλητών (**Use Stepwise Method**). Η δεύτερη επιλογή είναι πολύ χρήσιμη στην πράξη και εντοπίζει βήμα – βήμα τις ασήμαντες μεταβλητές για το διαχωρισμό και τις αφαιρεί από τη διαχωριστική συνάρτηση. Στο παράδειγμα μας επιλέγουμε την πρώτη επιλογή καθώς έχουμε μόλις δύο μεταβλητές.

Ακολουθούν τα υπό-μενού **Select**, **Statistics**, **Method**, **Classify** και **Save**. Στο υπο-μενού **Select** μπορούμε να επιλέξουμε υπο-ομάδα των δεδομένων προς ανάλυση.

Επιλογές Περιγραφικών Δεικτών(Υπο-μενού Statistics)

Προχωράμε στο υπο-μενού **Statistics** το οποίο είναι το ακόλουθο (εικόνα 10.2):



Εικόνα 10.2 Το menu Statistics

Χωρίζεται σε τρεις κατηγορίες: Περιγραφικοί Δείκτες (**Descriptives**), Συντελεστές της διαχωριστικής συνάρτησης (**Function Coefficients**) και Μήτρες – Πίνακες (**Matrices**).

Στους περιγραφικούς δείκτες έχουμε τις εξής επιλογές: Εμφάνιση Μέσων Τιμών (**Means**), Ανάλυση διακύμανσης κατά ένα παράγοντα (**Univariate Anova's**) και μια στατιστική δοκιμασία (τεστ) για την ισότητα των πινάκων συνδιακύμανσης (**Box's M**).

Η επιλογή των μέσων μας δίνει περιγραφικούς δείκτες (μέγεθος δείγματος, μέση τιμή και τυπική απόκλιση). Στο παράδειγμά μας το αποτέλεσμα δίδεται από τον Πίνακα 10.3:

Group Statistics

	GROUP	Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
INCOME	1 NAI	79.4750	18.7788	12	12.000
	2 OXI	57.4000	14.1671	12	12.000
	Total	68.4375	19.7931	24	24.000
LOTSIZE	1 NAI	20.2667	2.0205	12	12.000
	2 OXI	17.6333	2.1129	12	12.000
	Total	18.9500	2.4283	24	24.000

Πίνακας 10.3. Περιγραφικά μέτρα για τις 2 ομάδες

στον οποίο φαίνονται εμφανώς οι διαφορές στις μέσες τιμές των δύο ομάδων. Η επιλογή της ανάλυσης διακύμανσης παράγει τον ακόλουθο πίνακα 10.4:

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
INCOME	.676	10.568	1	22	.004
LOTSIZE	.693	9.736	1	22	.005

Πίνακας 10.4. Έλεγχος διακυμάνσεων

από τον οποίο βλέπουμε ότι και για τις δύο μεταβλητές οι μέσες τιμές στις δύο ομάδες διαφοροποιούνται σημαντικά (p.values 0.004 και 0.005 < 0.05). Επιπλέον ο δείκτης λάμδα του Wilks μας δίνει χρήσιμες πληροφορίες για τις διαφορές των ομάδων. Ο δείκτης αυτός είναι το ποσοστό της διακύμανσης το οποίο δεν εξηγείται από το μοντέλο της ανάλυσης διακύμανσης κατά ένα παράγοντα. Κυμαίνεται από το μηδέν (0) έως το ένα (1). Τιμές κοντά στο μηδέν (0) υποδεικνύουν ισχυρές διαφορές ενώ τιμές κοντά στο ένα (1) υποδεικνύουν ότι δεν υπάρχουν διαφορές.

Η επιλογή του τεστ για την ισότητα των πινάκων συνδιακύμανσης (**Box's M**) παράγει τους ακόλουθους πίνακες 10.5:

Log Determinants

GROUP	Rank	Log Determinant
1 NAI	2	7.170
2 OXI	2	6.790
Pooled within-groups	2	7.030

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

(α)

Test Results

Box's M		1.102
F	Approx.	.331
	df1	3
	df2	87120.000
	Sig.	.803

Tests null hypothesis of equal population covariance matrices

(β)

Πίνακας 10.5. Έλεγχος ισότητας πινάκων διακύμανσης

Το τεστ αυτό ελέγχει την υπόθεση H_0: $\mathbf{\Sigma}_1=\mathbf{\Sigma}_2=\ldots=\mathbf{\Sigma}_K$ άρα για p.value(Sig.)>0.05 όπως στο παράδειγμα μας δεν απορρίπτεται η υπόθεση της ισότητας των πινάκων διακυμάνσεων. Το τεστ αυτό είναι ευαίσθητο σε αποκλίσεις από την κανονική κατανομή.

Η δεύτερη κατηγορία ορίζει τον υπολογισμό και την εμφάνιση των συντελεστών της διαχωριστικής συνάρτησης (**Function Coefficients**). Για κάθε ομάδα υπολογίζουμε ένα σκορ με βάση μια συνάρτηση. Στην περίπτωση μας οι συναρτήσεις των είναι γραμμικές ως προς τις ανεξάρτητες μεταβλητές. Η επιλογή **Fisher's** υπολογίζει τους συντελεστές των γραμμικών συναρτήσεων των σκορ με τη μέθοδο του Fisher και για το παράδειγμα των θεριστικών μηχανών μας δίνει

Classification Function Coefficients

	GROUP	
	1 NAI	2 OXI
INCOME	.430	.329
LOTSIZE	5.467	4.682
(Constant)	-73.160	-51.421

Fisher's linear discriminant functions

Πίνακας 10.6. Οι διαχωριστικές συναρτήσεις του Fisher

που σημαίνει ότι το σκορ για την ομάδα που κατέχει θεριστικές μηχανές είναι:

$$w_1 = -73.16 + 0.43\ \text{INCOME} + 5.47\ \text{LOTSIZE}$$

ενώ για την ομάδα που δεν κατέχει θεριστικές μηχανές είναι:

$$w_2 = -51.42 + 0.33\ \text{INCOME} + 4.68\ \text{LOTSIZE}\ .$$

Κατατάσσουμε κάθε καινούριο αγρότη στην ομάδα όπου παρατηρείται το μέγιστο σκορ. Στις περίπτωση των δύο ομάδων έχουμε:

αν $w_1 > w_2$ κατατάσσουμε την παρατήρηση στην $1^η$ ομάδα αλλιώς στην $2^η$. Άρα αν $Z= w_1 - w_2$ τότε για $Z>0$ κατατάσσουμε στην $1^η$ ομάδα αλλιώς στην $2^η$. Όπου το Z μπορεί να δοθεί ως:

$Z = w_1 - w_2 =$ (-73.16+51.42) + (0.43 -0.33) INCOME + (5.47-4.68)LOTSIZE

$=$ -21.74 + 0.10 INCOME + 0.79 LOTSIZE

Οι συντελεστές αυτοί είναι ανάλογοι των μη τυποποιημένων συντελεστών (unstandardized function coefficients), οι οποίοι δίνονται από το υπο-μενού **statistics** επιλέγοντας την εμφάνιση των μη τυποποιημένων συντελεστών της κανονικοποιημένης διαχωριστικής συνάρτησης (unstandardized function coefficients). Το αποτέλεσμα για το παράδειγμα μας είναι

Canonical Discriminant Function Coefficients

	Function
	1
INCOME	.048
LOTSIZE	.380
(Constant)	-10.508

Unstandardized coefficients

Έτσι η διαχωριστική συνάρτηση μπορεί να γραφτεί ως Z=-10.51 + 0.05 INCOME + 0.38 LOTSIZE. Αυτή η διαχωριστική συνάρτηση έχει συντελεστές ανάλογους τους συντελεστές της διαχωριστικής συνάρτησης Z=-21.74 + 0.10 INCOME + 0.79 LOTSIZE που υπολογίσαμε παραπάνω (συγκεκριμένα είναι unstandardized × 2.08 για το παράδειγμα μας).

Οι αντίστοιχοι τυποποιημένοι συντελεστές της κανονικοποιημένης διαχωριστικής συνάρτησης (standardized canonical discrimination function coefficients) είναι χρήσιμοι όταν έχουμε ανεξάρτητες μεταβλητές διαφορετικής κλίμακας όπως και στο παράδειγμα μας. Οι συντελεστές αυτοί δίνουν μια ένδειξη της συνεισφοράς της κάθε μεταβλητής στη διαχωριστική συνάρτηση.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
INCOME	.806
LOTSIZE	.785

Οι συντελεστές αυτοί μπορούν να υπολογιστούν και από τους μη-τυποποιημένους συντελεστές πολλαπλασιάζοντας τους με την συνδυασμένη εκτίμηση των τυπικών τους αποκλίσεων οι οποίες δίνονται από τον πίνακα **within groups covariance matrix**.

Η τρίτη κατηγορία των επιλογών του υπο-μενού **statistics** δίνει την επιλογή υπολογισμού των συνδυασμένων πινάκων συνδιακύμανσης και συσχέτισης οι οποίοι για το παράδειγμα μας είναι:

Pooled Within-Groups Matrices[a]

		INCOME	LOTSIZE
Covariance	INCOME	276.675	-7.204
	LOTSIZE	-7.204	4.273
Correlation	INCOME	1.000	-.209
	LOTSIZE	-.209	1.000

a. The covariance matrix has 22 degrees of freedom.

Πίνακας 10.7. Ο σταθμισμένος πίνακας διακύμανσης και ο αντίστοιχος πίνακας συσχετίσεων

και την επιλογή υπολογισμού των πινάκων συνδιακύμανσης για κάθε ομάδα και συνολικά οι οποίοι για το παράδειγμα μας είναι:

Covariance Matrices[a]

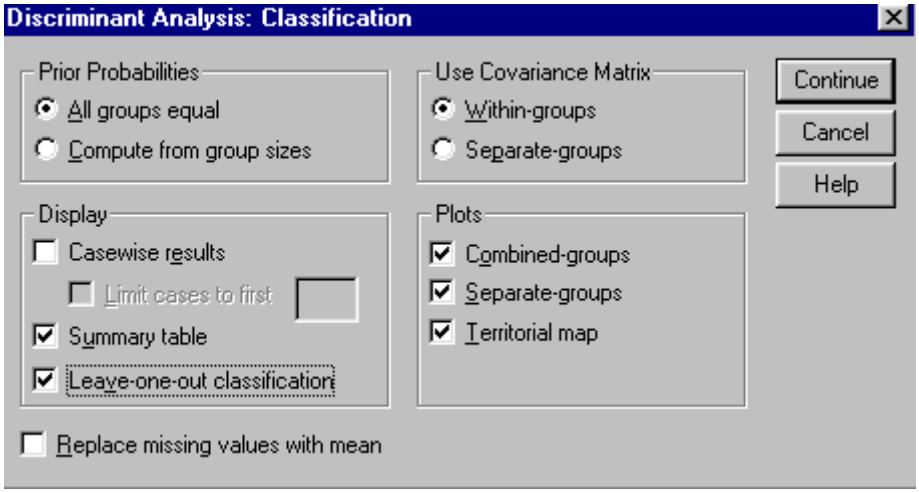
GROUP		INCOME	LOTSIZE
1 NAI	INCOME	352.644	-11.818
	LOTSIZE	-11.818	4.082
2 OXI	INCOME	200.705	-2.589
	LOTSIZE	-2.589	4.464
Total	INCOME	391.769	8.274
	LOTSIZE	8.274	5.897

a. The total covariance matrix has 23 degrees of freedom.

Πίνακας 10.8. Πίνακες διακύμανσης για τις 2 ομάδες

Επιλογές Ταξινόμησης (Υπο-μενού Classify)

Θα προχωρήσουμε στην περιγραφή των επιλογών **Classify** αφήνοντας για αργότερα την περιγραφή των επιλογών Method που αναφέρονται στην επιλογή των μεταβλητών. Το υπο-μενού αυτό χωρίζεται σε τέσσερις ομάδες επιλογών: **Prior Probabilities**, **Use Covariance Matrix, Display** και **Plots**.



Εικόνα 10.3 Το menu classify

Στην πρώτη ομάδα επιλογών (**Prior Probabilities**) μπορούμε να διαλέξουμε ανάμεσα σε ίσες πιθανότητες ή σε υπολογισμό από το μέγεθος των δειγμάτων. Δυστυχώς αν έχουμε άλλη πληροφόρηση δεν μπορούμε να τη χρησιμοποιήσουμε στο SPSS. Για το λόγο αυτό, εάν το SPSS είναι διαθέσιμο μόνο, καλό θα ήταν όταν έχουμε πληροφόρηση για τα πραγματικά ποσοστά των ομάδων στον πραγματικό πληθυσμό τότε και στο δείγμα να διατηρείται η ίδια αναλογία έτσι ώστε να επιλέγουμε **Compute from group sizes.**

Στην δεύτερη ομάδα επιλογών (**Use Covariance Matrix**) μπορούμε να επιλέξουμε εάν η ανάλυση θα γίνει με ίσες ή άνισες διακυμάνσεις. Αυτό προϋποθέτει την εφαρμογή του Box's M τεστ το οποίο στο παράδειγμα μας δεν απορρίπτει την υπόθεση ίσων πινάκων συνδιακύμανσης οπότε επιλέγουμε χρήση του συνδυασμένου πίνακα συνδιακυμάνσεων (pooled **Within-groups** estimate of covariance matrix).

Η τρίτη ομάδα επιλογών περιλαμβάνει εμφάνιση των αποτελεσμάτων ανά παρατήρηση (**casewise results**). Αν έχουμε πολλές παρατηρήσεις μπορούμε να περιορίσουμε την εμφάνιση στις πρώτες n (**limit cases to first …**). Επιπλέον μπορούμε να δούμε πως κατατάσσεται η κάθε παρατήρηση αν κάνουμε τη διαχωριστική ανάλυση χωρίς την συγκεκριμένη παρατήρηση και μετά την κατατάξουμε με βάση την διαχωριστική συνάρτηση όλων των υπόλοιπων παρατηρήσεων (**leave-one-out classification**). Στο παράδειγμά μας για τις 5 πρώτες παρατηρήσεις έχουμε τα ακόλουθα αποτελέσματα.

Casewise Statistics

			Highest Group					Second Highest Group			Discriminant Sc
	Case Number	Actual Group	Predicted Group	P(D>d | G=g) p	df	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Function 1
Original	1	1	2**	.677	1	.782	.174	1	.218	2.729	-.618
	2	1	1	.306	1	.506	1.048	2	.494	1.092	.011
	3	1	1	.838	1	.848	.042	2	.152	3.474	.830
	4	1	1	.504	1	.681	.447	2	.319	1.961	.366
	5	1	1	.103	1	.996	2.656	2	.004	13.679	2.664
Cross-validated[a]	1	1	2**	.900	2	.846	.210	1	.154	3.617	
	2	1	2**	.252	2	.618	2.753	1	.382	3.713	
	3	1	1	.550	2	.821	1.195	2	.179	4.247	
	4	1	1	.494	2	.640	1.409	2	.360	2.564	
	5	1	1	.110	2	.997	4.406	2	.003	16.104	

For the original data, squared Mahalanobis distance is based on canonical functions.
For the cross-validated data, squared Mahalanobis distance is based on observations.

**. Misclassified case

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

Ο παραπάνω πίνακα πίνακας χωρίζεται στα αποτελέσματα με όλα τα δεδομένα (**original**) και στα αποτελέσματα διασταυρωμένης επικύρωσης (**cross-validated = leave-one-out-classification**) τα οποία κατατάσσουν την κάθε παρατήρηση με τη διαχωριστική συνάρτηση που κατασκευάζουμε από όλες τις υπόλοιπες παρατηρήσεις. Στον πίνακα εμφανίζεται η πραγματική ομάδα (**actual group**), η προβλεπόμενη (**predicted** - με αστερίσκο υποδεικνύεται η λανθασμένη πρόβλεψη) , τη πιθανότητα να απόσταση μεγαλύτερη αυτής που έχουμε παρατηρήσει με δεδομένο ότι κατατάσσουμε στην παρατήρηση στο predicted group (**P(D>d|G=g), p**), την πιθανότητα να ανήκει στη g ομάδα με δεδομένη τη συγκεκριμένη απόσταση d (**P(G=g|D=d)**), την τετραγωνική απόσταση d του Mahalanobis από το κεντροειδές (**squared Mahalanobis distance from centroid**) και τέλος το σκορ της διαχωριστικής συνάρτησης (**discriminant score**) .

Τέλος η επιλογή του περιληπτικού πίνακα (**summary table**) εμφανίζει τους παρακάτω πίνακες

Prior Probabilities for Groups

GROUP	Prior	Specified Prior	Effective Prior	Cases Used in Analysis Unweighted	Weighted
1 NAI	.500			12	12.000
2 OXI	.500			12	12.000
Total	1.000			24	24.000

(α)

Classification Results[b,c]

			Predicted Group Membership 1 NAI	2 OXI	Total
		GROUP			
Original	Count	1 NAI	11	1	12
		2 OXI	2	10	12
	%	1 NAI	91.7	8.3	100.0
		2 OXI	16.7	83.3	100.0
Cross-validated[a]	Count	1 NAI	10	2	12
		2 OXI	3	9	12
	%	1 NAI	83.3	16.7	100.0
		2 OXI	25.0	75.0	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 87.5% of original grouped cases correctly classified.

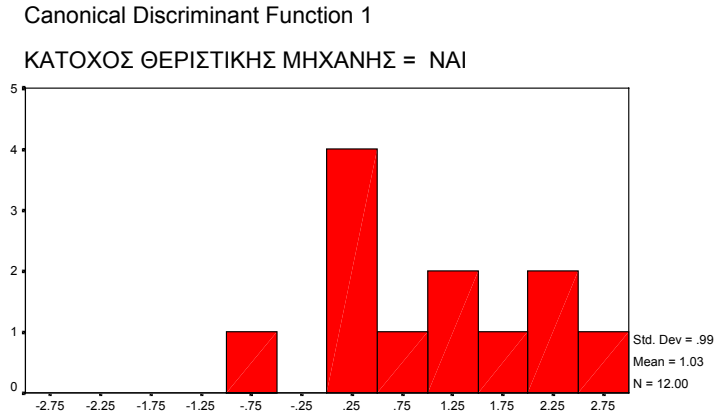
c. 79.2% of cross-validated grouped cases correctly classified.

(β)

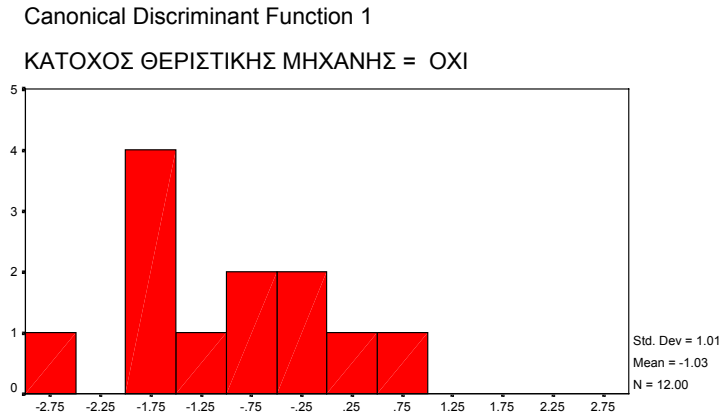
Πίνακας 10.9. Τα αποτελέσματα της κατάταξης των παρατηρήσεων

Ο δεύτερος πίνακας είναι χρήσιμος για τον υπολογισμό της επιτυχίας της διαχωριστικής ανάλυσης. Πιο συγκεκριμένα το ποσοστό σωστού διαχωρισμού είναι 87.5% για την συνολική διαχωριστική ανάλυση και 79.2% για την προσέγγιση της διασταυρωμένης επικύρωσης.

Τέλος η τέταρτη ομάδα επιλογών περιέχει τα γραφήματα όπου έχουμε επιλογή ιστογραμμάτων της κανονικοποιημένης συνάρτησης διαχωρισμού του συνολικού δείγματος , της κάθε ομάδας ξεχωριστά και του χάρτη περιοχών (όταν έχουμε πάνω από δύο ομάδες διαχωρισμού). Στο παράδειγμα μας έχουμε



Γράφημα 10.5α. Ιστόγραμμα των κανονικοποιημένων διαχωριστικών συναρτήσεων

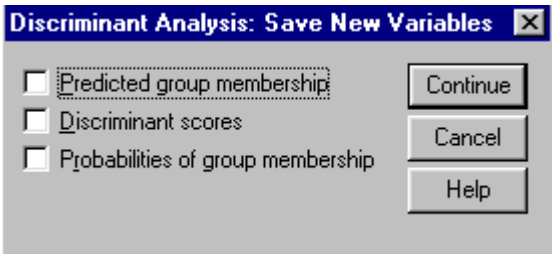


Γράφημα 10.5β. Ιστόγραμμα των κανονικοποιημένων διαχωριστικών συναρτήσεων

Επιλογές Αποθήκευσης (Υπο-μενού Save)

Η ομάδα επιλογών αποθήκευσης περιλαμβάνουν τις προβλεπόμενες με βάση τη διαχωριστική ανάλυση ομάδες (**predicted group membership**), τα σκορ διαχωρισμού (**discriminant scores**) και τις πιθανότητες να ανήκουν σε κάθε ομάδα (**probabilities of**

group memberships). Συνήθως μας ενδιαφέρει η αποθήκευση της προβλεπόμενης ομάδας για κατάταξη των παρατηρήσεων των οποίων η ομάδα δεν είναι γνωστή και για έλεγχο της επιτυχίας του διαχωρισμού μέσα από τον υπολογισμό του δείκτη συμφωνίας κάπα.

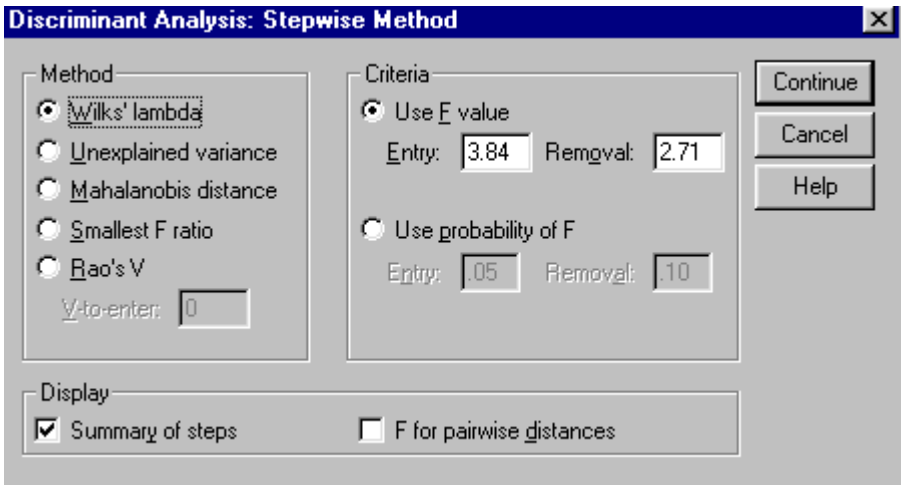


Εικόνα 10.4 Το menu Save

Επιλογές Μεθόδου Επιλογής Μεταβλητών (Υπο-μενού Method)

Οι επιλογές της κλιμακωτής επιλογής μεταβλητών ισχύουν μόνο εάν στο αρχικό μενού της διαχωριστικής συνάρτησης επιλέξουμε χρήση κλιμακωτής μεθόδου (**Use stepwise methods**). Η επιλογή αυτή είναι χρήσιμη για τον εντοπισμό κακών μεταβλητών και την αφαίρεση τους από την διαχωριστική ανάλυση. Έχουμε τρεις ομάδες επιλογών: μέθοδος (**method**), κριτήρια (**criteria**) και εμφάνιση (**display**).

Στην κλιμακωτή επιλογή μεταβλητών ξεκινάμε χωρίς καμία μεταβλητή στο μοντέλο και συνεχίζουμε προσθέτοντας τη μεταβλητή με τον καλύτερο δείκτη ανάλογα των μέθοδο που διαλέγουμε στην υπο-ομάδα **method** (ή το αντίστοιχο καλύτερο F) δεδομένου ότι ικανοποιείται και το όριο εισόδου της μεταβλητής στο μοντέλο που θέτουμε στις επιλογές **criteria** . Παράλληλα σε κάθε βήμα ελέγχουμε κάποια από τις μεταβλητές που ήδη είναι στο μοντέλο πρέπει να αφαιρεθεί σύμφωνα με το κριτήριο αφαίρεσης που θέτουμε στις επιλογές **criteria.** Αν παραπάνω από μια μεταβλητή πρέπει να αφαιρεθεί αφαιρούμε αυτή με το μικρότερο F.



Εικόνα 10.5 Το menu Method

Η πρώτη ομάδα επιλογών αναφέρεται στη μέθοδο επιλογής των μεταβλητών (**method**). Οι μέθοδοι που μπορούν να χρησιμοποιηθούν είναι

- **Wilks' lambda**: Σε κάθε βήμα επιλέγουμε ποια μεταβλητή θα εισάγουμε στο μοντέλο με βάση τη μείωση στο λάμδα του Wilks. Για κάθε ανεξάρτητη μεταβλητή υπολογίζεται ένα τεστ F το οποίο βασίζεται στη διαφορά μεταξύ των λάμδα του Wilks για τα μοντέλα με και χωρίς την αντίστοιχη μεταβλητή. Ο δείκτης του Wilks όπως ήδη αναφέραμε μετράει το ποσοστό της μη ερμηνεύσιμης από το μοντέλο διακύμανσης.
- **Unexplained Variance**: εδώ επιλέγουμε σαν δείκτη απόδοσης της κάθε μεταβλητής το άθροισμα της ερμηνεύσιμης διακύμανσης ανάμεσα στα ζευγάρια των ομάδων.
- **Mahalanobis Distance**: η μέθοδος αυτή βασίζεται στον υπολογισμό της απόστασης του Mahalanobis μεταξύ των δύο πιο κοντινών ομάδων. Σε κάθε βήμα εισάγεται η μεταβλητή που μεγιστοποιεί αυτή την απόσταση.
- **Smallest F ratio**: η μέθοδος αυτή βασίζεται στον υπολογισμό του F για όλα τα ζευγάρια των τιμών και από αυτές τις τιμές επιλέγουμε το μικρότερο F. Σε κάθε βήμα εισάγεται η μεταβλητή που μεγιστοποιεί αυτή το μικρότερο F.
- **Rao's V**: η μέθοδος αυτή βασίζεται στον υπολογισμό του της απόστασης Mahalanobis μεταξύ της κάθε ομάδας και του συνολικού δείγματος. Σε κάθε βήμα εισάγεται η μεταβλητή που μεγιστοποιεί αυτή την απόσταση.

Η δεύτερη ομάδα επιλογών αναφέρεται στα κριτήρια (**criteria**) εισαγωγής και αφαίρεσης μεταβλητών από το μοντέλο. Μπορούμε να ορίσουμε είτε το επίπεδο του F (**use F value**) είτε το επίπεδο των p.value (**use probability of F**).

Τέλος η ομάδα επιλογών εμφάνισης λεπτομερειών (**display**) μας δίνει τη δυνατότητα να παρακολουθήσουμε περιληπτικά τη εισαγωγή και απαλοιφή των μεταβλητών από το μοντέλο (**summary of steps**) και να δούμε τους δείκτες F για όλες τις συγκρίσεις ανά ζεύγη (**F for pairwise distances**).

Άλλα αποτελέσματα

Αν αποθηκεύσουμε τα σκορ διαχωρισμού και εφαρμόσουμε ανάλυση διακύμανσης κατά ένα παράγοντα ως προς τις πραγματικές ομάδες τότε έχουμε:

ANOVA

DIS1_1

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	25.681	1	25.681	25.681	.000
Within Groups	22.000	22	1.000		
Total	47.681	23			

Πίνακας 10.10. Πίνακας ανάλυσης διακύμανσης

Αν τώρα υπολογίσουμε τον λόγο των αθροισμάτων τετραγώνων ανάμεσα σε διαφορετικά γκρουπ και στο άθροισμα τετραγώνων μέσα σε κάθε γκρουπ τότε έχουμε 25.681/22.000= 1.167. Ο λόγος αυτός είναι ίσος με την ιδιοτιμή (eigenvalue) που δίνεται από τη διαχωριστική ανάλυση στο SPSS στον παρακάτω πίνακα

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.167[a]	100.0	100.0	.734

a. First 1 canonical discriminant functions were used in the analysis.

Πίνακας 10.11. Πίνακας ιδιοτιμής

Όταν έχουμε παραπάνω από δύο ομάδες οι ιδιοτιμές είναι χρήσιμες ως δείκτες μέτρησης της διασποράς των κεντρο-ειδών στον αντίστοιχο πολυμεταβλητό χώρο. Ο δείκτης κανονικής συσχέτισης (canonical correlation) μας δείχνει πόσο συσχέτιση υπάρχει μεταξύ των ομάδων και των σκορ της διαχωριστικής συνάρτησης και στη περίπτωση των δύο ομάδων υπολογίζεται ως τη ρίζα (Between Groups Sum of squares)/(Total Sum of squares) του παραπάνω πίνακα ανάλυσης διακύμανσης.

Από τον πίνακα που ακολουθεί επίσης δίδεται λάμδα του Wilks το οποίο όπως είπαμε είναι το ποσοστό της διακύμανσης που δεν εξηγείται από την προηγούμενη ανάλυση διακύμανσης και υπολογίζεται ως (Within Groups Sum of squares)/(Total Sum of squares). Μπορούμε να χρησιμοποιήσουμε το λάμδα για να ελέγξουμε την υπόθεση ότι οι μέσοι όλων των μεταβλητών ανά ομάδα είναι ίσοι. Αυτό το τεστ μπορεί να μας δώσει περιορισμένη διαγνωστική πληροφορία όταν οι μεταβλητές μας δεν είναι καλές για το διαχωρισμό των ομάδων (δηλαδή όταν δεν απορρίψουμε την H_0). Εδώ απορρίπτουμε την ισότητα των μέσων άρα δε φαίνεται να υπάρχει πρόβλημα με την εφαρμογή της διαχωριστικής ανάλυσης.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.461	16.243	2	.000

Πίνακας 10.12. Πίνακας με το Wilk's Lambda στατιστικό

Ο πίνακας δομής (**structure matrix**) μας δίνει τους δείκτες συσχέτισης κάθε ανεξάρτητης μεταβλητής με τις διαχωριστικές συναρτήσεις και μπορούν να

χρησιμοποιηθούν για να αξιολογήσουμε πόσο σημαντική είναι κάθε μεταβλητή για τη κατασκευή της διαχωριστικής συνάρτησης. Εδώ και οι δύο μεταβλητές φαίνεται να είναι το ίδιο σημαντικές αφού οι συσχετίσεις τους είναι πολύ κοντά και ίσες με 0.64 και 0.61.

Structure Matrix

	Function
	1
INCOME	.641
LOTSIZE	.616

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

Πίνακας 10.13. Πίνακας δομής

Τέλος ο πίνακας κεντροειδών μας δίνει τη μέση τιμή της κάθε κανονικοποιημένης διαχωριστικής συνάρτησης για κάθε ομάδα. Εδώ έχουμε δύο ομάδες άρα μια συνάρτηση και μέσους ίσους με 1.034 για τους κατόχους θεριστικών μηχανών και –1.034 για τους μη κατόχους.

Functions at Group Centroids

	Function
GROUP	1
1 NAI	1.034
2 OXI	-1.034

Unstandardized canonical discriminant functions evaluated at group means

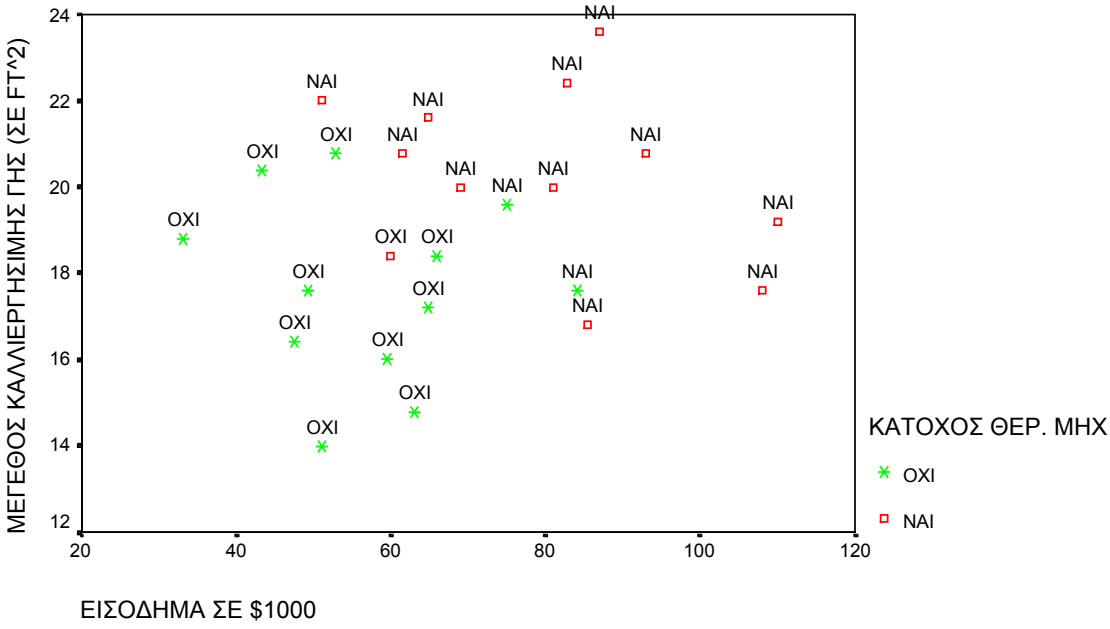
Πίνακας 10.14. Πίνακας κεντροειδών

Επικύρωση της διαχωριστική συνάρτησης.

Το ποσοστό του επιτυχημένου διαχωρισμού μπορούμε να το μετρήσουμε με το ποσοστό των σωστά καταχωρημένων παρατηρήσεων (87.5% στο παράδειγμα μας) αλλά και με τον υπολογισμό του δείκτη συμφωνίας κάπα αφού αποθηκεύσουμε τις προβλεπόμενες ομάδες. Κάνοντας τον αντίστοιχο πίνακας συνάφειας βρίσκουμε κάπα ίσο με 0.75 το οποίο μας δίνει πολύ καλή συμφωνία.

Διαγραμματική Απεικόνιση των Ομάδων

Τέλος αν θέλουμε να δούμε τις τιμές των μεταβλητών και τη σχέση πραγματικών και προβλεπόμενων ομάδων μπορούμε να κάνουμε το διάγραμμα σημείων με τύπο σημείων τις πραγματικές τιμές και ετικέτες τις προβλεπόμενες ομάδες.



Γράφημα 10.6. Διαγραμματική απεικόνιση των ομάδων

10.10Παράδειγμα Διαχωρισμού Τεσσάρων Ομάδων.

Εδώ θα ασχοληθούμε με το σετ δεδομένων world95 το οποίο έχει δημογραφικά χαρακτηριστικά 109 χωρών του κόσμου. Πριν ξεκινήσουμε κάνουμε ένα πρόχειρο υπολογισμό των περιγραφικών δεικτών. Βλέπουμε ότι σε τρεις τουλάχιστον μεταβλητές (CALORIES: ημερήσια κατανάλωση θερμίδων, LIT_MALE: ποσοστό ανδρών που ξέρουν να διαβάζουν, LIT_FEMA: ποσοστό γυναικών που ξέρουν να διαβάζουν) έχουμε πολλές missing τιμές και γι' αυτό τις αφαιρούμε από την επόμενη ανάλυση μας. Οι μεταβλητές οι οποίες θα χρησιμοποιηθούν για διαχωρισμό της μεταβλητής region δηλαδή της οικονομικής ή γεωγραφική ομάδας είναι ακόλουθες :

POPULATN: Πληθυσμός σε χιλιάδες

DENSITY Πυκνότητα Πληθυσμού (Πληθυσμός /τετραγωνικό χλμ.)
URBAN Ποσοστό Αστικού Πληθυσμού

LIFEEXPF Αναμενόμενος Χρόνος Ζωής Γυναικών LIFEEXPM Αναμενόμενος Χρόνος Ζωής Ανδρών LITERACY Ποσοστό κατοίκων που ξέρουν να διαβάζουν

POP_INCR Ποσοστό αύξησης του πληθυσμού ανά χρόνο

BABYMORT Βρεφική θνησιμότητα (θάνατοι ανά 1000 ζωντανές γεννήσεις)
GDP_CAP Ακαθάριστο εγχώριο προϊόν

AIDS Περιπτώσεις Aids

BIRTH_RT Ρυθμός γεννήσεων ανά 1000 κατοίκους

DEATH_RT Ρυθμός θανάτων ανά 1000 κατοίκους

AIDS_RT Αριθμός περιπτώσεων AIDS ανά 100000 κατοίκους

LOG_GDP Λογάριθμος με βάση το 10 του Ακαθάριστου εγχώριου προϊόντος
LG_AIDSR Λογάριθμος με βάση το 10 των περιπτώσεων AIDS ανά 100000 κατοίκους

B_TO_D Λόγος γεννήσεων προς θανάτους

FERTILTY Δείκτης γονιμότητας: μέσος αριθμός παιδιών ανά οικογένεια LOG_POP Λογάριθμος (με βάση το 10) του πληθυσμού

CROPGROW Δείκτης αγροτικής παραγωγής

Η μεταβλητή region έχει έξι επίπεδα – κατηγορίες: (1) Αναπτυγμένες Οικονομικά χώρες, (2) χώρες της ανατολική Ευρώπης, (3) χώρες του Ειρηνικού και της Ασίας, (4) χώρες της Αφρικής, (5) χώρες της Μέσης Ανατολής και (6) χώρες της Λατινικής Αμερικής. Σύνολο 103 χώρες έχουν πλήρη δεδομένα σε αυτές τις μεταβλητές.

Ξεκινώντας την διαχωριστική ανάλυση επιλέγουμε κλιμακωτή επιλογή μεταβλητών με τη χρήση του δείκτη λάμδα του Wilks.

Από τον πίνακα των συγκρίσεων των μέσων τιμών που ακολουθεί βλέπουμε ότι η μεταβλητή AIDS δεν είναι στατιστικά σημαντικά διαφορετική στις 6 ομάδες. Επίσης υπάρχουν αρκετές μεταβλητές (πληθυσμός, πυκνότητα πληθυσμού, AIDS, ρυθμός AIDS, λογάριθμος πληθυσμού, και δείκτης αγροτικής παραγωγής) με λάμδα μεγαλύτερο του 0.70 που είναι μια ένδειξη ότι δεν είναι καλές για τη διαχωριστική μας ανάλυση

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
POPULATN	.841	3.674	5	97	.004
DENSITY	.855	3.290	5	97	.009
URBAN	.531	17.143	5	97	.000
LIFEEXPF	.369	33.221	5	97	.000
LIFEEXPM	.381	31.545	5	97	.000
LITERACY	.416	27.256	5	97	.000
POP_INCR	.303	44.715	5	97	.000
BABYMORT	.425	26.219	5	97	.000
GDP_CAP	.278	50.419	5	97	.000
AIDS	.946	1.105	5	97	.363
BIRTH_RT	.304	44.454	5	97	.000
DEATH_RT	.557	15.454	5	97	.000
AIDS_RT	.712	7.843	5	97	.000
LOG_GDP	.387	30.742	5	97	.000
LG_AIDSR	.390	30.378	5	97	.000
B_TO_D	.428	25.920	5	97	.000
FERTILTY	.341	37.509	5	97	.000
LOG_POP	.832	3.921	5	97	.003
CROPGROW	.792	5.110	5	97	.000

Πίνακας 10.15. Πίνακας με αποτελέσματα ελέγχων μέσων τιμών

Προχωρώντας στα αποτελέσματα βλέπουμε από το τεστ του BOX ότι απορρίπτεται η υπόθεση των ίσων πινάκων συνδιακύμανσης.

Test Results

Box's M		1400.140
F	Approx.	4.586
	df1	225
	df2	9854.319
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

Πίνακας 10.16. Πίνακας με αποτελέσματα ελέγχου ίσων διακυμάνσεων

Για το λόγο αυτό ξανατρέχουμε τη διαχωριστική ανάλυση ορίζοντας τη χρήση διαφορετικών πινάκων συνδιακυμάνσεων για κάθε ομάδα (**separate groups**) στη επιλογή **Use Covariance Matrix** της ομάδας επιλογών **classify**.

Ακολουθεί ο πίνακας με τα τις λεπτομέρειες της κλιμακωτής επιλογής μεταβλητών η οποία ολοκληρώθηκε σε 9 βήματα:

Variables Entered/Removed[a,b,c,d]

			Wilks' Lambda											
							Exact F				Approximate F			
Step	Entered	Removed	Statistic	df1	df2	df3	Statistic	df1	df2	Sig.	Statistic	df1	df2	Sig
1	GDP_CAP		.278	1	5	97.000	50.419	5	97.000	.000				
2	LG_AIDSR		.104	2	5	97.000	40.247	10	192.000	.000				
3	POP_INCR		.041	3	5	97.000					38.418	15	262.655	
4	URBAN		.022	4	5	97.000					33.568	20	312.713	
5	DENSITY		.014	5	5	97.000					29.933	25	346.982	
6	DEATH_RT		.010	6	5	97.000					26.215	30	370.000	
7	LOG_GDP		.008	7	5	97.000					23.809	35	385.232	
8	AIDS_RT		.006	8	5	97.000					21.791	40	395.095	
9	BABYMORT		.005	9	5	97.000					20.185	45	401.221	

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 38.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

Πίνακας 10.17. Πίνακας με στοιχεία κλιμακωτής επιλογής μεταβλητών

Σύμφωνα με τον αλγόριθμο προστέθηκαν οι μεταβλητές: ακαθάριστο εγχώριο προϊόν, ο λογάριθμος των περιπτώσεων AIDS, το ποσοστό αύξησης του πληθυσμού, το ποσοστό αστικού πληθυσμού, η πυκνότητα του πληθυσμού, ο δείκτης θνησιμότητας, ο λογάριθμος του ΑΕΠ, ο ρυθμός εμφάνισης του AIDS και η παιδική θνησιμότητα. Επίσης δίδονται και αναλυτική πίνακες με τις τιμές του λάμδα ανά μεταβλητή σε κάθε βήμα.

Στην ανάλυση μας έχουμε έξι ομάδες διαχωρισμού άρα πέντε (5) διαχωριστικές συναρτήσεις. Για τη διαχωριστική συνάρτηση έχουμε μια ιδιοτιμή η οποία μπορεί να ερμηνευτεί σαν μέτρο της διασποράς των μέσων. Έτσι μπορούμε να πούμε ότι η $1^{η}$ κανονικοποιημένη διαχωριστική συνάρτηση εξηγεί το 5.742/(5.742+…+0.645) = 49.8% της συνολικής διακύμανσης.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	5.742[a]	49.8	49.8	.923
2	2.773[a]	24.1	73.9	.857
3	1.333[a]	11.6	85.4	.756
4	1.034[a]	9.0	94.4	.713
5	.645[a]	5.6	100.0	.626

a. First 5 canonical discriminant functions were used in the analysis.

Πίνακας 10.18. Πίνακας ιδιοτιμών

Ο παρακάτω πίνακας ελέγχει την υπόθεση την ισότητα των μέσων τιμών των διαχωριστικών συναρτήσεων (δηλαδή τα κεντροειδή) δηλαδή αν υπάρχει περίπτωση κακού διαχωρισμού. Σε όλες τις περιπτώσεις απορρίπτεται η υπόθεση.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 5	.005	500.046	45	.000
2 through 5	.034	319.708	32	.000
3 through 5	.128	194.227	21	.000
4 through 5	.299	114.158	12	.000
5	.608	47.045	5	.000

Πίνακας 10.19. Πίνακας με το Wilk's Lambda στατιστικό

Structure Matrix

	Function				
	1	2	3	4	5
GDP_CAP	-.614*	.142	.391	.405	.073
POP_INCR	.571*	.123	.311	.363	.446
BIRTH_RT[a]	.542*	.085	.305	-.264	.422
FERTILTY[a]	.481*	.117	.334	-.244	.457
LITERACY[a]	-.434*	-.043	-.387	.299	-.242
LG_AIDSR	-.039	.735*	.022	-.217	-.128
AIDS_RT	.062	.312*	.100	-.290	.137
AIDS[a]	.009	.164	.294*	.075	-.046
DENSITY	.017	-.178	.203*	-.071	-.175
DEATH_RT	.006	.202	.074	-.742*	.408
URBAN	-.250	-.038	-.249	.650*	-.055
B_TO_D[a]	.392	.031	.127	.589*	.181
LIFEEXPF[a]	-.403	-.166	-.219	.576*	-.317
LIFEEXPM[a]	-.359	-.207	-.180	.574*	-.309
LOG_GDP	-.473	.018	-.071	.519*	.139
BABYMORT	.400	.103	.285	-.494*	.266
CROPGROW[a]	-.057	.031	-.021	-.392*	-.130
POPULATN[a]	-.007	-.041	.067	-.034	-.281*
LOG_POP[a]	.025	-.013	.004	-.019	-.258*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

Πίνακας 10.20. Πίνακας δομής

Ο πίνακας δομής περιλαμβάνει τις συσχετίσεις μεταξύ των μεταβλητών και των κανονικοποιημένων διαχωριστικών συναρτήσεων. Στον πίνακα συμπεριλαμβάνονται και οι μεταβλητές που τελικά δεν χρησιμοποιήθηκαν. Επίσης με αστερίσκο υποδηλώνεται σε ποια διαχωριστική συνάρτηση έχουμε τη μεγαλύτερη συσχέτιση. Έτσι βλέπουμε ότι το ΑΕΠ, αύξηση του πληθυσμού, ο ρυθμός γεννήσεων, ο δείκτης γονιμότητας και το ποσοστό κατοίκων που γνωρίζουν να διαβάζουν επηρεάζουν πιο πολύ την $1^{η}$ διαχωριστική συνάρτηση, ο λογάριθμος του AIDS και ο ρυθμός του AIDS τη $2^{η}$ διαχωριστική συνάρτηση, το AIDS και ο πληθυσμός την $3^{η}$ διαχωριστική συνάρτηση, ο δείκτης θνησιμότητας, το ποσοστό αστικού πληθυσμού, λόγος γεννήσεων προς θανάτους, ο αναμενόμενος χρόνος ζωής ανδρών και γυναικών, λογάριθμος του ΑΕΠ, η βρεφική θνησιμότητα και η αγροτική παραγωγή επηρεάζουν πιο πολύ την $4^{η}$ διαχωριστική συνάρτηση. Τέλος ο πληθυσμό και ο λογάριθμος του πληθυσμού επηρεάζουν πιο πολύ την $5^{η}$ διαχωριστική συνάρτηση.

Ο πίνακας των συντελεστών των κανονικοποιημένων συναρτήσεων δίδεται όπως και στο προηγούμενο παράδειγμα

Canonical Discriminant Function Coefficients

	Function				
	1	2	3	4	5
DENSITY	.001	-.001	.000	-.001	.000
URBAN	.012	.010	-.032	.033	-.013
POP_INCR	.669	.395	.163	.493	1.192
BABYMORT	.029	.006	.028	.010	-.020
GDP_CAP	.000	.000	.000	.000	.000
DEATH_RT	-.203	.046	-.219	-.211	.385
AIDS_RT	.006	-.019	.007	.002	.004
LOG_GDP	.605	-.035	-1.664	.106	3.193
LG_AIDSR	.448	3.233	-.536	-.075	-1.020
(Constant)	-2.562	-5.905	6.342	-1.684	-12.997

Unstandardized coefficients

Πίνακας 10.21. Πίνακας συντελεστών κανονικοποιημένων συναρτήσεων

Ενώ ο πίνακας των συντελεστών των διαχωριστικών συναρτήσεων του Fisher δίδονται από τον πίνακα

Classification Function Coefficients

	REGION					
	1 Αναπτυγμένες Οικονομικά	2 Ανατολική Ευρώπη	3 Ειρηνικός/Ασία	4 Αφρική	5 Μέση Ανατολή	6 Λατινική Αμερική
DENSITY	-.004	-.002	.003	.001	.001	.000
URBAN	-.077	-.047	-.106	-.050	.022	.058
POP_INCR	12.337	11.942	12.833	16.747	17.101	14.270
BABYMORT	.109	.027	.240	.256	.233	.236
GDP_CAP	-.005	-.007	-.006	-.007	-.007	-.007
DEATH_RT	4.664	5.416	3.376	4.093	3.699	3.515
AIDS_RT	-.186	-.137	-.085	-.160	-.108	-.161
LOG_GDP	90.848	99.157	90.585	97.093	98.946	93.682
LG_AIDSR	19.186	10.963	9.098	23.377	13.977	22.499
(Constant)	-191.063	-202.349	-162.824	-215.097	-208.586	-194.443

Fisher's linear discriminant functions

Πίνακας 10.22. Πίνακας συντελεστών διαχωριστικών συναρτήσεων του Fisher

Ο πίνακας των κεντροειδών δίνει τις μέσες τιμές κάθε κανονικοποιημένης διαχωριστικής συνάρτησης

Functions at Group Centroids

	Function				
REGION	1	2	3	4	5
1 Αναπτυγμένες Οικονομικά	-3.91	1.19	.72	.34	-.06
2 Ανατολική Ευρώπη	-2.22	-1.88	-2.28	-1.09	.67
3 Ειρηνικός/Ασία	.64	-2.65	1.43	-.60	-.64
4 Αφρική	2.22	1.79	.46	-1.17	.66
5 Μέση Ανατολή	1.59	-.91	.07	1.80	1.03
6 Λατινική Αμερική	1.44	.84	-1.08	.46	-1.13

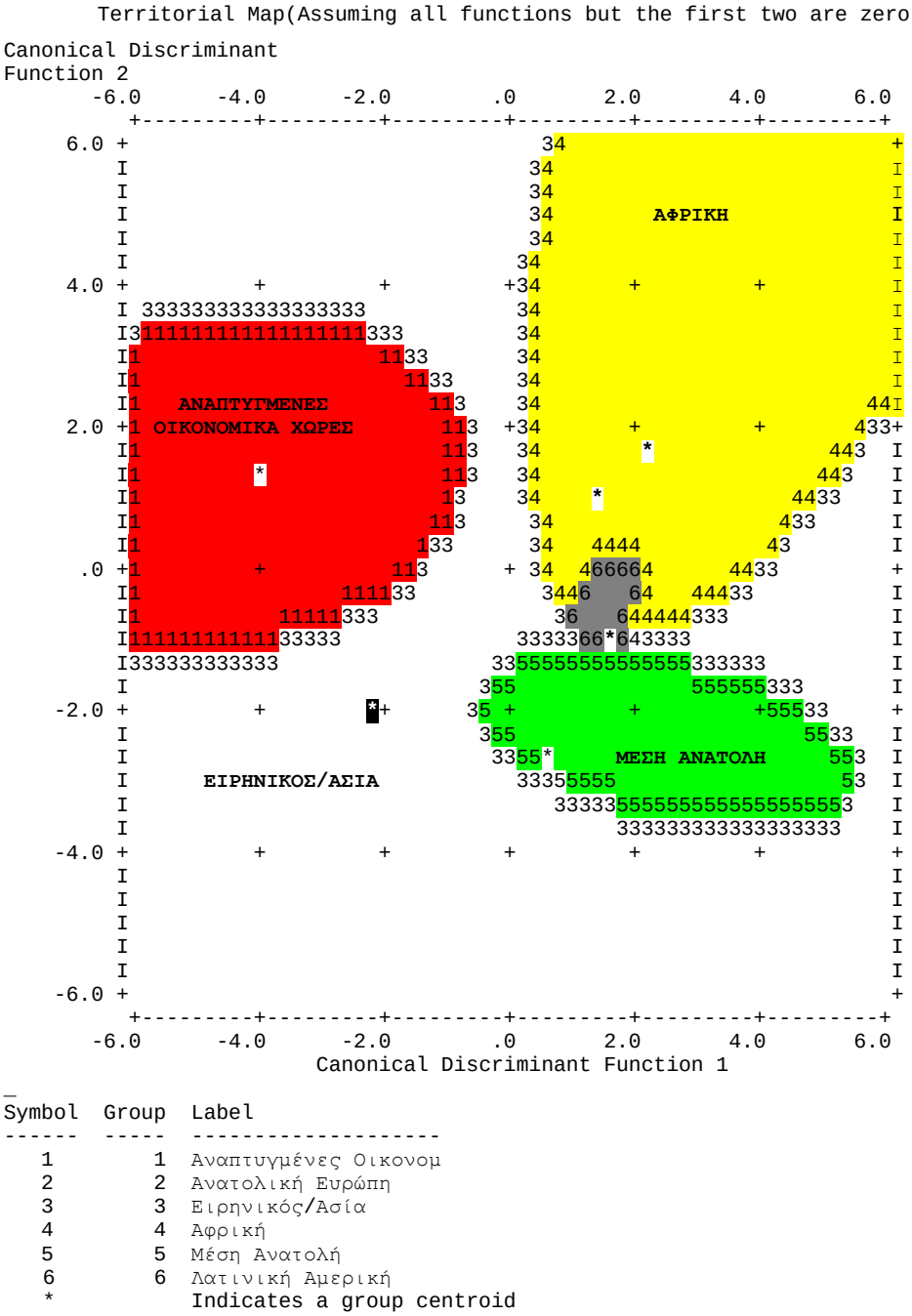
Unstandardized canonical discriminant functions evaluated at group means

Πίνακας 10.23. Πίνακας κεντροειδών

Βλέπουμε ότι οι αναπτυγμένες οικονομικά χώρες έχουν τη μικρότερη τιμή στην $1^η$ διαχωριστική συνάρτηση και ακολουθούν οι χώρες της ανατολικής Ευρώπης. Εάν δούμε τις συσχέτισης από τον πίνακα δομής βλέπουμε ότι η $1^η$ διαχωριστική συνάρτηση συσχετίζεται αρνητικά με το ΑΕΠ και θετικά με την αύξηση του πληθυσμού. Πράγματι στις αναπτυγμένες χώρες υπάρχει υψηλό ΑΕΠ και μικρός ρυθμός αύξησης του πληθυσμού. Η $2^η$ διαχωριστική συνάρτηση σχετίζεται θετικά με το AIDS και το λογάριθμό του. Παρατηρούμε πολύ αυξημένη τιμή στις χώρες της Αφρικής και αμέσως μετά στις

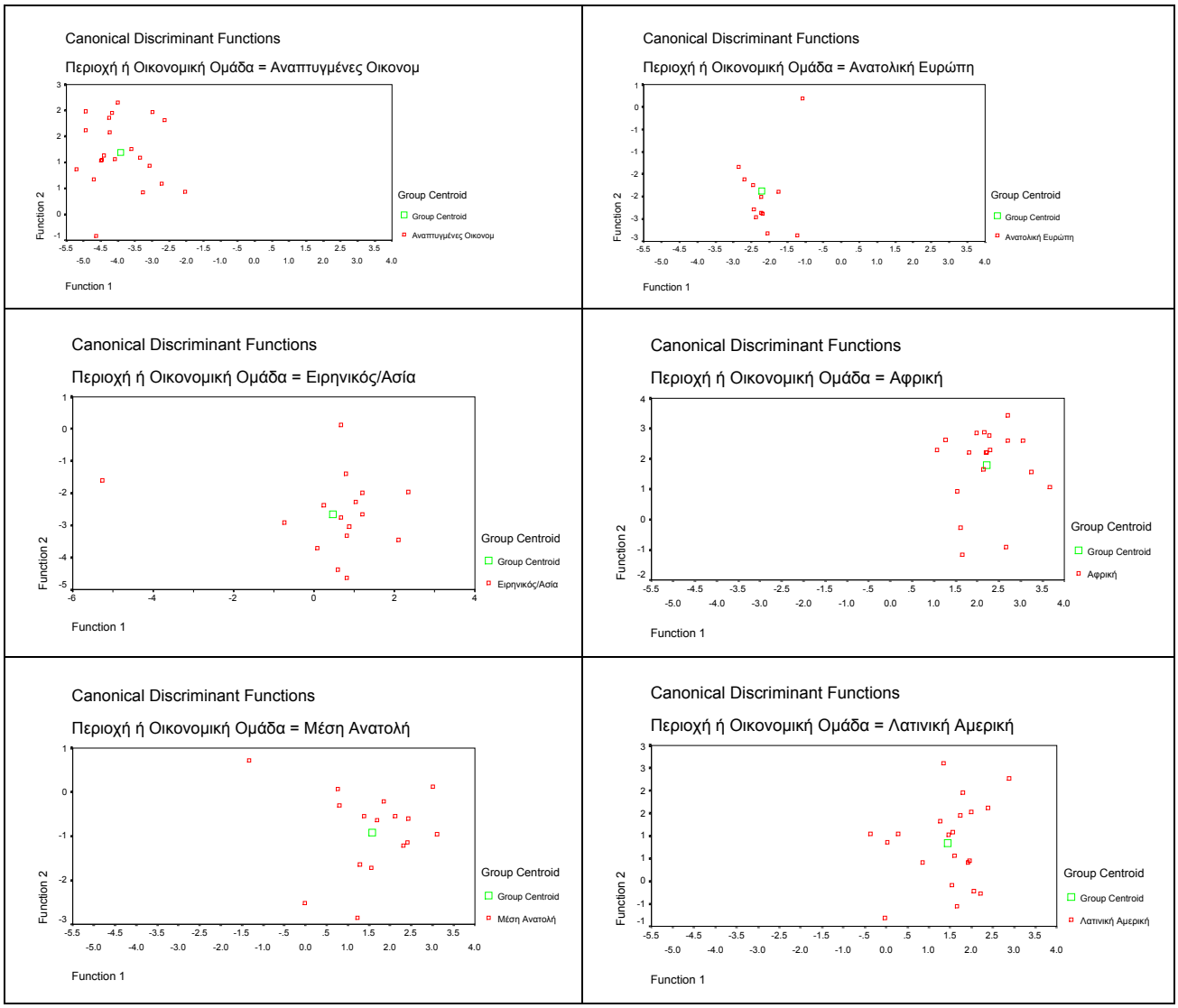
αναπτυγμένες χώρες. Στην ανατολική Ευρώπη, Ειρηνικο/Ασία και Μέση Ανατολή οι αντίστοιχες τιμές είναι πολύ μικρότερες. Όμοια μπορούμε να ερμηνεύσουμε και τις υπόλοιπες διαχωριστικές συναρτήσεις.

Πολύ χρήσιμος είναι ο χάρτης διαχωρισμού με τη χρήση των δύο πρώτων διαχωριστικών συναρτήσεων. Στο παρακάτω διάγραμμα μπορούμε να δούμε ότι η Αφρική έχει υψηλό σκορ στις 2 διαχωριστικές συναρτήσεις, τα αναπτυγμένα κράτη χαμηλό σκορ στην 1η διαχωριστική συνάρτηση και μέτρια επίπεδα σκορ στην 2η. Παρατηρήστε ότι η 2η ομάδα (χώρες ανατολικής Ευρώπης) δεν διαχωρίζονται ικανοποιητικά μόνο με τις δυο διαχωριστικές συναρτήσεις καθώς επίσης και τον μικρό χώρο που κατέχει η 6η ομάδα (χώρες της Λατινικής Αμερικής).



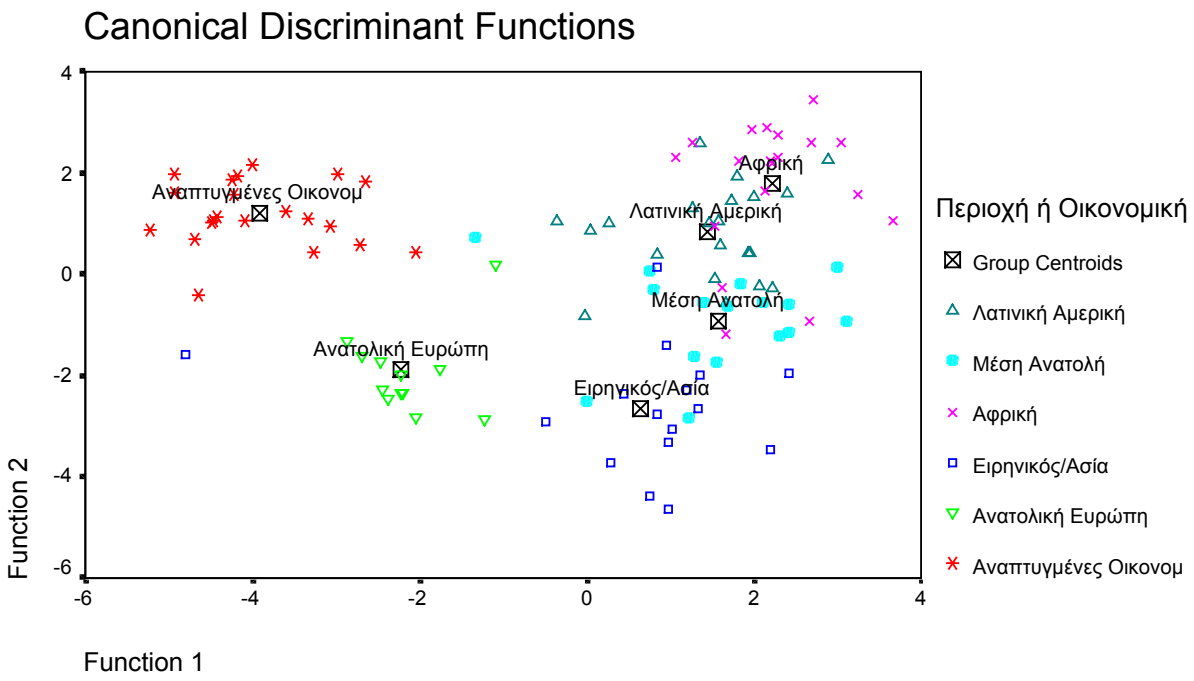
Γράφημα 10.7. Χάρτης διαχωρισμού

Τα διαγράμματα που ακολουθούν δίνουν μια εικόνα της κατανομής των σημείων των 2 πρώτων διαχωριστικών συναρτήσεων για κάθε ομάδα.



Γράφημα 10.8. Επιμέρους διαγράμματα κανονικοποιημένων διαχωριστικών συναρτήσεων

Τέλος το διάγραμμα που ακολουθεί δίνει μια συνολική εικόνα της κατανομής των σημείων των 2 πρώτων διαχωριστικών συναρτήσεων για κάθε ομάδα.



Γράφημα 10.9. Συνολικό διάγραμμα κανονικοποιημένων διαχωριστικών συναρτήσεων

Τέλος ο πίνακας ταξινόμησης μας δίνει υψηλό ποσοστό επιτυχίας ίσο με 90.5%. Μπορούμε να υπολογίσουμε και το κάπα το οποίο είναι ίσο με 0.885 (άριστη συμφωνία).

Classification Results[a]

			Predicted Group Membership						
		REGION	1 Αναπτυγμένες Οικονομικά	2 Ανατολική Ευρώπη	3 Ειρηνικός/Ασία	4 Αφρική	5 Μέση Ανατολή	6 Λατινική Αμερική	Total
Original	Count	1 Αναπτυγμένες Οικονομικά	21	0	0	0	0	0	21
		2 Ανατολική Ευρώπη	0	11	1	0	0	0	12
		3 Ειρηνικός/Ασία	0	0	13	0	2	1	16
		4 Αφρική	0	0	0	18	0	1	19
		5 Μέση Ανατολή	0	0	3	0	12	1	16
		6 Λατινική Αμερική	0	0	0	1	0	20	21
	%	1 Αναπτυγμένες Οικονομικά	100.0	.0	.0	.0	.0	.0	100.0
		2 Ανατολική Ευρώπη	.0	91.7	8.3	.0	.0	.0	100.0
		3 Ειρηνικός/Ασία	.0	.0	81.3	.0	12.5	6.3	100.0
		4 Αφρική	.0	.0	.0	94.7	.0	5.3	100.0
		5 Μέση Ανατολή	.0	.0	18.8	.0	75.0	6.3	100.0
		6 Λατινική Αμερική	.0	.0	.0	4.8	.0	95.2	100.0

a. 90.5% of original grouped cases correctly classified.

Πίνακας 10.24. Πίνακας ταξινόμησης

Βιβλιογραφία

Παρακάτω υπάρχει μια λίστα με βιβλιογραφία σχετική με τα θέματα που αναπτύσσονται στις σημειώσεις αλλά και άλλα θέματα πολυμεταβλητής ανάλυσης που δεν συζητήθηκαν στις σημειώσεις αυτές. Σε καμιά περίπτωση η λίστα δεν είναι πλήρης, η βιβλιογραφία σε θέματα πολυμεταβλητής στατιστικής είναι απέραντη. Το επίπεδο γνώσης που απαιτούν οι παρακάτω αναφορές ποικίλλει. Μερικά από τα βιβλία είναι πολύ τεχνικά και ο απλός αναγνώστης θα μπορούσε να τα παραλείψει. Μερικά όμως είναι εισαγωγικά και επομένως ιδιαίτερα εύχρηστα..

Ξένη

Afifi A., and V. Clark,(1996) Computer-Aided Multivariate Analysis, CRC Press.

Anderson T. W. (1984), An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, New York, 2nd edition

Anderberg, M.R. (1973). Cluster Analysis for Applications, New York: Academic Press, Inc.

Andrews, D. F. (1972) Plots of high-dimensional data. ***Biometrics***, 28:125-136.

Bartolomew, D.J., Steele, F., Moustaki, I. And Galbraith, J.I. (2001). The analysis and interpretation of multivariate data for science scientists. Chapman and Hall/ CRC

Basilevski, A. (1994) Statistical Factor Analysis and Related Methods. Theory and Applications. John Wiley & Sons

Benzécri, J-P. (1992). Correspondence Analysis Handbook, translated by T.K. Gopolan, Marcel Dekker.

Chatfield, C and Collins, A.J. (1992) Introduction to Multivariate Analysis. Chapman and Hall.

Chernoff, H. (1973). Using faces to represent points in k-dimensional space graphically. Journal of American Statistical Association, 68, 361-368.

Cooper, J. C. B (1983). Factor Analysis: An Overview. The American Statistician,37, 141-146

Efron, B. and Tibshirani, R.J. (1993) An introduction to the bootstrap. Chapman and Hall.

Everrit, B. S. (1993) Cluster Analysis, 3rd edition, Arnold, London

Giri, N.G. (1996) Multivariate Statistical Analysis. Dekker, New York

Gnanadesikan, R. (1977) Methods for Statistical Data Analysis of Multivariate Observations. John Wiley and Sons, New York.

Goldstein, M. (1982). Preliminary inspection of multivariate data. The American Statistician, 36, 358-363.

Greenacre, M. and Blasius, J. (1994). Correspondence Analysis in the Social Sciences. Academic Press

Gower, J.C. (1971). 'A General Coefficient of Similarity and Some of Its Properties,' Biometrics, 27, 857-871.

Hawkins, D.M. (1973). 'On the Investigation of Alternative Regressions by Principal Components Analysis,' Applied Statistics, 22, 275-286.

Hastie, T., Tibshirani, R., and Friedman, J.H. (2001). The Elements of Statistical Learning : Data Mining, Inference, and Prediction. New York: Springer.

Healy, J. R. (1968) Multivariate normal plotting. Applied Statistics, 17, 157-161.

Hotteling, H. (1933). 'Analysis of a Complex of Statistical Variables into Principal Components,' Journal of Educational Psychology, 24, 417-441, 498-520

Jackson J., (1991) *A User's Guide to Principal Components*, John Wiley & Sons, Inc., New York, NY.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data Clustering: A review. *ACM Computing Surveys,* 31,264-323

Jeffers, J.N.R. (1967). Two Case Studies on the Application of Principal Components Analysis, Applied Statistics, 16, 225-236.

Jolliffe, Y. (1986). Principal Component Analysis. Springer-Verlag, New York,

Kaufman, L. & Rousseeuw, P. J. (1990). Finding Groups in Data. An Introduction to Cluster Analysis. New York: Wiley.

Kruskal, J.B. and M. Wish 1978. *Multidimensional Scaling*. Sage.

Krzanowski, W. J. (1988) Principles of Multivariate Analysis. Oxford University Press.

Krzanowski, W.J. (1989). 'Cross-Validation in Principal Components Analysis,' Biometrics, 43, 575-584.

Lebart, L., Morineau, A., & Warwick, K. (1984). Multivariate Descriptive Statistical Analysis. New York: Wiley.

Mardia K.V. and Jupp, P.E. (1999) Directional Statistics. Wiley, Chichester.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika, 57, pp. 519-530.

Mardia, K.V. (1975). Assessment of multinormality and the Robustness of Hotelling's T2 test. Applied Statistics, 24, 163-171

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). Multivariate Analysis. London: Academic Press.

Massart, D.L. and Kaufmanns, L. (1983). The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, New York: John Wiley & Sons, Inc.

McLachlan, G. & Basford, K. (1988). Mixture Models: Inference and Applications to Clustering. New York: Marcel Dekker.

McLachlan, G.J. (1992). Discriminant analysis and statistical pattern recognition. Wiley, New York,

Rao, C.R. (1964). 'The Use and Interpretation of Principal Component Analysis in Applied Research,' *Sankhya* A 26 , 329 -358.

Rencher, (1997) Multivariate Statistical Inference and Applications. Wiley Interscience

Reyment R., and K. Joreskog, (1996) Applied Factor Analysis in the Natural Science, Cambridge University Press,.

Ripley, B.D. (1996). Pattern Recognition and Neural Networks, Cambridge University Press,Cambridge

Schatzoff, M. (1966). Exact distributions of Wilks's likelihood ratio criterion. Biometrika, 53, 347-358

Schiffman, S. M., Reynolds, L., and Young, F. W. (1981). ***Introduction to multidimensional scaling.*** New York: Academic Press.

Sneath, P.H.A. and Sokal, R.R. (1973). Numerical Taxonomy, San Francisco: W.H. Freeman.

Sokal, R.R. and Michener, C.D. (1958). 'A Statistical Method for Evaluating Systematic Relationships,' Univ. Kansas Sci. Mull., 38, 1409-1438.

Webb, A. (2001) Statistical Pattern Recognition, Wiley, NY

Ελληνική

Μαγδαληνός, Μ. (1990) Πολυμεταβλητή Στατιστική Ανάλυση. Ανώτατη Σχολή Οικονομικών και Εμπορικών Επιστημών.

Μπεχράκης Θ. (1999) Πολυδιάστατη Ανάλυση Δεδομένων, Εκδόσεις Λιβάνη.

Μτζούφρας Ι. (2002) Στοιχεία Πολυμεταβλητής Ανάλυσης Δεδομένων. Πανεπιστημιακές Παραδόσεις, Πανεπιστήμιο Αιγαίου

Πανάρετος, Ι. και Ξεκαλάκη, Ε. (1993) Εισαγωγή στην Πολυμεταβλητή Στατιστική Ανάλυση.

Παπαδημητρίου, Ι. (1998) Εισαγωγή στην Ανάλυση Δεδομένων. Πανεπιστημιακές Παραδόσεις, Πανεπιστήμιο Μακεδονίας.