ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΠΟΛΥΜΕΤΑΒΛΗΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

Διπλωματική Εργασία

Αθανάσιος Σταυρίδης

Επιβλέπουσα Καθηγήτρια:
Χρυσηΐς Καρώνη
Ομότιμη Καθηγήτρια Ε.Μ.Π.

Χρυσηΐς Καρώνη	**Βασίλης Παπανικολάου**	**Καλλιόπη Παυλοπούλου**
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Ε.ΔΙ.Π Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2023

Στην οικογένειά μου,
καθώς και στην καθηγήτριά μου,
κυρία Χρυσηΐδα Καρώνη,
για την πολύτιμη συμβολή της

Περιεχόμενα

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με δύο κύριες τεχνικές, οι οποίες χρησιμοποιούνται ευρέως για τη στατιστική ανάλυση πολυμεταβλητών δεδομένων, εστιάζοντας στη διερεύνηση των μεταβλητών του εκάστοτε δείγματος δεδομένων.

Στο πρώτο κεφάλαιο, εισάγονται οι τεχνικές της Ανάλυσης Κύριων Συνιστωσών και της Ανάλυσης Συστάδων ως αναπόσπαστα μέρη της μη επιβλεπόμενης μάθησης. Επίσης, αναφέρονται τα πεδία των επιστημών, στα οποία αυτές έχουν εφαρμογή.

Στο δεύτερο κεφάλαιο, παρουσιάζεται θεωρητικά και αναλυτικά η Ανάλυση Κύριων Συνιστωσών, η οποία θέτει ως στόχο την αντικατάσταση (συνήθως συσχετισμένων) μεταβλητών του αρχικού δείγματος με τις κύριες συνιστώσες, οι οποίες είναι σε αριθμό λιγότερες από αυτόν των αρχικών μεταβλητών και οι οποίες έχουν την ιδιότητα να είναι ασυσχέτιστες μεταξύ τους.

Στο τρίτο κεφάλαιο, παρουσιάζεται η Ανάλυση Συστάδων, που προσφέρει την ομαδοποίηση των παρατηρήσεων σε συστάδες έτσι, ώστε τα στοιχεία, τα οποία ανήκουν στην ίδια συστάδα, να εμφανίζουν μεγάλη ομοιότητα. Στη συνέχεια, αναπτύσσονται οι τεχνικές της Ανάλυσης Συστάδων, οι οποίες διακρίνονται στη μη Ιεραρχική ανάλυση συστάδων, όπως η K-means, στην Ιεραρχική και τέλος στην Ανάλυση Συστάδων βασισμένη σε μοντέλο. Οι πρώτες δύο δεν προϋποθέτουν την ύπαρξη κατανομών, οι οποίες δύνανται να περιγράψουν το πληθυσμό του δείγματος, σε αντίθεση με την τρίτη, η οποία βασίζεται στην εύρεση στατιστικού μοντέλου, στο οποίο υπακούει το δείγμα. Επίπροσθετα, παραθέτονται τα πλεονεκτήματα και τα μειονεκτήματα της κάθε τεχνικής.

Στο τελευταίο κεφάλαιο, περιγράφεται πώς εφαρμόζονται οι παραπάνω μέθοδοι με τη βοήθεια του στατιστικού προγράμματος R-studio, αφενός στο δείγμα USArrests, το οποίο αφορά στις συλλήψεις, που έλαβαν χώρα σε 50 πολιτείες των Ηνωμένων Πολιτειών της Αμερικής, και αφετέρου στο δείγμα heptathlon, στο οποίο είναι καταγεγραμμένες οι επιδόσεις 25 αθλητριών στο έπταθλο στους Ολυμπιακούς Αγώνες στη Σεούλ της Νότιας Κορέας το 1988. Με αυτόν τον τρόπο, εξάγονται συμπεράσματα τόσο για αυτά τα δύο δείγματα όσο και για τις τεχνικές της Ανάλυσης Κύριων Συνιστωσών και της Ανάλυσης Συστάδων εν γένει.

Λέξεις-Κλειδιά: Μη επιβλεπόμενη μάθηση, πολυμεταβλητά δεδομένα, Ανάλυση Κύριων Συνιστωσών, αλγόριθμος συμπλήρωσης πινάκων, Ανάλυση Συστάδων, Συσταδοποίηση.

Abstract

The present thesis deals with two main techniques, which are widely used in the realm of multivariate analysis, focusing on the exploration of the variables of each data sample.

In the first chapter, the techniques of Principal Component Analysis and Cluster Analysis are introduced as substantial parts of Unsupervised Learning. Furthermore, the fields of sciences in which they are applied are mentioned.

The second chapter gives an extensive theoretical presentation of Principal Component Analysis. To give a deeper insight, Principal Component Analysis aims at the replacement of (usually correlated) variables of the initial sample with the principal components, which are fewer in comparison with the initial variables. What is more, these principal components have the property of not being correlated with each other.

Cluster Analysis is presented in the third chapter. This analysis yields the classification of each observation into clusters, in such a way that the individuals of each cluster appear to have high similarity. After that, the main techniques of Clustering are developed: Non-Hierarchical clustering, such as K-means, Hierarchical clustering and finally Model Based clustering. The first two do not require the existence of any distributions that could describe the population of the sample. On the contrary, the third one is feasible as long as the sample follows a formal statistical model. In addition, advantages and drawbacks of each method are cited.

In the last chapter, the aforesaid methods are implemented by using

the software package R-studio in both data sets USArrests and heptathlon. The former refers to arrests in the 50 states of the United States of America, and the latter to the performances of 25 women athletes in the heptathlon at the Olympic Games in Seoul in South Korea in 1998. Inferences are made on the above data sets and the techniques of Principal Component Analysis and Cluster Analysis in general.

Keywords: Unsupervised learning, multivariate data, Principal Component Analysis, algorithm for matrix completion, Cluster Analysis, Clustering.

Κεφάλαιο 1

Εισαγωγή στη μη επιβλεπόμενη μάθηση

1.1 Ιδιότητες μη επιβλεπόμενης μάθησης

Η μη επιβλεπόμενη μάθηση (Unsupervised learning) αποτελεί πεδίο μηχανικής μάθησης μεγάλου όγκου δεδομένων, ήτοι δείγματα δεδομένων, που εξαρτώνται από αρκετά μεγάλο αριθμό παραγόντων και κατά συνέπεια από πολλές τυχαίες μεταβλητές (πολυμεταβλητά δεδομένα). Συγκεκριμένα, ιδιαίτερο χαρακτηριστικό των προβλημάτων, τα οποία υπόκεινται σε αυτού του είδους μάθησης, συνιστά η μελέτη ενός δείγματος με n παρατηρήσεις και με p επεξηγηματικές μεταβλητές $X_1, X_2, X_3, ..., X_p$, και στα οποία σε αντίθεση με τα προβλήματα επιβλεπόμενης μάθησης (για παράδειγμα μοντέλα παλινδρόμησης) είναι δύσκολη η εκτίμηση της τιμής απόκρισης Y και, ως εκ τούτου, η πρόβλεψη λόγω της υποκειμενικότητας των εφαρμοσμένων μηχανικών υπολογισμού. Ωστόσο, η μάθηση αυτή επικεντρώνεται σε τεχνικές διερευνητικής ανάλυσης δεδομένων (exploratory data analysis), όπου δίνεται έμφαση στη διερεύνηση των μεταβλητών, που περιγράφουν τα δεδομένα (Hastie et al., 2021).

1.2 Στόχοι και αξία μη επιβλεπόμενης μάθησης

Η μη επιβλεπόμενη μάθηση θέτει ως σκοπό την περιγραφή και την ανάλυση των μεταβλητών X_i με $i = 1, ..., p$, που προσδιορίζουν τα δεδομένα.

Αναλυτικότερα, κατέχουν θεμελιώδη θέση σε αυτήν η οπτικοποίηση των δεδομένων και η ομαδοποίηση των παρατηρήσεων και των μεταβλητών X_i. Για την επίτευξη των παραπάνω διαδικασιών υπάρχουν ποικίλες μέθοδοι. Αυτή η διπλωματική εργασία εστιάζει σε δύο τεχνικές, σε αυτήν που βασίζεται στην Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis) και έχει ως αντικείμενο τη μείωση των χαρακτηριστικών μεταβλητών των δεδομένων χωρίς να διαταράσσεται η διακύμανση των παρατηρήσεων και η μεταβλητότητα του αρχικού δείγματος, και σε αυτήν που έχει ως αναφορά την Ανάλυση Συστάδων(Cluster Analysis) και συντελεί στην ομαδοποίηση καθώς και στην εν γένει γραφική απεικόνιση των δεδομένων.

Είναι φανερό ότι η μη επιβλεπόμενη μάθηση παίζει πρωταρχικό ρόλο στη σύγχρονη καθημερινότητα, δεδομένου ότι ως τομέας του επίκαιρου παρά ποτέ αντικειμένου της Ανάλυσης Μεγάλων Δεδομένων (Big Data Analysis) δύναται να «απαντήσει» σε ερωτήματα τόσο οικονομικών προβλημάτων επιχειρήσεων όσο ζητημάτων, που προκύπτουν σε βιολογικά συστήματα (για παράδειγμα μέτρηση συγκεντρώσεων mRNA, ταυτοποίηση γονιδίων, εύρεση νουκλεοσωμικών θέσεων και γονιδιακών ρυθμίσεων κ.λ.π) και σε ιατρικά πειράματα (λόγου χάριν συσχέτηση του DNA με αρκετές αρρώστειες και νοσήματα). Επιπλέον, αποτελεί σημαντικό εργαλείο σε διάφορες εφαρμογές της πληροφορικής και της τεχνητής νοημοσύνης, όπως σε κλάδους της υπολογιστικής ή τεχνητής όρασης, της αναγνώρισης ομιλίας, των συστημάτων συστάσεων και της ανάκτησης πληροφοριών, καθώς επίσης και της επιστήμης της Μηχανικής (Hochreiter, 2015).

Κεφάλαιο 2

Ανάλυση Κύριων Συνιστωσών

2.1 Εισαγωγή

Όπως έχει αναφερθεί στο προηγούμενο κεφάλαιο, η ερμηνεία πολυμεταβλητών δεδομένων αποτελεί διακύβευμα καθώς και επιτακτική ανάγκη για τη σύγχρονη Επιστήμη. Αυτή η ανάγκη, λοιπόν, έχει οδηγήσει σε ανακάλυψη τρόπων και μηχανισμών απλοποίησης των αρχικών στατιστικών δεδομένων ελαττώνοντας, δηλαδή, τη διάσταση του αρχικού πολυμεταβλητού δείγματος. Η τεχνική της Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis) συνιστά μία από τις παλαιότερες και τις πιο ευρέως διαδεδομένες και χρησιμοποιημένες, και υπόκειται στην κατηγορία της μη επιβλεπόμενης μάθησης, εφόσον δεν καταλήγει στην πρόβλεψη της τιμής απόκρισης Y. Όταν πρωτοδιατυπώθηκε σε πολλά έργα, όπως αυτό των Cauchy και Jordan, λόγω της μη εύκολης προσβασιμότητας σε Ηλεκτρονικούς Υπολογιστές παρέμεινε σε αρχικό στάδιο. Μετά από κάποιες δεκαετίες αναπτύχθηκε περαιτέρω με σχετικές εφαρμογές σε προβλήματα μεγαλυτέρων διαστάσεων στα δεδομένα. Η σημερινή μορφή της εισήχθη αρχικά από τον Pearson κατά τη χρονική περίοδο του 1901 και στη συνέχεια εμπλουτίσθηκε από τον Hotelling το 1933 (Jolliffe, 2002).

Η μέθοδος της Ανάλυσης Κύριων Συνιστωσών θέτει ως βάση της τη μείωση του (μεγάλου) όγκου μεταβλητών του αρχικού δείγματος δεδομένων έχοντας ως σημεία αναφοράς την αυξημένη μαθηματική και στατιστική ερμηνευσιμότητα (Interpretability) και την ελάχιστη απώλεια στατιστικής πληροφορίας μέσω της διατήρησης όσο το δυνατόν περισσότερης μεταβλητότητας και, άρα, διακύμανσης από τα αρχικά δεδομένα. Αυτό το γεγονός επιτυγχάνεται δια

μέσου της εύρεσης καινούριων μεταβλητών, που αποτελούν γραμμικούς συνδυασμούς των αρχικών μεταβλητών του δείγματος και που είναι ασυσχέτιστες μεταξύ τους (Everitt και Hothorn, 2011). Αυτή η διαδικασία αποφέρει τους παρακάτω καρπούς: πρωτίστως, είναι εφικτή η οπτικοποίηση των παρατηρήσεων και των μεταβλητών των δεδομένων και, επίσης, είναι δυνατή η στατιστική διαδικασία (Imputation), όπου, δηλαδή, συμπληρώνονται ελλιπή στοιχεία ενός πολυδιάστατου πίνακα δεδομένων με μη καταγεγραμμένα και μη συμπληρωμένα στοιχεία (Hastie et al., 2021). Παράλληλα, η μέθοδος των κυρίων συνιστωσών δε δέχεται υποθέσεις, ότι δηλαδή δεν υπακούει σε κάποια κατανομή, και εξαιτίας αυτού, αποτελεί τεχνική διερευνητικής ανάλυσης δεδομένων (exploratory data analysis) με την προϋπόθεση, όμως, ότι οι αρχικές μεταβλητές των δεδομένων είναι συσχετισμένες μεταξύ τους. Επιπλέον, κρίνεται σκόπιμο να τονισθεί ότι η μέθοδος αυτή «παράγει» νέες (ορθογώνιες) μεταβλητές, οι οποίες δεν μπορούν να ξεπερνούν σε αριθμό τις αρχικές μεταβλητές του προβλήματος, και οι οποίες χαρακτηρίζονται από μία διάταξη, στην οποία οι πρώτες μεταβλητές εκφράζουν το μεγαλύτερο ποσοστό της αρχικής διακύμανσης των δεδομένων (Jolliffe, 2002). Πρόκειται, λοιπόν, για μία διάταξη, η οποία επιτρέπει να κατανοήσουμε με «μία πρώτη ματιά» τη δομή των δεδομένων μέσω της γραφικής παράστασης των παρατηρήσεων προβαλλόμενες στις κύριες συνιστώσες, οι οποίες περιγράφουν το δείγμα.

2.2 Κατάταξη δεδομένων υπό κλίμακα - Τυποποίηση δεδομένων

Έστω X το διάνυσμα στήλη των μεταβλητών X_i, όπου $i = 1, ..., p$ του δείγματος . Κρίνεται αναγκαίο τα στοιχεία του διανύσματος X_i να τυποποιηθούν έτσι, ώστε κάθε στοιχείο να έχει είτε μέσο όρο μηδέν και τυπική απόκλιση ένα, σε αντίθεση με εφαρμογές της επιβλεπόμενης μάθησης, όπου το κεντράρισμα των επεξηγηματικών μεταβλητών δεν επηρεάζει το αποτέλεσμα στα μοντέλα. Αυτό το βήμα θεωρείται μείζον, όταν οι μεταβλητές είναι μετρημένες σε διαφορετικές μονάδες μέτρησης, προκειμένου οι κύριες συνιστώσες να μη μεταβάλλονται από μεταβλητές με υψηλές μονάδες μέτρησης και ως εκ τούτου με υψηλή τιμή διασποράς, και να μη «ρίχουν το βάρος τους» σε αυτές. Για παράδειγμα, σε ένα δείγμα δεδομένων με μεταβλητή πρώτη τον αριθμό πληθυσμού, μετρημένο ανά μονάδα 10.000 ανθρώπων, και δεύτερη τα διαφημιστικά έξοδα μίας εταιρίας, μετρημένα ανά 1.000 ευρώ σε 100 πόλεις, είναι απαραίτητο οι μεταβλη-

τές αυτές να κεντραριστούν, ώστε η πρώτη μεταβλητή του παραδείγματος να μην αλλοιώσει το αποτέλεσμα και τη γραφική αναπαράσταση των κύριων συνιστωσών (Hastie et al., 2021). Στη σπάνια περίπτωση που οι μεταβλητές είναι μετρημένες στην ίδια κλίμακα, τότε μπορεί να παραληφθεί το παραπάνω στάδιο.

2.3 Υπολογισμός Κύριων Συνιστωσών

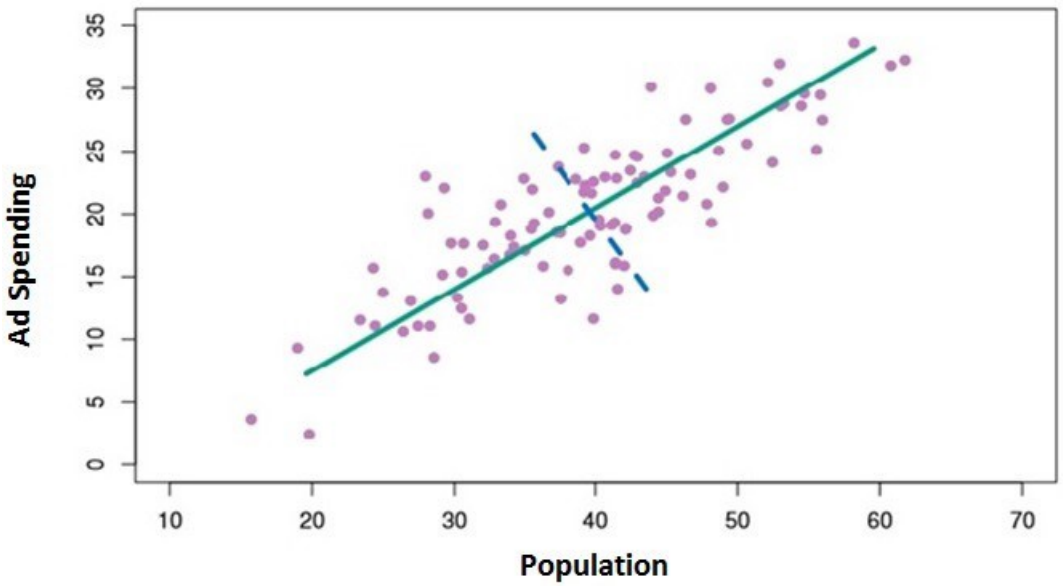
Όπως έχει προλεχθεί, η μέθοδος των Κύριων Συνιστωσών συμβάλλει στη μείωση της διάστασης του διανύσματος των μεταβλητών X_i, όπου $i = 1, ..., p$ του αρχικού δείγματος και κατ΄ επέκταση στην οπτικοποίηση των n παρατηρήσεων εξαιτίας αυτής της μείωσης της διάστασης. Με αυτόν τον τρόπο, αποκτούμε μία δισδιάστατη αναπαράσταση των δεδομένων · γεγονός, που επιτρέπει τη γραφική αναπαράσταση των παρατηρήσεων σε μία πολύ χαμηλότερη διάσταση διατηρώντας τη στατιστική πληροφορία σταθερή (Jolliffe και Cadima, 2016).

Έτσι, το διακύβευμα της μεθόδου είναι η εύρεση νέων μεταβλητών $Z_{1,2,...,j}$ με διάσταση μικρότερη ή ίση με $p(j \leq p)$ και με την ιδιότητα ότι οι πρώτες μεταβλητές Z κατέχουν το μεγαλύτερο βαθμό της αρχικής διακύμανσης. Σε ό,τι αφορά το μαθηματικό υπολογισμό των Z μεταβλητών, είναι γνωστό ότι οι μεταβλητές αυτές αποτελούν γραμμικό συνδυασμό των X_i του αρχικού προβλήματος, όπου σε συμφωνία με την προηγούμενη ενότητα θεωρούμε ότι τα X_i έχουν τεθεί υπό κλίμακα (Hastie et al., 2021). Έτσι, αποκτούμε εύκολα την πρώτη κύρια συνιστώσα Z_1 από τη επίλυση της γραμμικής εξίσωσης:

$$Z_1 = a_{11}X_1 + a_{21}X_2 + a_{31}X_3 + ... + a_{p1}X_p,$$

όπου οι όροι $a_{11}, a_{21}, a_{31}, ..., a_{p1}$ ονομάζονται φορτίσεις (loadings) της πρώτης κύριας συνιστώσας και τηρούν τον περιορισμό της κανονικοποίησής τους για την αποσόβηση του κινδύνου της ανεξέλεγκτης αύξησης του Z_1, και συνεπώς απαιτείται το άθροισμα των τετραγώνων να ισούται πάντα με 1. Ισχύει, δηλαδή, ότι $\sum_{l=1}^{p} a_{l1}^2 = 1$. Γεωμετρικά, οι a_{l1} αναπαριστούν την κατεύθυνση σε ένα χώρο, όπου υπάρχει η μέγιστη μεταβλητότητα των δεδομένων, με αποτέλεσμα η πρώτη κύρια συνιστώσα να ορίζει τη γραμμή καλής προσαρμογής των παρατηρήσεων του δείγματος.

Με το ίδιο σκεπτικό, υπολογίζεται η δεύτερη κύρια συνιστώσα από τη γραμ-



Σχήμα 2.1: *Ανάλυση Κύριων Συνιστωσών για δισδιάστατο δείγμα δεδομένων (Αριθμός Πληθυσμού, μετρημένος ανά μονάδα 10.000 ανθρώπων Population, και Διαφημιστικά Έξοδα μίας εταιρίας, μετρημένα ανά 1.000 ευρώ Ad Spendings σε 100 πόλεις). Η πράσινη γραμμή δηλώνει την κατεύθυνση της πρώτης κύριας συνιστώσας Z_1 , τα σημεία προβάλλονται στη Z_1 και μετράται η απόσταση από τα αρχικά σημεία (Hastie et al., 2021).*

μική εξίσωση:

$$Z_2 = a_{12}X_1 + a_{22}X_2 + a_{32}X_3 + ... + a_{p2}X_p$$

Οι φορτίσεις της δεύτερης κύριας συνιστώσας, εκτός του ότι είναι κανονικοποιημένες και, επομένως, το άθροισμα των τετραγώνων a_{i2} ισούται με 1, ισχύει, επίσης, ότι είναι ορθογώνιες με τις φορτίσεις της πρώτης κύριας συνιστώσας, ήτοι $\sum_{i=1}^{p} a_{i1} \cdot a_{i2} = 0$. Η τελευταία προϋπόθεση εξασφαλίζει ότι η δεύτερη κύρια συνιστώσα είναι ασυσχέτιστη με την πρώτη κύρια συνιστώσα.

Επαγωγικά, λοιπόν, υπολογίζεται εύκολα η $k(\leq j)$ κύρια συνιστώσα Z_k με $k = 1, 2, .., j$ με τη βοήθεια της εξίσωσης:

$$Z_k = a_{1k}X_1 + a_{2k}X_2 + a_{3k}X_3 + ... + a_{pk}X_p$$

με τους περιορισμούς της κανονικοποίησης $\sum_{l=1}^{p} a_{lk}^2 = 1$ και της μη συσχέτισης $\sum_{l=1}^{p} a_{lm} \cdot a_{lk} = 0$, $\forall m \in [1, k-1]$.

Είναι σαφές ότι για την εύρεση των κύριων συνιστωσών $Z_{1,2,...,j}$ χρειάζεται να υπολογιστούν οι φορτίσεις των j κύριων συνιστωσών, με σκοπό να επιτευχθεί η επίλυση των παραπάνω γραμμικών εξισώσεων. Συγκεκριμένα, ξέρουμε ότι οι φορτίσεις της πρώτης κύριας συνιστώσας $a_{11}, a_{21}, a_{31}, ..., a_{p1}$, τις οποίες θεωρούμε ότι είναι στοιχεία του πίνακα στήλης A_1, πρέπει να εκφράζουν τη μέγιστη δειγματική διασπορά.Το γεγονός αυτό σημαίνει ότι οι $a_{11}, a_{21}, a_{31}, ..., a_{p1}$ θα αποτελούν λύσεις του παρακάτω προβλήματος βελτιστοποίησης:

$$\max_{a_{11},a_{21},...,a_{p1}} \{A_1^T S A_1\} \tag{2.1}$$

δοθέντος $A_1^T A_1 = 1$ (κανονικοποιημένες φορτίσεις). Ο πίνακας S θεωρούμε ότι είναι ο τετραγωνικός πίνακας συνδιασποράς των X_i μεταβλητών, και δεδομένου ότι τα στοιχεία X_i είναι μέχρι p σε αριθμό, είναι επόμενο ότι ο πίνακας S έχει διάσταση $p \times p$.
Επαγωγικά, οι φορτίσεις της $k(\leq j)$ κύριας συνιστώσας θα αποτελούν ρίζες του προβλήματος:

$$\max_{a_{1k},a_{2k},...,a_{pk}} \{A_k^T S A_k\} \tag{2.2}$$

με τον περιορισμό $A_k^T A_k = 1$ (Everitt και Hothorn, 2011).

Σε ό,τι αφορά την επίλυση του προβλήματος 2.1, κάνοντας χρήση της μεθόδου του πολλαπλασιαστή Lagrange από τη μαθηματική βελτιστοποίηση καταλήγουμε ότι ο πίνακας A_1 με στοιχεία $a_{11}, a_{21}, ..., a_{p1}$ αποτελεί το ιδιοδιάνυσμα του πίνακα συνδιασποράς $S(v_1)$, το οποίο προέρχεται από την πιο αυξημένη ιδιοτιμή του $S(\lambda_1)$. Αντίστοιχα, από το πρόβλημα βελτιστοποίησης 2.2 λαμβάνουμε την πληροφορία ότι ο πίνακας A_k ισούται με το ιδιοδιάνυσμα v_k του πίνακα S με την αντίστοιχη ιδιοτιμή λ_k,η οποία είναι η k μεγαλύτερη σε σειρά ιδιοτιμή. Παράλληλα, εκτός του γεγονότος ότι οι ιδιοτιμές $(\lambda_{1,2,..k,...p})$ οδηγούν στην επίλυση των παραπάνω προβλημάτων βελτιστοποίησης μέσω των ιδιοδιανυσμάτων και παράγουν τα διανύσματα των φορτίσεων των κύριων συνιστωσών, παίζουν σημαντικό ρόλο και για το λόγο ότι αποτελούν διασπορές των κύριων συνιστωσών $Z_{1,2,..,k,...,j}$. Με αυτόν τον τρόπο, έχουμε ότι $\mathrm{Var}(Z_1) = \lambda_1$, ..., $\mathrm{Var}(Z_k) = \lambda_k$, κ.λ.π., καθώς και ότι το $\sum_{i=1}^{p} \lambda_i = trace(S) = S_1^2 + S_2^2 + ... + S_p^2$, όπου $S_1^2 + S_2^2 + ... + S_p^2$ οι δειγματικές διασπορές των αρχικών μεταβλητών X_i λαμ-

βάνοντάς τες από τον πίνακα συνδιασποράς S. Οπότε, είναι φανερό ότι το άθροισμα των ιδιοτιμών ισούται με τη συνολική διασπορά του αρχικού δείγματος.

Οι παραπάνω υπολογισμοί των φορτίσεων από τον πίνακα συνδιασποράς ισχύουν με την απαίτηση ότι κάθε μία από τις μεταβλητές του αρχικού δείγματος X_i να έχει τεθεί υπό κλίμακα, είτε με σκοπό να έχει μέσο όρο μηδέν, είτε να έχει τυπική απόκλιση ίση με ένα. Στην υπόθεση που αυτό το βήμα έχει παραληφθεί, τότε η «εξαγωγή» και ο υπολογισμός των φορτίσεων των κύριων συνιστωσών από τον πίνακα συνδιασποράς θα οδηγήσουν σε λανθασμένη οπτικοποίηση των κύριων συνιστωσών και κατά συνέπεια σε εσφαλμένα συμπεράσματα. Για αυτόν το λόγο, η χρήση του πίνακα συσχέτισης αντί του πίνακα συνδιασποράς κρίνεται σκόπιμη (Everitt και Hothorn, 2011). Με αυτόν τον τρόπο, υπολογίζονται οι φορτίσεις ως τα ιδιοδιανύσματα του πίνακα συσχέτισης και έπειτα με βάση αυτές υπολογίζονται οι k κύριες συνιστώσες από τη γραμμική σχέση που έχει ήδη αναφερθεί στην αρχή, δηλαδή:

$$Z_k = a_{1k}X_1 + a_{2k}X_2 + a_{3k}X_3 + ... + a_{pk}X_p$$

με τους περιορισμούς της κανονικοποίησης $\sum_{l=1}^{p} a_{lk}^2 = 1$ και της μη συσχέτισης $\sum_{l=1}^{p} a_{lm} \cdot a_{lk} = 0$, $\forall m \in [1, k-1]$.
Έπειτα, όμως, πρέπει οι επαγόμενες φορτίσεις να τεθούν αυτές υπό κλίμακα έτσι, ώστε να αποτελούν συσχετίσεις ή συνδιακυμάνσεις των αρχικών μεταβλητών X_i με τις Z_k κύριες συνιστώσες, με στόχο να γίνει σωστή στατιστική ερμηνεία. Τονίζεται, επίσης, ότι όλη αυτή η παραπάνω διαδικασία ελλοχεύει τον κίνδυνο της μετατροπής όλων των αρχικών μεταβλητών σε εξίσου στατιστικά σημαντικές με αυθαίρετο τρόπο (Everitt και Hothorn, 2011).

Δεδομένου ότι έχουν υπολογιστεί οι φορτίσεις των κύριων συνιστωσών και οι κύριες συνιστώσες, δίνεται η ευκαιρία οπτικοποίησης των δεδομένων σε χαμηλή διάσταση μέσω των γραφικών παραστάσεων της πρώτης κύριας συνιστώσας Z_1 με τη δεύτερη κύρια συνιστώσα Z_2, της πρώτης κύριας συνιστώσας με την τρίτη κύρια συνιστώσα Z_3 αντίστοιχα, της δεύτερης κύριας συνιστώσας με την τρίτη κ.ο.κ. Αυτό σημαίνει ότι σχεδιάζονται σημεία προβαλλόμενα στους υποχώρους, που εκτείνονται κατά τις φορτίσεις των κύριων συνιστωσών a_{l1}, a_{l2}, a_{l3} με $l = 1, 2, ..., p$ κ.ο.κ.

2.4 Αναλογία εξηγούμενης διασποράς

Είναι πλέον αντιληπτό ότι η «πρόκληση» της ανάλυσης κύριων συνιστωσών είναι η μείωση των μεταβλητών του δείγματος με τον περιορισμό ότι οι πρώτες νέες μεταβλητές θα κατέχουν το μεγαλύτερο ποσοστό διασποράς του δείγματος. Έτσι, προκύπτει ο εύλογος προβληματισμός της διαπίστωσης του ποσοστού απώλειας στατιστικής πληροφορίας κατά τη διαδικασία αυτή. Υπονοείται, δηλαδή, πόση διασπορά του αρχικού δείγματος δεν περιέχουν οι πρώτες νέες μεταβλητές του νέου δείγματος, ήτοι οι κύριες συνιστώσες (Hastie et al., 2021).

Για την απάντηση αυτού του ερωτήματος θεωρείται απαραίτητος ο υπολογισμός της αναλογίας διασποράς κάθε κύριας συνιστώσας. Από την προηγούμενη ενότητα έχουμε δει ότι η συνολική διασπορά του αρχικού δείγματος ισούται με το άθροισμα των ιδιοτιμών του πίνακα συνδιασποράς S, καθώς επίσης ότι $\mathrm{Var}(Z_1) = \lambda_1$, ..., $\mathrm{Var}(Z_j) = \lambda_j$, όπου $\lambda_1 > \lambda_2 > ..... > \lambda_j$. Ως εκ τούτου, προκύπτει εύκολα ο υπολογισμός της αναλογίας διασποράς της Z_k κύριας συνιστώσας από τον ακόλουθο τύπο:

$$P_k = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}.$$

Επομένως, οι πρώτες m$< p$ κύριες συνιστώσες κατέχουν την αναλογία:

$$P^{(m)} = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}.$$

Αν πρόκειται για μεταβλητές, οι οποίες δεν έχουν τεθεί υπό κλίμακα και δεν έχουν τυποποιηθεί, τότε υπενθυμίζεται ότι οι φορτίσεις των κύριων συνιστωσών εξάγονται ως τα ιδιοδιανύσματα του πίνακα συσχέτισης και η διασπορά κάθε κύριας συνιστώσας ισούται με την αντίστοιχη ιδιοτιμή του πίνακα συσχέτισης R (Everitt και Hothorn, 2011). Επίσης είναι γνωστό ότι $trace(R) = p$ και άρα καταλήγουμε ότι

$$P^{(m)} = \frac{\sum_{i=1}^{m} \lambda_i}{p}.$$

Είναι φανερό ότι για $m = p$ και, άρα, όταν οι κύριες συνιστώσες είναι σε αριθμό όσες και οι μεταβλητές του αρχικού δείγματος, λαμβάνουμε $P^{(m)} = 1$ και αυτό είναι λογικό, δεδομένου ότι δε θα έχει χαθεί καμία πληροφορία από το

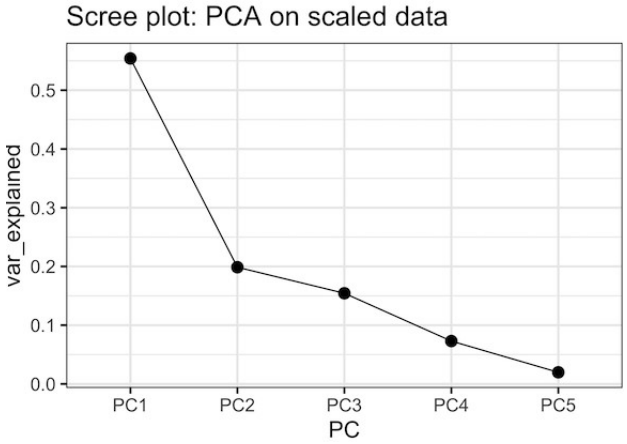
δείγμα, και συνεπώς οι κύριες συνιστώσες θα εξηγούν 100% τη διασπορά του δείγματος. Επιπλέον, αναντίρρητα είναι επιθυμητό οι πρώτες m$< p$ κύριες συνιστώσες να έχουν μία υψηλή αναλογία $P^{(m)}$ κοντά στο 1, ώστε να χάνεται ένας μικρός αριθμός διασποράς. Μάλιστα, ο παράγοντας της αναλογίας διασποράς αποτελεί αρωγό για την εκτίμηση και τον προσδιορισμό της διάστασης m των κύριων συνιστωσών, που επιλέγονται για την περιγραφή του δείγματος.

2.5 Προσδιορισμός αριθμού κύριων συνιστωσών

Ο προσδιορισμός του αριθμού των κύριων συνιστωσών, που χρειάζεται να υπολογιστούν σε ένα πολυμεταβλητό μοντέλο αποτελεί ακανθώδες ζήτημα, εφόσον υπόκειται σε υποκειμενικά κριτήρια και, για αυτό άλλωστε, δεν υπάρχει ένας απλός και ορθός τρόπος, ο οποίος να χρησιμοποιείται αλάνθαστα και να είναι κοινώς αποδεκτός. Το μόνο σίγουρο είναι ότι επιδιώκεται η αναζήτηση όσο δυνατών λιγότερων κύριων συνιστωσών, οι οποίες περιγράφουν με τον καταλληλότερο τρόπο τα δεδομένα της αρχικής ανάλυσης.

Ένας «οδηγός», ο οποίος θα μπορούσε να συμβάλει στην απόφαση αυτή είναι τα διαγράμματα των κύριων συνιστωσών με τις αντίστοιχές τους αναλογίες διασποράς (scree plot) (Hastie et al., 2021). Ειδικότερα, εξετάζουμε αν υπάρχει σημείο, όπου η τιμή της εξηγούμενης διασποράς μειώνεται απότομα με μεγάλη κλίση διαγραμματικά και, στη συνέχεια, μένει σχεδόν επίπεδη δημιουργώντας εμπειρικά ένα βραχίονα. Οπότε, επιλέγουμε το σημείο του διαγράμματος πριν εμφανιστεί η αλλαγή της κλίσης και συνεπώς επιλέγουμε την αντίστοιχη κύρια συνιστώσα, που αντιπροσωπεύει το σημείο αυτό. Η ίδια ακριβώς μέθοδος μπορεί να εφαρμοστεί και σε διαγράμματα των ιδιοτιμών λ_i με το πλήθος i των ιδιοτιμών αντίστοιχα.

Ένας εναλλακτικός τρόπος περιορισμού του αριθμού των κύριων συνιστωσών συνιστά η μέθοδος, κατά την οποία παραλείπονται οι κύριες συνιστώσες, των οποίων οι ιδιοτιμές είναι μικρότερες από τη μέση διασπορά του δείγματος $\frac{\sum_{i=1}^{p}\lambda_i}{p}$. Στην περίπτωση κύριων συνιστωσών και ιδιοτιμών, οι οποίες έχουν υπολογιστεί από τον πίνακα συσχέτισης R και με βάση το γεγονός ότι $\frac{\sum_{i=1}^{p}\lambda_i}{p} = \frac{p}{p} = 1$, προκύπτει η συμφωνία να αγνοούνται οι κύριες συνιστώσες με ιδιοτιμές χαμηλότερες από τη μονάδα (Sharma, 1996). Παράλληλα, αξιο-



Σχήμα 2.2: *Διάγραμμα αναλογιών διασποράς έναντι κύριων συνιστωσών σε Palmer Penguins σετ δεδομένων (Fraser et al., 2014). Θα συμπεραίναμε ίσως ότι η πρώτη κύρια συνιστώσα περιγράφει αρκετά καλά τα δεδομένα.*

σημείωτος είναι και ο εμπειρικός κανόνας, ο οποίος ορίζει ότι είναι χρήσιμο να περιλαμβάνονται μόνο οι κύριες συνιστώσες που περιγράφουν ένα ποσοστό διασποράς των δεδομένων της τάξης των $70\% - 90\%$ (Jolliffe και Cadima, 2016).

2.6 Μοναδικότητα κύριων συνιστωσών

Η μέθοδος της ανάλυσης κύριων συνιστωσών παράγει φορτίσεις κύριων συνιστωσών, οι οποίες είναι μοναδικές, με δεδομένο ότι οι ιδιοτιμές του πίνακα συνδιασποράς διαφέρουν μεταξύ τους. Ενδέχεται οι φορτίσεις που έχουν υπολογιστεί από δύο διαφορετικά υπολογιστικά πακέτα να διαφέρουν κατά πρόσημο αλλά οι απόλυτες τιμές τους να είναι ίσες. Αυτό συμβαίνει, επειδή τα πρόσημα εκφράζουν τον προσανατολισμό σε ένα χώρο, όπου υπάρχει η μέγιστη μεταβλητότητα των δεδομένων. Η εναλλαγή των προσήμων στις φορτίσεις δεν επιφέρει καμία αλλαγή στις παραγόμενες κύριες συνιστώσες, εφόσον ο προσανατολισμός και η κατεύθυνση στον χώρο δε μεταβάλλονται (Hastie et al., 2021).

2.7 Γραφικές παραστάσεις Biplots

Οι γραφικές παραστάσεις Biplots αποτελούν ακρογωνιαίο λίθο της συμπερασματολογίας

σε προβλήματα ανάλυσης κύριων συνιστωσών. Πρόκειται για δισδιάστατα γραφήματα, τα οποία αποτελούν αναπαράσταση των πολυδιάστατων δεδομένων και συνδέουν τις διασπορές, τις συνδιασπορές και τις αποστάσεις των παρατηρήσεων του προβλήματος μεταξύ τους.

Ειδικότερα, η κατασκευή του γραφήματος αποκτάται σχεδιάζοντας τις n(=μέγεθος του αρχικού δείγματος) γραμμές του πίνακα $(\sqrt{n}P_1, \sqrt{n}P_2)$, με τις q (= διάσταση του πίνακα συνδιασποράς) γραμμές του πίνακα $(\sqrt{\frac{\lambda_1}{n}}V_1, \sqrt{\frac{\lambda_2}{n}}V_2)$, όπου :

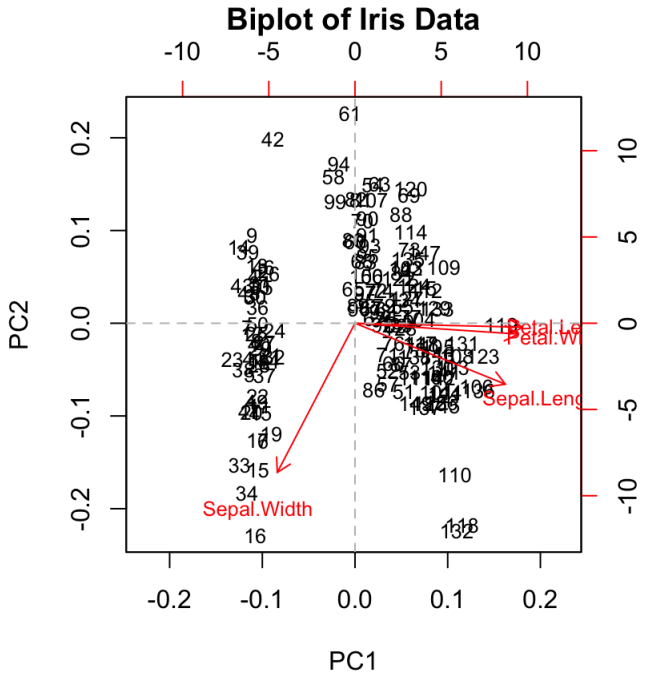
$$P_1 = \frac{1}{\sqrt{\lambda_1}}XV_1, P_2 = \frac{1}{\sqrt{\lambda_2}}XV_2$$

και $\lambda_1, \lambda_2, V_1, V_2$ οι δύο πρώτες ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα του πίνακα nS, όπου S ο πίνακας συνδιασποράς (Everitt και Hothorn, 2011).

Με αυτόν τον τρόπο, μέσω της γραφικής αυτής παράστασης είναι εμφανής η μεταξύ τους Ευκλείδεια απόσταση (αναλογική με την απόσταση Mahalanobis) των παρατηρήσεων, που αντικατροπτίζονται ως σημεία στο γράφημα. Παρατηρείται το μήκος του επαγόμενου διανύσματος, το οποίο αναπαριστά μία μεταβλητή και υποδηλώνει τη διασπορά της μεταβλητής αυτής, καθώς και η γωνία των επαγόμενων διανυσμάτων δύο μεταβλητών, η οποία αντανακλά τη συσχέτιση των μεταβλητών αυτών, εφόσον η σχέση της γωνίας με το συντελεστή συσχέτισης των μεταβλητών είναι αντιστρόφως ανάλογη. Αντίστοιχα, το ίδιο ισχύει και με τη γωνία μεταξύ ενός επαγόμενου διανύσματος μίας μεταβλητής με τους άξονες, που αναπαριστούν τις κύριες συνιστώσες. Έτσι, όσο μεγαλύτερη είναι αυτή η γωνία τόσο μικρότερη είναι η συσχέτιση της εκάστοτε μεταβλητής με την αντίστοιχη κύρια συνιστώσα. Για αυτά τα παραπάνω χαρακτηριστικά, τα γραφήματα Biplots κρίνονται ως μία διαδικασία υψίστης σημασίας για την ανάλυση κύριων συνιστωσών.

Εφαρμόζοντας τις παραπάνω σκέψεις στο σχήμα 2.3 για το δείγμα δεδομένων Iris, το οποίο αποτελείται από 50 δείγματα των τριών ειδών του φυτού Ίρις ((Iris setosa, Iris virginica και Iris versicolor), και στα οποία αυτά τα δείγματα έχουν μετρηθεί το μήκος και το πλάτος των σεπάλων και των πετάλων τους, διαπιστώνεται ότι η γωνία των επαγόμενων διανυσμάτων των μεταβλητών Petal.Length και Petal.Width είναι πάρα πολύ μικρή. Αυτό το γεγονός οδηγεί στο συμπέρασμα ότι υπάρχει πολύ υψηλή συσχέτιση των μεταβλητών αυτών και,

άρα, φαίνεται ότι το μήκος των πετάλων της Ίριδος είναι άρρικτα συνδεδμένο με το πλάτος των πετάλων της. Επιπλέον, είναι εμφανές ότι οι γωνίες των δύο προαναφερθεισών μεταβλητών με τη μεταβλητή Sepal.Length είναι μικρές, με αποτέλεσμα να υπάρχει, εν τέλει, υψηλή συσχέτιση μεταξύ αυτών των τριών μεταβλητών. Αντίθετα, οι γωνίες που σχηματίζονται από αυτές τις τρεις μεταβλητές με τη μεταβλητή Sepal.Width είναι μεγάλες καταλήγοντας στο συμπέρασμα ότι η μεταβλητή Sepal.Width έχει πιθανόν χαμηλή συσχέτιση με τις υπόλοιπες. Συμπερασματικά, θα μπορούσε κάποιος να αποφανθεί με τη βοήθεια του σχήματος 2.3 ότι υπάρχει υψηλή εξάρτηση μεταξύ του μήκους και του πλάτους των πετάλων και του μήκους των σεπάλων, ενώ αυτό το παραπάνω γενονός φαίνεται να μην ισχύει όσον αφορά στο πλάτος των σεπάλων.



Σχήμα 2.3: *Διάγραμμα Biplot στο σετ δεδομένων Iris (Anderson, 1935) και (Fisher, 1936).*

2.8 Χρήση Κύριων Συνιστωσών κατά τη συμπλήρωση πινάκων δεδομένων

Είναι πλέον σύνηθες να λείπουν παρατηρήσεις σε δείγματα δεδομένων, και να μην είναι συμπληρωμένα,δηλαδή, όλα τα στοιχεία του πίνακα δεδομένων της μελέτης. Αυτό μπορεί να συμβαίνει για διάφορους λόγους, είτε επειδή κατά τη διάρκεια της μελέτης δεν έχει επιτευχθεί η συλλογή όλων των επιθυμητών δεδομένων εξαιτίας της υψηλής πολυμεταβλητότητας του προβλήματος, είτε επειδή τα βοηθητικά μηχανήματα και τα όργανα του πειράματος εμφανίζουν κάποια βλάβη με αποτέλεσμα τη μη λήψη της αντίστοιχης πληροφορίας από τη μέτρηση αυτή, είτε επειδή η πληροφορία μπορεί να έχει ουδέποτε καταγραφεί. Ωστόσο, αυτά τα παραπάνω δείγματα αποτελούν αναπόσπαστο στοιχείο της ανάλυσης τόσο της μη επιβλεπόμενης όσο και της επιβλεπόμενης μάθησης όπως στη γραμμική παλινδρόμηση, με αποτέλεσμα η συμπλήρωση των κενών στοιχείων να είναι αναγκαία. Ιδιαίτερο ρόλο, επιπλέον, παίζει η συμπλήρωση πινάκων δεδομένων στα συστήματα σύστασης.

Οι κύριες συνιστώσες συνδράμουν στη μέθοδο της συμπλήρωσης πινάκων υπολογίζοντας τα στοιχεία x_{ij}, που λείπουν. Θεωρείται κατάλληλο να οριστεί εναλλακτικά το πρόβλημα βελτιστοποίησης των κύριων συνιστωσών (εξίσωση 2.2) με τη βοήθεια της Ευκλείδειας απόστασης με στόχο τη διευκόλυνση της συμπλήρωσης των πινάκων. Έστω X ο πίνακας δεδομένων και x_{ij} τα στοιχεία του. Τότε τα στοιχεία z_{ik} της Z_k κύριας συνιστώσας θα δίνονται ως εξής:

$$z_{ik} = a_{1k}x_{i1} + a_{2k}x_{i2} + a_{3k}x_{i3} + ... + a_{pk}x_{ip}.$$

Οπότε, δεδομένου ότι οι κύριες συνιστώσες είναι m σε αριθμό, προκύπτει εύκολα η αναπαράσταση των στοιχείων ως:

$$x_{ij} = \sum_{m=1}^{M} z_{im}a_{jm}.$$

Το πρόβλημα βελτιστοποίησης της εξίσωσης 2.2 μετατρέπεται με στόχο την ελαχιστοποίηση της απόστασης των στοιχείων των δεδομένων από τα στοιχεία των κύριων συνιστωσών ως εξής:

$$\min_{B\in\mathbb{R}^{n\times M}, C\in\mathbb{R}^{p\times M}} \{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \sum_{m=1}^{M} b_{im}c_{jm})^2\}, \tag{2.3}$$

όπου B και C οι πίνακες με τα στοιχεία τους να αποτελούν λύσεις του παραπάνω προβλήματος ελαχιστοποίησης (Hastie et al., 2021). Αποδεικνύεται ότι για $b_{im} = z_{im}$ και $c_{jm} = a_{jm}$ παίρνουμε τη λύση του προβλήματος βελτιστοποίησης και επομένως έχουμε:

$$\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \sum_{m=1}^{M} z_{im}a_{jm})^2\}. \tag{2.4}$$

Στην περίπτωση ελλιπουσών παρατηρήσεων, τότε το πρόβλημα βελτιστοποίησης 2.4 λαμβάνει τη μορφή:

$$\min_{B\in\mathbb{R}^{n\times M}, C\in\mathbb{R}^{p\times M}}\{\sum_{(i,j)\in O}(x_{ij} - \sum_{m=1}^{M} b_{im}c_{jm})^2\}, \tag{2.5}$$

όπου O είναι ο πίνακας με όλες τις παρατηρήσεις x_{ij} που είναι συμπληρωμένες (Hastie et al., 2021). Έτσι, από την επίλυση του παραπάνω προβλήματος 2.5 βρίσκουμε τις λύσεις b_{im} και c_{jm} και κατα συνέπεια τις ελλιπούσες x_{ij}, δεδομένου ότι $x_{ij} = \sum_{m=1}^{M} b_{im}c_{jm}$. Τέλος, γίνεται εφικτός μέσω των b_{im} και c_{jm} ο προσδιορισμός των κύριων συνιστωσών z_{im} και των φορτίσεών τους a_{jm}.

Εφόσον η «δια χειρός» επίλυση του προβλήματος 2.5 κρίνεται δύσκολη έως ακατόρθωτη, υπάρχει ο παρακάτω υπολογιστικός αλγόριθμος που διευκολύνει τη διαδικασία εύρεσης της λύσης.

Ο αλγόριθμος έχει ως εξής σύμφωνα με τους Hastie et al. (2010):

1. Έστω $Q \in \mathbb{R}^{n\times p}$ με στοιχεία x_{ij} τις παρατηρήσεις του δείγματος και στις κενές θέσεις του πίνακα, όπου δεν υπάρχει η παρατηρήση του δείγματος εισάγουμε τη μέση δειγματική τιμή της αντίστοιχης στήλης, όπου υπάρχει η κενή θέση.

2. Έπειτα, ορίζουμε την επαναληπτική διαδικασία επίλυσης του προβλήματος

$$\min_{B\in\mathbb{R}^{n\times M}, C\in\mathbb{R}^{p\times M}}\{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \sum_{m=1}^{M} b_{im}c_{jm})^2\},$$

όπου μοιάζει με το πρόβλημα βελτιστοποίησης της εξίσωσης 2.5, με τη διαφορά ότι στην προκειμένη περίπτωση όλες οι τιμές των στοιχείων x_{ij} είναι υπαρκτές.

(α′) Με αυτόν τον τρόπο, βρίσκονται οι τιμές των b_{im} και c_{jm} και άρα προσδιορίζονται οι κύριες συνιστώσες του πίνακα X.

(β′) Επιβάλλουμε **στις μη παρατηρούμενες τιμές** x_{ij}, τις οποίες στο βήμα 1 είχαμε ορίσει ως τη μέση τιμή των παρατηρήσεων της αντίστοιχης στήλης, να ισούνται με:

$$x_{ij} = \sum_{m=1}^{M} b_{im} c_{jm}$$

(γ′) Ύστερα υπολογίζεται η τιμή του αθροίσματος

$$\sum_{(i,j)\in O} (x_{ij} - \sum_{m=1}^{M} b_{im} c_{jm})^2,$$

όπου O είναι ο πίνακας με όλες τις παρατηρήσεις x_{ij} που είναι συμπληρωμένες, b_{im} και c_{jm} είναι τα στοιχεία, που έχουν υπολογιστεί στο βήμα (2)-(α). Αφ′ ης στιγμής η ευκλείδεια απόσταση των παρατηρήσεων από τις κύριες συνιστώσες δε μικραίνει άλλο, και άρα η τιμή του παραπάνω αθροίσματος δε μειώνεται άλλο, η επαναληπτική διαδικασία τερματίζει, και κατά συνέπεια λαμβάνουμε τις τιμές x_{ij}, που έλειπαν από το δείγμα έχοντας ήδη εκτελέσει το βήμα (2)-(β).

Είναι σαφές ότι για τη χρήση του παραπάνω αλγόριθμου χρειάζεται η εκ των προτέρων επιλογή της τιμής της μεταβλητής M, ήτοι του πλήθους των κύριων συνιστωσών. Μία προσέγγιση αυτού του προβλήματος αποτελεί η τυχαία επιλογή της τιμής της με γνώμονα τη μη επιρροή σε μεγάλο βαθμό στις γνωστές τιμές του πίνακα από την αφαίρεση των επιπρόσθετων στοιχείων του πίνακα X (Hastie et al., 2021).

Κεφάλαιο 3

Ανάλυση Συστάδων

3.1 Εισαγωγή

Η Ανάλυση Συστάδων (Cluster analysis) κατέχει, και αυτή με τη σειρά της, ιδιαίτερη και εξίσου σημαντική θέση στο πλαίσιο της μη επιβλεπόμενης μάθησης. Ως μέρος της μη επιβλεπόμενης μάθησης, δε θέτει ως στόχο την πρόβλεψη κάποιας τιμής απόκρισης, αλλά αποτελεί μηχανισμό απλοποίησης πολυμεταβλητών δεδομένων, και αποσκοπεί στην εύρεση και στην ανακάλυψη συστάδων και υποομάδων ανάμεσα στις ποικίλες παρατηρήσεις του πολυμεταβλητού δείγματος. Η παραπάνω ομαδοποίηση γίνεται με βάση κάποια κοινά χαρακτηριστικά που εμφανίζουν οι παρατηρήσεις και οι μεταβλητές στο δείγμα και, με αυτόν τον τρόπο, οι παρατηρήσεις, οι οποίες ανήκουν στην ίδια ομάδα έχουν κοινό γνώρισμα, ενώ αυτές που ανήκουν σε διαφορετικές, δεν έχουν (Hastie et al., 2021). Οι τεχνικές της Ανάλυσης Συστάδων στα πλαίσια της μη επιβλεπόμενης μάθησης πρωτοδιατυπώθηκαν και δημοσιεύτηκαν από τους Jain και Dubes (1988).

Η ειδοποιός διαφορά μεταξύ της Ανάλυσης Κύριων Συνιστωσών και της Ανάλυσης Συστάδων είναι ότι η πρώτη έχει ως κεντρικό άξονα τη μείωση της διάστασης των μεταβλητών του δείγματος και την αντίστοιχη οπτικοποίηση των παρατηρήσεων σε έναν χώρο πολύ μικρότερης διάστασης, στον οποίο η χαμένη στατιστική πληροφορία είναι αμελητέα και η εξηγούμενη διασπορά του δείγματος είναι αρκετά υψηλή. Στον αντίποδα, η Ανάλυση Συστάδων έχει συνήθως ως «μέλημα» την ομογενή ομαδοποίηση των παρατηρήσεων του δείγματος.

Οι εφαρμογές της Ανάλυσης Συστάδων είναι πολλαπλές σε διάφορα επιστημονικά πεδία. Για παράδειγμα, οι τεχνικές της Ανάλυσης Συστάδων αποτελούν εργαλεία στους κόλπους ιατρικών και βιολογικών ερευνών και μελετών βρίσκοντας ομοιότητες ανάμεσα στις κλινικές ή πειραματικές μετρήσεις καθώς και ανάμεσα στους ασθενείς ή στα πειραματόζωα. Επιπλέον, συνιστούν αναπόσπαστο κομμάτι του Οικονομικού τομέα του marketing, δεδομένου ότι διευκολύνουν τον προσδιορισμό του καταμερισμού της αγοράς κατηγοριοποιώντας τους καταναλωτές με κοινά χαρακτηριστικά, με στόχο την πιο αποτελεσματική διαφήμιση των προϊόντων σε κοινό, το οποίο πιθανώς να ενδιαφέρεται περισσότερο για τα συγκεκριμένα είδη αγαθών (Hastie et al., 2021). Παράλληλα, μείζονα ρόλο παίζουν και σε εφαρμογές τμηματοποίησης εικόνων, αναγνώρισης αντικειμένων, ανάκτησης πληροφοριών και εξόρυξης δεδομένων (Flynn et al., 1999). Έτσι, αν και η αντίληψη ότι η σπουδαιότητα της Ανάλυσης Συστάδων στην καθημερινή ζωή είναι σπουδαία δεν κομίζει γλαύκα ες Αθήνας, χρειάζεται να σημειωθεί ότι απαιτείται ιδιαίτερη προσοχή στην εφαρμογή των μεθόδων και των τεχνικών της Ανάλυσης Συστάδων, καθώς δεν υπάρχει μία αποκλειστική μέθοδος, η οποία να έχει σωστή και κατάλληλη εφαρμογή σε όλους τους τύπους δεδομένων (Everitt και Hothorn, 2011).

Οι προσεγγίσεις των τεχνικών της Ανάλυσης Συστάδων διακρίνονται κυρίως σε δύο βασικές, ήτοι σε αυτήν, η οποία είναι γνωστή ως K-means ανάλυση συστάδων, η οποία αποτελεί τμήμα της διαχωριστικής ανάλυσης συστάδων, καθώς και στην Ιεραρχική Ανάλυση Συστάδων (hierarchical clustering). Βασικό γνώρισμα της πρώτης προσέγγισης είναι ότι επιδιώκει την ομαδοποίηση των παρατηρήσεων γνωρίζοντας a priori τον αριθμό των συσταθέντων ομάδων, βρίσκοντας βέβαια τον αριθμό αυτόν που κρίνεται καταλληλότερος έπειτα από (πολλαπλές) δοκιμές. Αντίθετα, το παραπάνω γεγονός δεν ισχύει στην ιεραρχική ανάλυση συστάδων, η οποία αναπαριστά τα δεδομένα σε μορφή δενδρογράμματος διαχωρίζοντας τα δεδομένα σε κλαδιά χωρίς την εκ των προτέρων πληροφορία του αριθμού των ομάδων (Hastie et al., 2021). Επίσης, σημαντική μέθοδος της Ανάλυσης Συστάδων συνιστά και η Ανάλυση Συστάδων Βασισμένη σε μοντέλο (Model-based clustering), η οποία ασχολείται με την εύρεση στατιστικού μοντέλου, το οποίο χαρακτηρίζει το δείγμα των δεδομένων και, έτσι, με τη βοήθεια του μοντέλου καθορίζεται ο αριθμός των συστάδων (Everitt και Hothorn, 2011).

3.2 K-means Ανάλυση Συστάδων

Η K-means ανάλυση συστάδων αποτελεί, όπως έχει προειπωθεί, μία από τις μεθόδους τμηματοποίησης και ομαδοποίησης δεδομένων μεγάλου όγκου σε Κ ξεχωριστές συστάδες, οι οποίες οφείλουν να μην αλληλοκαλύπτονται η μία από την άλλη. Όσον αφορά στην εφαρμογή της μεθόδου απαιτείται η a priori εκτίμηση του αριθμού Κ των συστάδων κάνοντας δοκιμές και βρίσκοντας τον αριθμό αυτόν, που ταιριάζει καλύτερα στο μοντέλο, με σκοπό την κατάταξη των δεδομένων σε αυτές.

Η K-means ανάλυση συστάδων έχει τρεις βασικές αρχές. Η πρώτη απαιτεί κάθε παρατήρηση του δείγματος να ανήκει σε τουλάχιστον μία συστάδα. Έτσι, αν $C_1, C_2, ..., C_K$ συνιστούν δείγματα, τα οποία περιλαμβάνουν τους δείκτες των παρατηρήσεων κάθε συστάδας, τότε προκύπτει ότι πρέπει να ισχύει ότι $C_1 \cup C_2 \cup ... \cup C_K = \{1, ..., n\}$, όπου n ο αριθμός των παρατηρήσεων του δείγματος. Η δεύτερη αρχή σχετίζεται με τη μη αλληλοκάλυψη των συστάδων μεταξύ τους και, άρα, πρέπει κάθε παρατήρηση να μην ανήκει σε περισσότερες από μία συστάδα. Επομένως, υπάρχει η αναγκαιότητα να ισχύει ότι $C_1 \cap C_2 \cap ... \cap C_K = \emptyset$. Τέλος, η τρίτη βασίζεται στον περιορισμό της διασποράς εντός μίας συστάδας να είναι αρκετά μικρή (Hastie et al., 2021). Με άλλα λόγια, κρίνεται αναγκαία η επίλυση του παρακάτω προβλήματος βελτιστοποίησης:

$$\min_{C_1, C_2, ..., C_K} \{\sum_{k=1}^{K} W(C_k)\}, \tag{3.1}$$

όπου $W(C_k)$ η διασπορά κάθε $k(\leq K)$ συστάδας. Για την επίλυση του προβλήματος 3.2 είναι πασιφανές ότι πρέπει πρώτα να ορισθεί μαθηματικά η έννοια της διασποράς εντός συστάδας και άρα να δοθεί ο μαθηματικός τύπος του $W(C_k)$. Με τη συνδρομή της Ευκλείδειας απόστασης το $W(C_k)$ ορίζεται ως

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{3.2}$$

όπου $|C_k|$ ο αριθμός των παρατηρήσεων εντός της k συστάδας.

Συνεπώς, με τη βοήθεια της εξίσωσης 3.2 το πρόβλημα βελτιστοποίησης παίρνει τη μορφή:

$$\min_{C_1, C_2, ..., C_K} \{\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\} \tag{3.3}$$

Εξαιτίας του δυσεπίλυτου προβλήματος 3.3, δεδομένου ότι υπάρχουν K^n τρόποι ομαδοποίησης των n παρατηρήσεων σε K συστάδες, ειδικά όταν πρόκειται για δεδομένα μεγάλου όγκου, έχει κατασκευαστεί ο παρακάτω αλγόριθμος, ο οποίος δίνει μία αρκετά καλή προσέγγιση της λύσης του προβλήματος βελτιστοποίησης (Hastie et al., 2021).

Ο αλγόριθμος έχει ως εξής σύμφωνα με τους Hastie et al.(2021):

1. Επιλέγουμε αυθαίρετα για κάθε παρατήρηση του δείγματος σε ποιά συστάδα ανήκει. Έτσι, συνδέουμε κάθε παρατήρηση με έναν αριθμό k, όπου $1 \leq k \leq K$, όπου Κ ο αριθμός συστάδων, που έχει ήδη οριστεί.

2. Έπειτα, για κάθε συστάδα υπολογίζεται το κέντρο βάρους της. Το κέντρο βάρους της k συστάδας αποτελεί το διάνυσμα της μέσης τιμής των p στοιχειώντων παρατηρήσεων, οι οποίες ανήκουν στην k συστάδα. Έτσι, επιτυγχάνεται η ελαχιστοποίηση της Ευκλείδειας απόστασης του προβλήματος 3.3.

3. Ύστερα, τοποθετείται κάθε παρατήρηση σε συστάδα, της οποίας το κέντρο βάρους είναι πιο κοντά στην παρατήρηση σε συμφωνία με την έννοια της Ευκλείδειας απόστασης.

Η παραπάνω επαναληπτική διαδικασία τερματίζει τη στιγμή, στην οποία κάθε παρατήρηση ανήκει σε μία συστάδα και δεν υπάρχει λόγος επανατοποθέτησής της σε άλλη με γνώμονα το βήμα (3) του αλγόριθμου, και κατά συνέπεια, στην οποία το αποτέλεσμα του αλγορίθμου δεν αλλάζει πλέον. Παράλληλα, κρίνεται σκόπιμο να επισημανθεί ότι ο αλγόριθμος προσδιορίζει περισσότερο τοπικά ελάχιστα παρά ολικά. Αυτό συμβαίνει λόγω της αυθαίρετης τοποθέτησης των στοιχείων στις συστάδες κατά το βήμα (1) του αλγορίθμου. Ως εκ τούτου, είναι χρήσιμο κάποιος να τρέξει πολλαπλές φορές τον εν λόγω αλγόριθμο με διαφορετικές αρχικές συνθήκες κάθε φορά έτσι, ώστε να αποκτηθεί η καλύτερη προσέγγιση του προβλήματος ελαχιστοποίησης του προβλήματος 3.3.

Η K-means ανάλυση συστάδων, αν και θεωρείται αρκετά κατάλληλη προς εφαρμογή σε πολλά δείγματα, εγκυμονεί δύο βασικά προβλήματα, τα οποία δεν μπορούν να παραληφθούν. Το πρώτο αφορά στη μη ανεξαρτησία της μεθόδου ως προς την κλίμακα, στην οποία έχουν τεθεί τα δεδομένα. Κατ΄ αυτόν τον τρόπο, ελλοχεύει ο κίνδυνος της εύρεσης διαφορετικών αποτελεσμάτων για τυποποιημένα δείγματα δεδομένων. Επιπλέον, ένα άλλο μειονέκτημα είναι ότι η

μέθοδος αυτή προσδίδει στις συστάδες εξ ορισμού ένα σφαιρικό σχήμα, το οποίο ενδέχεται να μην προσιδιάζει σε αυτές στην πραγματικότητα, εφόσον είναι πιθανό να έχουν μία διαφορετική δομή (Everitt και Hothorn, 2011).

3.3 Ιεραρχική Ανάλυση Συστάδων

Η ιεραρχική ανάλυση συστάδων συνιστά, όπως έχει ήδη αναφερθεί, μία εναλλακτική μέθοδος ομαδοποίησης και ταξινόμησης των δεδομένων σε συστάδες. Σε αντιδιαστολή με τη μέθοδο K-means, η ιεραρχική δεν απαιτεί την εκ των προτέρων πρόβλεψη του αριθμού των συστάδων, ήτοι του αριθμού K. Ιδιαίτερο χαρακτηριστικό της αποτελεί το εφόδιο, που αποφέρει, το οποίο δεν είναι άλλο παρά η απεικόνιση των παρατηρήσεων σε δενδρογράμματα. Η πιο συνηθισμένη μορφή της αποτελεί η ιεραρχική σωρευτική ανάλυση συστάδων agglomerative hierarchical clustering, στην οποία το δενδρόγραμμα, που προκύπτει, κατασκευάζεται πρώτα από τα φύλλα θέτοντας τις συστάδες πάνω στους κορμούς. Παράλληλα η ιεραρχική ανάλυση συστάδων διακρίνεται και στη διχαστική ιεραρχική ανάλυση συστάδων divisive hierarchical clustering, της οποίας η διαδικασία είναι εκ διαμέτρου αντίθετη με αυτήν της σωρευτικής, με την έννοια ότι το απεικονισμένο δενδρόγραμμα σχεδιάζεται αρχικά από τον κορμό και καταλήγει προς τα φύλλα (Hochreiter, 2014).

Για να αποκτηθεί ένα δενδρόγραμμα της ιεραρχικής σωρευτικής ανάλυσης, γίνεται χρήση ενός αλγορίθμου, ο οποίος βασίζεται στη μέτρηση της διαφορετικότητας κάθε ζεύγους παρατηρήσεων του δείγματος, εφόσον πρόκειται για ποσοτικές μεταβλητές, μέσω συνήθως του μαθηματικού εργαλείου της Ευκλείδειας απόστασης, καθώς και αρχίζει ανάποδα από τα φύλλα του δενδρογράμματος. Με αυτόν τον τρόπο, αρχικά, κάθε παρατήρηση ανήκει σε μία ξεχωριστή δική της συστάδα και, εξού, υπάρχουν συνολικά n συστάδες, όσο δηλαδή και το πλήθος των παρατηρήσεων. Έπειτα, δύο συστάδες, οι οποίες φαίνεται να είναι όμοιες, ενώνονται μεταξύ τους αφαιρώντας, έτσι, μία συστάδα και μετατρέποντας το συνολικό αριθμό συστάδων σε $n-1$. Αντίστοιχα, με την ίδια λογική ξανά, δύο συστάδες με ομοιότητες συνδέονται μεταξύ τους και πλέον υπάρχουν $n-2$. Ο αλγόριθμος συνεχίζει με το ίδιο σκεπτικό και τερματίζει όταν τελικά όλες οι παρατηρήσεις ανήκουν σε μία μόνο συστάδα. Πιο αναλυτικά, ο αλγόριθμος της ιεραρχικής ανάλυσης συστάδων έχει ως εξής σύμφωνα με τους Hastie et al. (2021) :

1. Έστω n παρατηρήσεις και έστω μία μέτρηση κατά ζεύγη διαφορετικότητας (Ευκλείδειας απόστασης) όλων των $n(n-1)/2$. Κάθε παρατήρηση ανήκει σε μία συστάδα μόνη της.

2. Έπειτα, για κάθε $i = n, n-1, n-2, ..., 2$

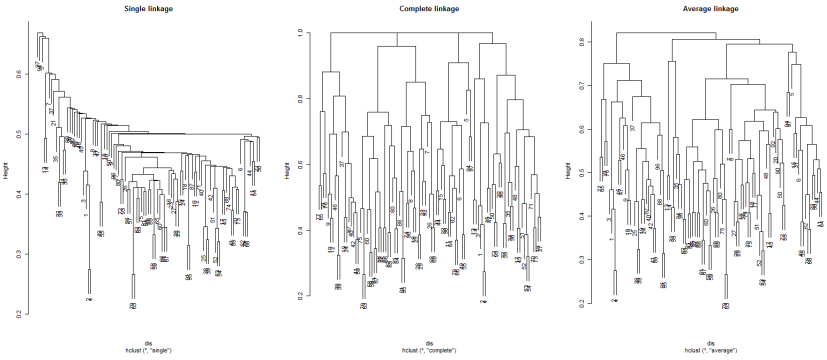
   (α′) εξετάζεται η διαφορετικότητα της συστάδας i με τις υπόλοιπες και εντοπίζεται το ζεύγος συστάδων με τη μικρότερη έως αμελητέα διαφορετικότητα. Ύστερα, αυτό το ζεύγος ενώνεται σε ύψος του δενδρογράμματος, το οποίο σχετίζεται με το μέγεθος διαφορετικότητάς τους.

   (β′) Μετά, υπολογίζονται εκ νέου οι διαφορετικότητες μεταξύ των άλλων συστάδων, που έχουν απομείνει, ήτοι των $i-1$.

Δυστυχώς, ένα αρκετά σημαντικό μειονέκτημα του παραπάνω αλγορίθμου είναι ότι δεν εμπεριέχει και δε συνυπολογίζει την περίπτωση, στην οποία σε μία συστάδα ανήκουν δύο ή περισσότερες παρατηρήσεις και, κατ΄ επέκταση, χρειάζεται να υπολογισθεί η διαφορετικότητα αυτής της συστάδας με, πλέον, μία ομάδα παρατηρήσεων με μία άλλη συστάδα με μία ή πολλαπλές παρατηρήσεις. Εξ αυτού ορμώμενοι, και για την αποσόβηση ενός κινδύνου τέτοιου σφάλματος, έχει αναπτυχθεί η έννοια της ζεύξης, η οποία ορίζει τη διαφορετικότητα μεταξύ δύο ομάδων παρατηρήσεων. Οι τέσσερεις πιο σημαντικές κατηγορίες της έννοιας αυτής είναι η ολοκληρωμένη, μονή, μέση - ή αλλιώς γνωστή ως Unweighted Pair Group Method using arithmetic Averages (UPGMA) στον κόσμο της Βιοπληροφορικής- και η κέντρου βάρους ζεύξη.

Αρχικά, στην ολοκληρωμένη ζεύξη δίνεται έμφαση στη μέγιστη διαφορετικότητα μεταξύ των συστάδων, και για αυτόν το λόγο, υπολογίζονται όλες οι διαφορετικότητες ανάμεσα στις ομάδες παρατηρήσεων και καταγράφεται αυτή, που έχει μεγαλύτερη τιμή. Έτσι, η ζεύξη αυτή αναζητά για δύο συστάδες Α και Β $d_{max}(A,B) = \max_{\alpha \in A, \beta \in B} ||\alpha - \beta||$, όπου α και β στοιχεία των αντίστοιχων συστάδων και $||.||$ η αντίστοιχη απόσταση- είτε Ευκλείδεια, είτε Μανχάταν, είτε Mahalanobis. Αναφορικά με τη μονή ζεύξη, η διαδικασία είναι εκ διαμέτρου αντίθετη με αυτήν της ολοκληρωμένης και, έτσι, καταγράφεται η διαφορετικότητα των ομάδων με τη μικρότερη τιμή κυριαρχώντας στη διαδικασία αυτή η ελάχιστη διαφορετικότητα μεταξύ συστάδων, και άρα ισχύει ότι $d_{min}(A,B) = \min_{\alpha \in A, \beta \in B} ||\alpha - \beta||$. Επιπρόσθετα, αυτή η σύνδεση θα μπορούσε να χαρακτηρισθεί εκτεταμένη και χρονοβόρα, εφόσον οι μονές

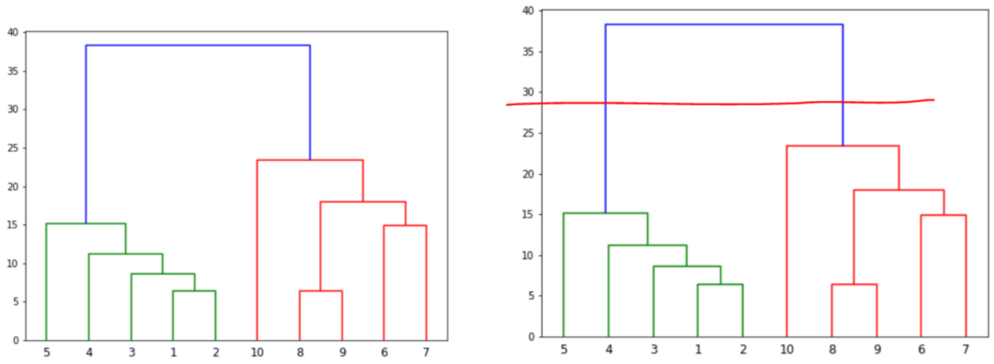
παρατηρήσεις ενώνονται μία μία κάθε φορά. Σχετικά με τη μέση ζεύξη, υπερισχύει η μέση τιμή της διαφορετικότητας εντός συστάδας, δεδομένου ότι υπολογίζονται οι διαφορετικότητες μεταξύ των ομάδων και αποθηκεύεται η μέση τιμή αυτών των διαφορετικοτήτων. Σε αυτήν την περίπτωση αναζητείται $d_{avg}(A, B) = \frac{1}{n_A n_B} \sum_{\alpha \in A, \beta \in B} ||\alpha - \beta||$, όπου n_A και n_B το πλήθος των στοιχείων στην αντίστοιχη συστάδα. Αυτές οι τρεις διαδικασίες, ήτοι η ολοκληρωμένη, μονή και η μέση ζεύξη κατέχουν πρωταρχική θέση στις μελέτες των στατιστικών, και μεταξύ αυτών συχνά προτιμώνται η μονή και η μέση ζεύξη, δεδομένου ότι τείνουν να προσφέρουν πιο ισορροπημένα δενδρογράμματα. Όσον αφορά στην κέντρου βάρους ζεύξη, γίνεται ο υπολογισμός της διαφορετικότητας μεταξύ των κέντρου βάρους της κάθε ομάδας παρατηρήσεων, αλλά έχει ως παράπλευρο κόστος τον κίνδυνο της αντιμετάθεσης ή της αντιστροφής, όπου το ύψος των συστάδων, που ενώνονται, είναι χαμηλότερο από αυτό των συστάδων αυτών, όταν ήταν ξεχωριστές, και οδηγώντας τυχόν σε εσφαλμένη δημιουργία και ερμηνεία ενός δενδρογράμματος (Hastie et al., 2021).



Σχήμα 3.1: *Δενδρόγραμμα ενός δείγματος δεδομένων με μονή, ολοκληρωμένη και μέση ζεύξη.*

Ως εκ τούτου, είναι αντιληπτό και από το σχήμα 3.1, ότι ένα δενδρόγραμμα, είναι πολύ πιθανό, να διαφέρει για το ίδιο δείγμα δεδομένων αλλά για διαφορετικό τρόπο ζεύξης. Έτσι, κρίνεται σκόπιμο να επισημανθεί η σπουδαιότητα και η σημασία της επιλογής της κατάλληλης διαδικασίας σύνδεσης, αφού φαίνεται να ασκεί αρκετά μεγάλη επιρροή στο τελικό αποτέλεσμα. Για αυτόν το λόγο, θα ήταν χρήσιμες οι πολλαπλές δοκιμές εφαρμογών διαφορετικών διαδικασιών σύνδεσης στα ίδια δεδομένα, όταν πρόκειται για αυθαίρετη επιλογή σύνδεσης.

Ο αλγόριθμος, ο οποίος ανταποκρίνεται στις απαιτήσεις της διχαστικής ιεραρχικής ανάλυσης αφίσταται παρασάγγας του παραπάνω αλγορίθμου, ο οποίος έχει ισχύ στα πλαίσια της ιεραρχικής σωρευτικής ανάλυσης συστάδων. Αυτό το γεγονός οφείλεται πρωτίστως στον τρόπο κατασκευής του δενδρογράμματος, το οποίο προκύπτει αρχίζοντας από το πάνω μέρος του δενδρογράμματος (κορμός του δένδρου) και καταλήγει στα φύλλα του δένδρου. Έτσι, το πρώτο βήμα του αλγορίθμου είναι να συμπεριλάβει όλες τις παρατηρήσεις του δείγματος σε μία συστάδα. Έπειτα, γίνεται ο διαχωρισμός των συστάδων με γνώμονα τη διαφορετικότητα των παρατηρήσεων, και η εν λόγω διαδικασία τερματίζει αν και μόνο αν έχουν ομαδοποιηθεί έτσι, ώστε να έχουν μείνει αποκλειστικά συστάδες με μία μονή παρατήρηση (Hochreiter, 2014).



Σχήμα 3.2: *Δενδρόγραμμα παρατηρήσεων του δείγματος δεδομένων Iris (Anderson, 1935) και (Fisher, 1936).*

Στο αριστερό σχήμα 3.2 απεικονίζεται το δενδρόγραμμα, το οποίο έχει προκύψει από το δείγμα δεδομένων Iris με τη μέθοδο της μέσης ζεύξης. Αν κάποιος θα ήθελε να ερμηνεύσει το δενδρόγραμμα αυτό σύντομα δίχως να μπει σε βάθος, τότε θα επισήμανε αρχικά ότι καθώς παρατηρεί το σχήμα από κάτω προς τα πάνω διαπιστώνεται ότι τα φύλλα του δενδρογράμματος, τα οποία αναπαριστούν τις παρατηρήσεις του δείγματος, ενώνονται μεταξύ τους με κλαδιά. Όσο προχωράει προς τα πάνω, γίνεται επίσης εμφανές ότι τα κλαδιά ενώνονται και αυτά με τη σειρά τους με άλλα κλαδιά. Οι παραπάνω διαπιστώσεις οδηγούν στον ισχυρισμό ότι τα φύλλα, τα οποία συνδέονται το συντομότερο δυνατόν, εμφανίζουν αρκετές ομοιότητες, ενώ όσο μετέπειτα συνδέονται μεταξύ τους, τόσο μεγαλύτερο χάσμα τα χωρίζει. Ομοίως, το γεγονός της συγχώνευσης

των κλαδιών σε σύντομο χρονικό διάστημα συνηγορεί στο συμπέρασμα ότι οι συστάδες είναι πανομοιότυπες μεταξύ τους. Έτσι, με την χρήση του κάθετου άξονα καθίσταται ευδιάκριτη η απόσταση και άρα η διαφορά μεταξύ των παρατηρήσεων. Είναι, λοιπόν, αυτόδηλο ότι οι παρατηρήσεις οι οποίες ενώνονται στο τέλος του δενδρογράμματος, ήτοι στο κάτω μέρος του γραφήματος, είναι αρκετά όμοιες, στον αντίποδα με τις παρατηρήσεις οι οποίες ενώνονται στην αρχή του δενδρογράμματος, για τις οποίες υπάρχει η ένδειξη ότι αυτές διαφέρουν. Επιπρόσθετα, αναφορικά με την εύρεση του αριθμού των συστάδων με τη συνδρομή του δενδρογράμματος, η κεντρική ιδέα σχετίζεται με τη σχεδίαση μίας οριζόντιας γραμμής, η οποία θα χωρίζει το δενδρόγραμμα. Με αυτόν τον τρόπο, οι κάθετοι κορμοί του δέντρου, που βρίσκονται λίγο πιο κάτω από την οριζόντια γραμμή, αντιπροσωπεύουν την κάθε συστάδα και, έτσι, εξάγεται το συμπέρασμα του αριθμού των συστάδων ενός δείγματος. Κάποιος θα μπορούσε εύλογα να αναρωτηθεί πού πρέπει να σχεδιάσει αυτήν την εν λόγω γραμμή. Η αλήθεια είναι ότι δεν υπάρχει κάποιος περιορισμός ή κάποιος κανόνας, στον οποίο υπακούει η μέθοδος αυτή, και η σχεδίαση της γραμμής εναπόκειται στην κρίση του κάθε ενός, καθιστώντας την τεχνική αυτή πιο υποκειμενική χωρίς να υπάρχει μία αντικειμενική αλήθεια. Επομένως, θεωρητικά μπορεί η γραμμή να μη σχεδιαστεί καθόλου και, οπότε, να υπάρχει αποκλειστικά μία συστάδα για τα δεδομένα, ενώ όσο πιο κάτω στο δενδρόγραμμα σχεδιαστεί, τόσο περισσότερες σε αριθμό θα είναι οι συστάδες. Επίσης και στην περίπτωση που αυτή σχεδιαστεί στη βάση του δενδρογράμματος, ήτοι στον αριθμό μηδέν του κάθετου άξονα, τότε αυτό το γεγονός θα έχει ως συνέπεια τη δημιουργία n συστάδων, όπου n ο αριθμός των παρατηρήσεων του δείγματος. Πρακτικά, η διαλογή του σημείου κοπής γίνεται με το μάτι έτσι, ώστε να δημιουργείται ένας λογικός αριθμός από συστάδες και από διακλαδώσεις και συγχωνεύσεις (Hastie et al., 2021).

Λαμβάνοντας όλα τα παραπάνω υπ΄ όψιν και εφαρμόζοντάς τα στο δενδρόγραμμα του σχήματος 3.2, κρίνεται εύκολα αντιληπτό ότι, για παράδειγμα, οι παρατηρήσεις 1 και 2, όπως και οι 8 και 9, ίσως είναι αρκετά όμοιες μεταξύ τους. Ωστόσο, αυτή η ομοιότητα δε φαίνεται να ισχύει για τις παρατηρήσεις 5 και 2 ή για τις 10 και 8. Επιπλέον, ίσως λόγω κεκτημένης ταχύτητας και απροσεξίας, θα μπορούσε κάποιος να ισχυριστεί ότι η παρατήρηση 5 εμφανίζει κάποια ομοιότητα με την παρατήρηση 4, δεδομένου ότι αυτές βρίσκονται αρκετά κοντά μεταξύ τους. Αυτή η διαπίστωση, όμως, θεωρείται εσφαλμένη, μιας και η παρατήρηση 5 είναι τόσο όμοια με την παρατήρηση 4 όσο είναι και με τις παρατηρήσεις 3, 1 και 2. Για αυτόν το λόγο, θεωρείται σκόπιμο τα συμπεράσματα ως

προς την ομοιότητα δύο παρατηρήσεων να εξάγονται αποκλειστικά βάσει του κάθετου άξονα και όχι του οριζόντιου, προς αποφυγήν τέτοιων λαθών. Παράλληλα, όσον αφορά στον αριθμό των συστάδων, που δύναται να περιγράψουν τις παρατηρήσεις, φαίνεται ότι η σχεδίαση της γραμμής στην τιμή περίπου 28 του κάθετου άξονα να είναι μία κατάλληλη επιλογή δημιουργώντας, με αυτόν τον τρόπο δύο συστάδες για το δείγμα.

Καθοριστικό ρόλο στη διαμόρφωση του δενδρογράμματος και κατ΄ επέκταση στην ερμηνεία των αποτελεσμάτων ενός δείγματος δεδομένων παίζει και ο παράγοντας της κατάλληλης επιλογής ως προς τη μέτρηση διαφορετικότητας των ζευγών. Όπως έχει προλεχθεί, με την προϋπόθεση ότι το δείγμα περιέχει ποσοτικές μεταβλητές, η Ευκλείδεια, η Μανχάταν και η Mahalanobis απόσταση συνήθως προτιμώνται, προκειμένου να προσδιορίσει τη διαφορετικότητα των συστάδων και των παρατηρήσεων. Συγκεκριμένα, η Ευκλείδεια απόσταση μεταξύ δύο διανυσμάτων δίνεται από τον τύπο:

$$d(A,B) = (\sum_{i=1}^{p} (\alpha_i - \beta_i)^2)^{1/2}.$$

Επιπλέον, σε ό,τι αφορά την απόσταση Μανχάταν, αυτή υπολογίζεται εύκολα με τη βοήθεια της παρακάτω ισότητας:

$$d(A,B) = \sum_{i=1}^{p} |\alpha_i - \beta_i|.$$

Τέλος, η Mahalanobis απόσταση μπορεί να υπολογιστεί εύκολα από τη σχέση

$$d(A,B) = [(A-B)^T * S^{-1} * (A-B)]^{1/2},$$

όπου $A = [\alpha_i], B = [\beta_i], i = 1, .., p$ και S ο θετικά ορισμένος πίνακας συνδιασποράς των A και B.

Ωστόσο, υπάρχουν στη διάθεση ενός στατιστικού και άλλοι τρόποι μέτρησης της διαφορετικότητας, οι οποίοι θα μπορούσαν να χαρακτηριστούν αρκετά εύχρηστοι και ευπροσάρμοστοι για ένα δείγμα δεδομένων. Ένας τέτοιος τρόπος είναι το κριτήριο της απόστασης με βάση τη συσχέτιση, κατά το οποίο ορίζει ότι δύο παρατηρήσεις είναι όμοιες στην περίπτωση της υψηλής μεταξύ τους συσχέτισης, μη λογαριάζοντας τη μεταξύ τους Ευκλείδεια απόσταση. Επιπλέον, το κριτήριο αυτό εστιάζεται στο σχήμα των προφίλ των παρατηρήσεων παρά στη

σπουδαιότητά τους, καθώς και έχει ευρεία χρήση στους κόλπους του διαδικτυακού εμπορίου, εφόσον η κατηγοριοποίηση των καταναλωτών πραγματοποιείται βάσει της ομοιότητας του είδους των αγαθών, τα οποία έχουν αποκτηθεί ή για τα οποία ενδιαφέρονται εν γένει. Στην περίπτωση που είχε επιλεχθεί η Ευκλείδεια απόσταση για την εύρεση της διαφορετικότητας των καταναλωτών, τότε η ομαδοποίηση θα είχε προκύψει με γνώμονα την ποσότητα των αγαθών, που έχουν αγοραστεί από αυτούς. Επιπρόσθετα, στην περίπτωση της εφαρμογής της απόστασης με βάση τη συσχέτιση, θεωρείται απαραίτητη η τυποποίηση των παρατηρήσεων έτσι, ώστε η τυπική απόκλισή τους να ισούται με τη μονάδα, επειδή υπάρχει σε μεγάλο βαθμό το ρίσκο των διαφορετικών μονάδων μέτρησης των παρατηρήσεων (Hastie et al., 2021).

3.4 Ανάλυση Συστάδων Βασισμένη σε μοντέλο (Model-based clustering)

Στις προηγούμενες ενότητες έχει γίνει λόγος για την ιεραρχική καθώς και για την K- means ανάλυση συστάδων• μέθοδοι, οι οποίες θα μπορούσαν να χαρακτηρισθούν διερευνητικές και ταυτόχρονα διαισθητικές, μιας και προϋποθέτουν την εκτίμηση του αριθμού των συστάδων, που χαρακτηρίζει το δείγμα. Αυτό το γεγονός, καθώς και η «απουσία» ενός μοντέλου, το οποίο δύναται να περιγράψει τη δομή των συστάδων σε ένα δείγμα, καθιστά τη μελέτη ενός μεγάλου δείγματος αρκετά δύσκολη έως απαιτητική, με συνέπεια συχνά να μη γίνεται δυνατή η εξαγωγή στατιστικών συμπερασμάτων. Ως εκ τούτου, θεωρείται απαραίτητη και πολύ χρήσιμη η εύρεση των συστάδων δια μέσου ενός στατιστικού μοντέλου για έναν πληθυσμό, από τον οποίο έχουν εξαχθεί τα δεδομένα ενός δείγματος, με την υπόθεση ότι αυτός ο πληθυσμός εμπεριέχει υποπληθυσμούς (συστάδες) καθώς και ότι ακολουθεί διαφορετικές κατανομές πιθανοτήτων. Επιπλέον, κάθε υποπληθυσμός αντιπροσωπεύεται από μεταβλητές με διαφορετική πολυμεταβλητή συνάρτηση πυκνότητα πιθανότητας. Έτσι, με το παραπάνω σκεπτικό οδηγείται κάποιος στην χρήση της πυκνότητας πεπερασμένης ανάμειξης (finite mixture density), η οποία θα προσφέρει ένα κατάλληλο στατιστικό μοντέλο, προκειμένου να περιορισθεί ο αριθμός των παραμέτρων της εκτιμώμενης ανάμειξης και, κατά συνέπεια, να γίνει ο υπολογισμός της μεταγενέστερης πιθανότητας με τη συνεισφορά των συστάδων. Η παραπάνω διαδικασία της ανάλυσης συστάδων βάσει της πυκνότητας πεπερα-

σμένης ανάμειξης είναι γνωστή ως ανάλυση συστάδων βασισμένη σε μοντέλο (Model-based clustering). Επιπρόσθετα, η ανάλυση συστάδων βασισμένη σε μοντέλα, όπου οι υποπληθυσμοί περιγράφονται από διαφορετικές υποβόσκουσες κατηγορικές μεταβλητές και οι υποβοσκούμενες τάξεις περιγράφονται από διαφορετικές συνιστώσες πυκνότητας ανάμειξης, είναι γνωστή και ως ανάλυση συστάδων υποβόσκουσας τάξης (latent class cluster analysis) (Everitt και Hothorn, 2011).

Οι πυκνότητες πεπερασμένης ανάμειξης ορίζονται μαθηματικά ως μία οικογένεια συναρτήσεων πυκνότητας πιθανότητας με μορφή:

$$f(\mathbf{x}; \mathbf{p}, \mathbf{y}) = \sum_{i=1}^{c} p_i g_i(\mathbf{x}; y_i),$$

όπου $\mathbf{x}$ τυχαία μεταβλητή διάστασης p, $\mathbf{p}$ διάνυσμα στήλης διάστασης $c-1$ και $\mathbf{y}$ διάνυσμα στήλης διάστασης c. Τα στοιχεία p_i είναι γνωστά ως ανάμεικτες αναλογίες και τα στοιχεία g_i, τα οποία είναι παραμετρικοποιημένα ως προς τις μεταβλητές y_i είναι γνωστά ως συνιστώσες πυκνότητας. Έτσι, ο αριθμός των συνιστωσών συντελεί στο σχηματισμό της ανάμειξης. Οι ανάμεικτες αναλογίες δεν μπορούν να είναι αρνητικές, καθώς και υπακούουν στην ιδιότητα ότι το άθροισμά τους ισούται με τη μονάδα.

Για να υπολογιστούν οι παράμετροι της πυκνότητας πεπερασμένης ανάμειξης, γίνεται χρήση της μεθόδου μεγίστης πιθανοφάνειας. Αποδεικνύεται σε συμφωνία με τους Everitt και Hothorn (2011) ότι αυτή υπολογίζει τις παραμέτρους (την i συνιστώσα, το i μέσο διάνυσμα και τον πίνακα συνδιασποράς για την i συνιστώσα) ως:

$$\hat{p}_i = \frac{1}{n} \sum_{j=1}^{n} \hat{P}(i|x_j) \tag{3.4}$$

$$\hat{\mu}_i = \frac{1}{n\hat{p}_i} \sum_{j=1}^{n} x_j \hat{P}(i|x_j) \tag{3.5}$$

$$\hat{\Sigma}_i = \frac{1}{n} \sum_{j=1}^{n} (x_j - \mu_i)(x_j - \mu_i)^T \hat{P}(i|x_j) \tag{3.6}$$

Με αυτόν τον τρόπο, γίνεται αντιληπτό ότι για τον υπολογισμό των εξισώσεων 3.4, 3.5 και 3.6 χρειάζεται αρχικά να γίνει γνωστή η πιθανότητα $\hat{P}(i|x_j)$, γνωστή ως εκτιμώμενη τιμή μεταγενέστερης πιθανότητας.

Η εκτιμώμενη μεταγενέστερη πιθανότητα ορίζεται ως:

$$\hat{P}(i|x_j) = \frac{\hat{p}_i g_i(x_j; \hat{y}_i)}{f(x_j; \hat{\mathbf{p}}, \hat{\mathbf{y}})},$$

όπου i η κάθε συστάδα με $i = 1, 2, ..., c$.

Επιπλέον, για την εκτίμηση των 3.4, 3.5 και 3.6 έχουν γίνει αρκετές προτάσεις με τις Μπεϋζιανές μεθόδους εκτίμησης καθώς επίσης και με τον επαναλαμβανόμενο αλγόριθμο αναμενόμενης μεγιστοποίησης (EM algorithm) από τους Dempster, Laird, και Rubin (1977) να επικρατούν γενικότερα.

Έπειτα, έχοντας υπολογίσει όπως παραπάνω τις παραμέτρους, αντιστοιχίζονται οι παρατηρήσεις με τις συστάδες σύμφωνα με τη μέγιστη τιμή της μεταγενέστερης πιθανότητας.

Αναφορικά με την επιλογή του κατάλληλου μοντέλου, αυτό καθορίζεται από δύο παραμέτρους. Από τη μια πλευρά, πρέπει να ληφθεί υπόψη ποιό στατιστικό μοντέλο δύναται να περιγράψει ορθά τον πληθυσμό και τα δεδομένα. Από την άλλη, θεωρείται απαραίτητο να επιβεβαιωθεί αν αυτό το μοντέλο προσφέρει το βέλτιστο αριθμό συστάδων. Η Μπεϋζιανή προσέγγιση θέτει σε εφαρμογή το κριτήριο BIC για την επιλογή του μοντέλου (Everitt και Hothorn, 2011).

3.5 Υποβόσκοντες Κίνδυνοι στις Μεθόδους Ανάλυσης Συστάδων

Είναι αυταπόδεικτο γεγονός ότι η Ανάλυση Συστάδων συνιστά πολύτιμο και ισχυρό εργαλείο στο οπλοστάσιο της μη επιβλεπόμενης ανάλυσης (μεγάλων) δεδομένων. Όμως, πρέπει να εφιστάται η προσοχή σε ορισμένους τομείς κατά την πραγματοποίηση κάποιας μεθόδου ανάλυσης συστάδων.

Αρχικά, οι μέθοδοι αυτές εμπεριέχουν αρκετά την υποκειμενικότητα αυτού, ο οποίος αναλύει το δείγμα, με συνέπεια το αποτέλεσμα της ανάλυσης να είναι άρρηκτα συνδεδεμένο και εξαρτημένο από τις επιλογές και τις αποφάσεις του ατόμου αυτού. Έτσι, ούτε αποκλείουν κάποιο συμπέρασμα ως σίγουρα λανθασμένο ούτε υπαγορεύουν μία κοινή και καθολική αντικειμενική αλήθεια, αφού δεν αποσαφηνίζουν κάποια βασικά ερωτήματα, όπως αν πρέπει τα δεδομένα να τυποποιούνται έτσι, ώστε να έχουν τυπική απόκλιση ίση με τη μονάδα και πού πρέπει να κοπεί το δενδρόγραμμα, προκειμένου τα συμπεράσματα της ανάλυσης να είναι εύλογα, καθώς επίσης και δεν καθορίζει έναν τρόπο ζεύξης των παρατηρήσεων και της μέτρησης της διαφορετικότητας των ζευγών στην περίπτωση της ιεραρχικής ανάλυσης συστάδων. Παράλληλα, κατά την χρήση της μεθόδου K-means ένας άλλος λανθάνων κίνδυνος σφάλματος και παραπλάνησης από την πραγματικότητα οφείλεται στο γεγονός του μη προκαθορισμού του εκ των προτέρων αριθμού συστάδων, που απαιτούνται για το δείγμα.

Επιπρόσθετα, ένα άλλο γεγονός, που εγείρει προβληματισμούς, είναι ότι τόσο η μέθοδος K-means όσο και η ιεραρχική μέθοδος υποχρεώνει κάθε παρατήρηση να ανήκει σε μία συστάδα, αμελώντας την περίπτωση, όπου το δείγμα περιέχει μία μικρή υποομάδα παρατηρήσεων, οι οποίες διαφέρουν και μεταξύ τους και με όλες τις υπόλοιπες του δείγματος (έκτροπα σημεία). Έτσι, απόηχος αυτής της υποχρέωσης είναι η παραποίηση των συστάδων και κατ΄ επέκταση των στατιστικών συμπερασμάτων.

Τέλος, ένα μειονέκτημα των εν λόγω μεθόδων είναι ότι τα αποτελέσματα δεν παραμένουν αναλλοίωτα σε μεταβολές, τις οποίες θα μπορούσαν να υφίστανται τα αρχικά δεδομένα. Με άλλα λόγια, μία αφαίρεση μικρής υποομάδας των δεδομένων εγκυμονεί τον κίνδυνο μίας από αρκετά μεγάλης έως τεράστιας διαφοράς στα τελικά αποτελέσματα της ανάλυσης.

Με αυτόν τον τρόπο, εκτιμώντας όλους τους υποβόσκοντες κινδύνους, κρίνεται αναγκαίο οι μέθοδοι ανάλυσης συστάδων να εκτελούνται πολλαπλές φορές για διαφορετικές παραμέτρους συγκρίνοντας τις μεταξύ τους διαφορές, καθώς και να μη θεωρούνται τα συμπεράσματα ως αντικειμενική και μοναδική αλήθεια με πλήρη ορθότητα (Hastie et al., 2021).

Κεφάλαιο 4

Εφαρμογές

4.1 Δεδομένα USArrests

Το δείγμα δεδομένων USArrests (McNeil, 1977) αφορά σε αριθμούς συλλήψεων, οι οποίες έχουν λάβει χώρα στις 50 πολιτείες των Ηνωμένων Πολιτειών της Αμερικής. Οι αριθμοί των συλλήψεων είναι καταγεγραμμένοι ανά 100.000 κατοίκους των πολιτειών και οι συλλήψεις αυτές διακρίνονται σε Βιαιοπραγία (Assault), Φόνο (Murder) και Βιασμό (Rape). Επιπλέον, στο δείγμα είναι καταγεγραμμένος και ο πληθυσμός της εκάστοτε πολιτείας (UrbanPop).
Με τη βοήθεια των στατιστικών πακέτων της R και των παρακάτω εντολών, δίνεται η δυνατότητα να παρατηρήσουμε και να μελετήσουμε αυτά τα δεδομένα.

> summary(USArrests)

	Murder	**Assault**	**UrbanPop**	**Rape**
Min.	0.800	45.0	32.00	7.30
1st Qu.	4.075	109.0	54.50	15.07
Median	7.250	159.0	66.00	20.10
Mean	7.788	170.8	65.54	21.23
3rd Qu.	11.250	249.0	77.75	26.18
Max.	17.400	337.0	91.00	46.00

Πίνακας 4.1: *Σύνοψη δεδομένων USArrests*

Από τον παραπάνω πίνακα 4.1 παρατηρείται η δειγματική μέση τιμή και ο δειγματικός μέσος των παρατηρήσεων για κάθε κατηγορία, ήτοι Murder, Assault, UrbanPop και Rape. Παρατηρείται ότι η δειγματική μέση τιμή διαφέρει αρκετά σε κάθε κατηγορία.

> *apply(USArrests, 2, var)*

Murder Assault UrbanPop Rape
18.97047 6945.16571 209.51878 87.72916

Σημειώνεται ότι οι δειγματικές διασπορές κάθε κατηγορίας διαφέρουν αρκετά. Ωστόσο, δεν μπορεί να γίνει σύγκριση μεταξύ των μεταβλητών, δεδομένου ότι διαφέρουν οι μονάδες μέτρησής τους. Για αυτόν το λόγο, απαιτείται η κατάταξη των δεδομένων σε κλίμακα, όπως έχει αναφερθεί στο κεφάλαιο 2.2.

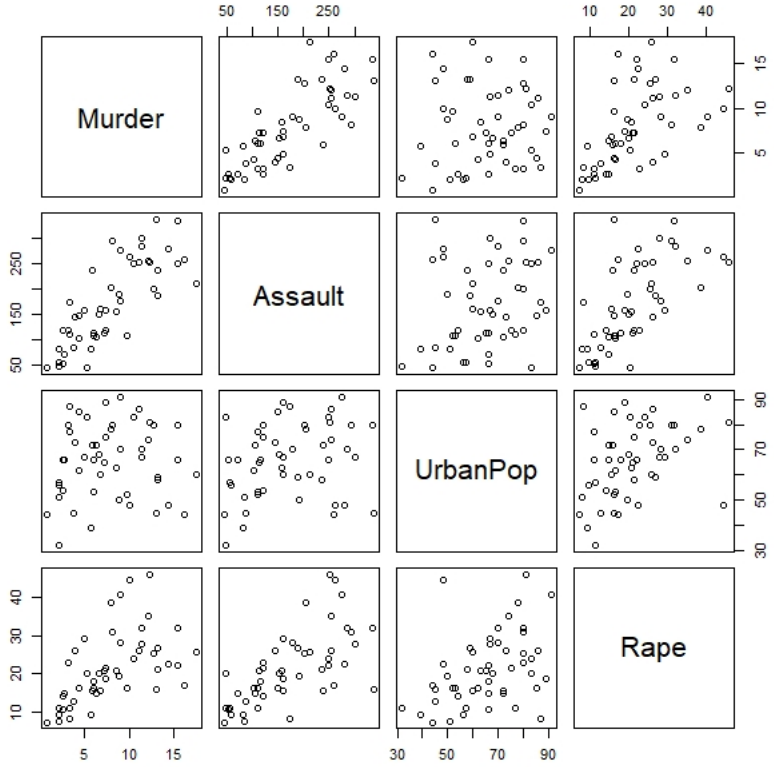
Έπειτα, ελέγχονται οι συσχετίσεις των μεταβλητών μεταξύ τους.
> *round(cor(USArrests), 2)*

	Murder	**Assault**	**UrbanPop**	**Rape**
Murder	1.00	0.80	0.07	0.56
Assault	0.80	1.00	0.26	0.67
UrbanPop	0.07	0.26	1.00	0.41
Rape	0.56	0.67	0.41	1.00

Πίνακας 4.2: *Πίνακας συσχετίσεων μεταβλητών του δείγματος USArrests*

> *plot(USArrests)*
Το γράφημα εμφανίζεται στην επόμενη σελίδα.

Παρατηρείται ότι με τη βοήθεια του πίνακα 4.2 οι μεταβλητές Murder, Assault και Rape είναι υψηλά συσχετισμένες. Η μεταβλητή UrbanPop είναι συσχετισμένη σε πολύ χαμηλότερο επίπεδο με τις υπόλοιπες μεταβλητές με εξαίρεση τη μεταβλητή Murder με την οποία εμφανίζεται εντελώς ασυσχέτιστη. Στο ίδιο συμπέρασμα οδηγείται κάποιος και από το διάγραμμα 4.1, όπου οι παρατηρήσεις των τριών συσχετισμένων μεταβλητών δύνανται να σχηματίσουν μία νοητή ευ-



Σχήμα 4.1: *Διάγραμμα συσχετίσεων στο δείγμα USArrests*

θεία γραμμή. Στην περίπτωση των παρατηρήσεων για τη μεταβλητή UrbanPop αυτή η νοητή ευθεία σχηματίζεται με αρκετή δυσχέρεια.

4.1.1 Εφαρμογές Ανάλυσης Κύριων Συνιστωσών στο δείγμα δεδομένων USArrests

Μετά την αναγκαία κατάταξη των δεδομένων σε κλίμακα καθιστώντας κάθε μεταβλητή να έχει μέση τιμή 0 και τυπική απόκλιση 1, ακολουθεί η εύρεση των κύριων συνιστωσών και των φορτίσεών τους για το δείγμα USArrests.

> pca.res <- prcomp(USArrests, scale = TRUE)

> *pca.res$rotation*

	PC1	**PC2**	**PC3**	**PC4**
Murder	-0.5358995	0.4181809	-0.3412327	0.64922780
Assault	-0.5831836	0.1879856	-0.2681484	-0.74340748
UrbanPop	-0.2781909	-0.8728062	-0.3780158	0.13387773
Rape	-0.5434321	-0.1673186	0.8177779	0.08902432

Πίνακας 4.3: *Φορτίσεις κύριων συνιστωσών του δείγματος USArrests*

Με την παρατήρηση των τιμών των κύριων φορτίσεων διαπιστώνεται ότι μερικές έχουν αρνητικό πρόσημο. Γενικά, όπως έχει προαναφερθεί στο κεφάλαιο 2.6, τα πρόσημα των φορτίσεων δεν παίζουν κάποιο σημαντικό ρόλο, καθώς η εναλλαγή τους δεν επιφέρει κάποια διαφορά στην κατεύθυνση στον χώρο, όπου υπάρχει μέγιστη μεταβλητότητα. Ωστόσο, αυτό που πρέπει να ληφθεί υπ΄ όψιν είναι η διαφορά των προσήμων των φορτίσεων για κάθε συνιστώσα.

> *pca.var =pca.res$sdev ^ 2*
> *pca.var*

2.4802416 0.9897652 0.3565632 0.1734301

Με την παραπάνω εντολή διαπιστώνεται η διασπορά, που έχει ως τιμή κάθε κύρια συνιστώσα. Συγκεκριμένα, η πρώτη κύρια συνιστώσα έχει διασπορά περίπου 2.5, η δεύτερη έχει πολύ μικρότερη με τιμή σχεδόν 1, η τρίτη έχει ακόμα μικρότερη περίπου 0.4 και τέλος η τέταρτη έχει την ελάχιστη και φτάνει στην τιμή 0.2 κατά προσέγγιση. Συμπερασματικά, με μία πρώτη εκτίμηση φαίνεται ότι οι δύο πρώτες κύριες συνιστώσες εκφράζουν ένα αρκετά υψηλό βαθμό διασποράς και μεταβλητότητας στο δείγμα. Αυτό το γεγονός επικυρώνεται εύκολα βρίσκοντας την αναλογία διασποράς κάθε συνιστώσας στο συγκεκριμένο δείγμα, τη γραφική αναπαράσταση της αναλογίας καθώς επίσης τη συσσωρευμένη αναλογία της διασποράς από κάθε συνιστώσα:

> *var.ratio=pca.var/sum(pca.var)*
> *var.ratio*
0.62006039 0.24744129 0.08914080 0.04335752

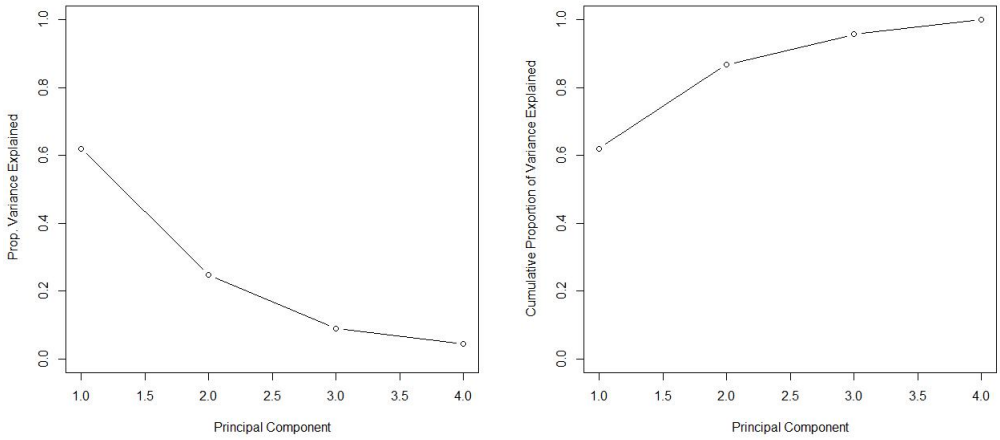
Μετατρέποντας προσεγγιστικά τις τιμές σε ποσοστά η διασπορά της πρώτης κύριας συνιστώσας εκφράζει το 62.0% του δείγματος, της δεύτερης το 24.7%, της τρίτης μόνο το 8.9% και τέλος της τέταρτης μόνο το 4.3%.

> plot(var.ratio , xlab=" Principal Component ", ylab=" Prop. Variance Explained" , ylim=c(0,1) ,type="b")
> plot(cumsum (var.ratio), xlab=" Principal Component ", ylab =" Cumulative Proportion of Variance Explained ", ylim=c(0,1) ,type="b")

Οι γραφικές παραστάσεις απεικονίζονται στο σχήμα 4.2. Στα αριστερά βρίσκεται η γραφική παράσταση αναλογίας των συνιστωσών, ενώ δεξιά διακρίνεται η γραφική παράσταση της συσσωρευμένης αναλογίας τους. Ειδικότερα, στο αριστερό γράφημα επαληθεύεται ότι η πρώτη κύρια συνιστώσα εκφράζει το μεγαλύτερο μέρος της διασποράς και η δεύτερη κύρια συνιστώσα εκφράζει μικρότερο ποσοστό από την πρώτη αλλά ευρύτερα ένα ικανοποιητικό ποσοστό σε αντίθεση με τις άλλες δύο κύριες συνιστώσες. Στο δεξιό γράφημα γίνεται κατανοητό ότι οι πρώτες δύο συνιστώσες μαζί συντελούν στη δέουσα περιγραφή των δεδομένων του αρχικού προβλήματος έχοντας συσσωρευμένη διασπορά πάνω από 0.8. Ωστόσο, παρατηρείται ότι οι άλλες δύο συνιστώσες συμβάλλουν σε ένα πολύ μικρότερο βαθμό στη περιγραφή του προβλήματος.

Αδιαφιλονίκητα, λοιπόν, καταλήγει κάποιος στο συμπέρασμα ότι οι δύο πρώτες κύριες συνιστώσες επαρκούν για την περιγραφή του δείγματος, μιας και εκφράζουν περίπου το 87% της διασποράς του αρχικού δείγματος. Πρόκειται για ένα αρκετά υψηλό ποσοστό μεγαλύτερο του 70% και αρκετά κοντά στο 90%, σε αντιδιαστολή με τις άλλες δύο, οι οποίες εκφράζουν ένα πολύ χαμηλό ποσοστό της τάξης του 13 με 14%. Παράλληλα, σε συμφωνία με τη γραφική παράσταση της αναλογίας διασπορών είναι εμφανές ότι στο δεύτερο σημείο (δεύτερη κύρια συνιστώσα) σημειώνεται ένας βραχίονας, αλλάζει δηλαδή απότομα η κλίση της ευθείας. Συνυπολογίζοντας το γεγονός ότι η τρίτη κύρια συνιστώσα έχει διασπορά 0.4 και η τέταρτη 0.2, παρατηρείται ότι αυτές οι τιμές είναι μικρότερες από τη μέση διασπορά του δείγματος, η οποία ισούται με 1. Όλοι οι παραπάνω ισχυρισμοί, επιρρωνύουν τη θέση ότι χρειαζόμαστε μόνο τις δύο πρώτες κύριες συνιστώσες για την περιγραφή, την επεξήγηση και την οπτικοποίηση του αρχικού δείγματος χάνοντας αμελητέα στατιστική πληροφορία.

Έτσι, έχοντας επιλέξει δύο κύριες συνιστώσες, αφαιρούνται οι δύο



Σχήμα 4.2: *Διάγραμματα αναλογιών - συσσωρευμένων αναλογιών διασποράς κύριων συνιστωσών στο δείγμα USArrests*

τελευταίες, και ακολουθεί η οπτικοποίηση των παρατηρήσεων του δείγματος USArrests με τη βοήθεια του γραφήματος biplot. Οι εντολές, που εκτελούνται στην R είναι οι ακόλουθες.

> pca.res <- prcomp(USArrests, scale = TRUE, rank =2)
> pca.res$rotation

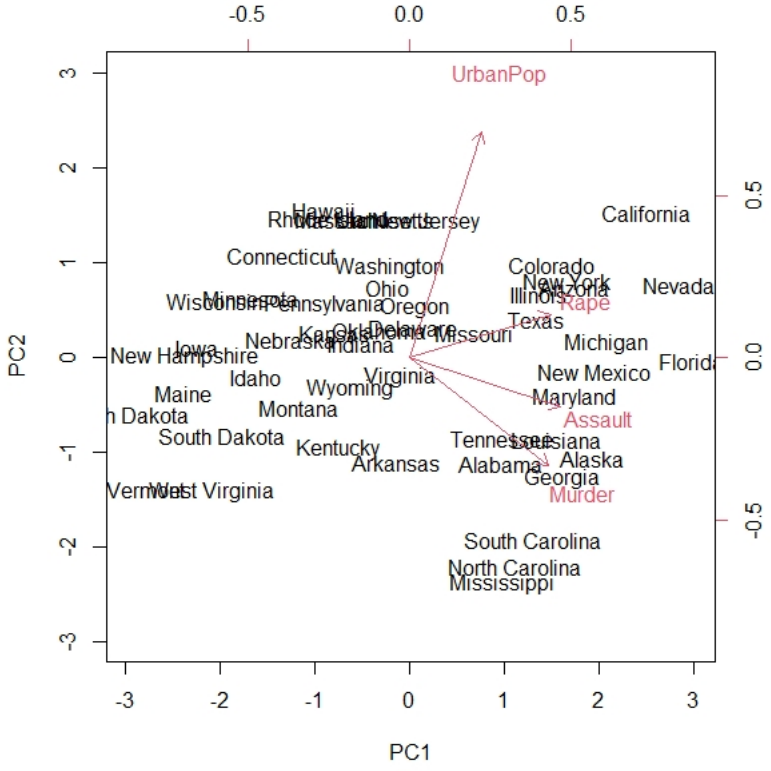
	PC1	**PC2**
Murder	-0.5358995	0.4181809
Assault	-0.5831836	0.1879856
UrbanPop	-0.2781909	-0.8728062
Rape	-0.5434321	-0.1673186

Πίνακας 4.4: *Δύο πρώτες κύριες συνιστώσες του δείγματος USArrests*

> pca.res$rotation=-pca.res$rotation
> pca.res$x=-pca.res$x
> biplot (pca.res , scale =0)

Η γραφική παράσταση εμφανίζεται στο σχήμα 4.3.

Αν κάποιος θα ήθελε να σχολιάσει τα αποτελέσματα του πίνακα 4.4, τότε θα επεσήμανε ότι οι φορτίσεις της πρώτης κύριας συνιστώσας είναι παραπλήσιες στις μεταβλητές- κατηγορίες Murder, Assault και Rape με εξαίρεση τη μεταβλητή UrbanPop, η οποία αντιστοιχεί σε φόρτιση με αλγεβρική τιμή μικρότερη περίπου κατά το μισό σε σχέση με τις άλλες φορτίσεις. Αντίθετα, η φόρτιση της δεύτερης κύριας συνιστώσας, η οποία αντιστοιχεί στη μεταβλητή UrbanPop είναι η μεγαλύτερη κατά απόλυτη τιμή, ενώ οι φορτίσεις της δεύτερης κύριας συνιστώσας, που αντιστοιχούν στις άλλες κατηγορίες είναι πολύ μικρότερες.Από τους παραπάνω ισχυρισμούς θα μπορούσε να εξαχθεί το συμπέρασμα, ότι η πρώτη κύρια συνιστώσα «δίνει το βάρος της» στις 3 μεταβλητές Murder, Assault και Rape ισομερισμένα, ενώ η δεύτερη κύρια συνιστώσα στην τέταρτη μεταβλητή, ήτοι UrbanPop.



Σχήμα 4.3: *Διάγραμμα Biplot στο δείγμα USArrests με δύο κύριες συνιστώσες.*

Επεξεργαζόμενοι τώρα το σχήμα 4.3, αρχικά, παρατηρείται ότι οι γωνίες με-

ταξύ των διανυσμάτων Murder, Assault και Rape είναι περίπου ίσες και μικρές. Αξίζει να σημειωθεί ότι η συσχέτιση των εγκλημάτων Murder και Assault είναι μεγαλύτερη, εφόσον η γωνία μεταξύ τους είναι μικρή. Αυτό το γεγονός υποδηλώνει ότι υπάρχει περίπου ίση και υψηλή συσχέτιση μεταξύ των μεταβλητών των εγκλημάτων στο δείγμα. Ωστόσο, η γωνία κάθε ενός από τα διανύσματα αυτά με το διάνυσμα UrbanPop είναι αρκετά μεγάλη καθιστώντας εμφανή τη χαμηλή συσχέτιση των μεταβλητών αυτών μεταξύ τους. Επιπρόσθετα, η γωνία των διανυσμάτων Murder, Assault και Rape με τον άξονα PC1 είναι μικρή σε αντίθεση με τη γωνία, που σχηματίζουν με τον άλλον άξονα PC2. Ως εκ τούτου, η συσχέτιση, που έχουν οι μεταβλητές αυτές με την πρώτη κύρια συνιστώσα, είναι μεγάλη σε αντιδιαστολή με αυτήν, που έχουν με τη δεύτερη κύρια συνιστώσα. Ακριβώς το αντίθετο ισχύει για τη μεταβλητή UrbanPop, η οποία είναι υψηλά συσχετισμένη μόνο με τη δεύτερη κύρια συνιστώσα. Αυτά τα πορίσματα, άλλωστε, επιβεβαιώνουν τα σχόλια του παραπάνω πίνακα 4.4.

Παράλληλα, η γραφική αυτή παράσταση οδηγεί και σε άλλα εξίσου σημαντικά συμπεράσματα. Εύκολα διαπιστώνεται ότι οι τιμές της πρώτης κύριας συνιστώσας για τις πολιτείες California, Nevada και Florida είναι οι πιο μεγάλες, εν αντιθέσει με τις τιμές της πρώτης κύριας συνιστώσας για τις πολιτείες North Dakota, Vermont και New Hampshire, οι οποίες είναι οι πιο χαμηλές. Με γνώμονα αυτόν τον ισχυρισμό, γίνεται αντιληπτό ότι οι τιμές στις μεταβλητές Murder, Assault και Rape είναι οι πιο υψηλές στις California, Nevada και Florida, ενώ οι μεταβλητές αυτές είναι οι πιο μικρές στις North Dakota, Vermont και New Hampshire. Συνεπώς, γίνεται κατανοητό οτι οι τρεις πρώτες πολιτείες έχουν εν γένει υψηλή εγκληματικότητα, ενώ οι άλλες τρεις έχουν την πιο χαμηλή εγκληματικότητα στις Η.Π.Α.
Με αντίστοιχη λογική, οι πολιτείες California και Hawaii είναι οι πιο κατοικημένες από πληθυσμό, δεδομένου ότι η τιμή της δεύτερης κύριας συνιστώσας,η οποία σχετίζεται με τη μεταβλητή UrbanPop, είναι η πιο μεγάλη στο δείγμα. Στον αντίποδα, οι πολιτείες Mississippi και North Carolina έχουν τους λιγότερους κατοίκους στο δείγμα.

Οι πολιτείες (όπως Virginia, Indiana κ.λ.π.), οι οποίες βρίσκονται κοντά στην τιμή του μηδενός και από τους δύο άξονες, κατέχουν τη μέση τιμή των μεταβλητών του δείγματος Murder, Assault , Rape και UrbanPop και κατ΄ επέκταση γίνεται αντιληπτό ότι αυτές οι πολιτείες έχουν μέσα επίπεδα όσον αφορά στην εγκληματικότητα και στον αριθμό του πληθυσμού τους.

Όπως έχει προαναφερθεί στην αρχή της ανάλυσης των δεδομένων USArrests, οι μεταβλητές έχουν τεθεί υπό κλίμακα με τέτοιον τρόπο, ώστε η δειγματική μέση τιμή κάθε μεταβλητής να ισούται με μηδέν και η τυπική απόκλιση κάθε μίας να ισούται με τη μονάδα. Οι εντολές, που έχουν χρησιμοποιηθεί πιο πάνω στη R, συμπεριλαμβάνουν αυτόν τον περιορισμό. Θεωρείται αρκετά ενδιαφέρον να φανερωθεί ποιά διαφορά θα υπάρξει, αν ακολουθηθεί η ίδια διαδικασία παραλείποντας την κλιμάκωση των μεταβλητών και τί συμπεράσματα επάγονται. Αν ήταν επιθυμητή μία a priori εκτίμηση για τις συσχετίσεις των μεταβλητών με τις κύριες συνιστώσες θα ήταν ότι η μεταβλητή Assault θα έχει αρκετά υψηλή τιμή φόρτισης πρώτης κύριας συνιστώσας, αφού ο πίνακας 4.2 θυμίζει ότι αυτή η μεταβλητή έχει την πιο υψηλή διασπορά, και άρα η πρώτη κύρια συνιστώσα θα δίνει έμφαση σε αυτή τη μεταβλητή, ενώ η δεύτερη θα δίνει βάρος στη μεταβλητή UrbanPop με τη δεύτερη μεγαλύτερη διασπορά. Εύκολα ελέγχονται αυτές οι εικασίες με τη βοήθεια της R.

> pca.res <- prcomp(USArrests, scale = FALSE, rank =2)
> pca.res$rotation

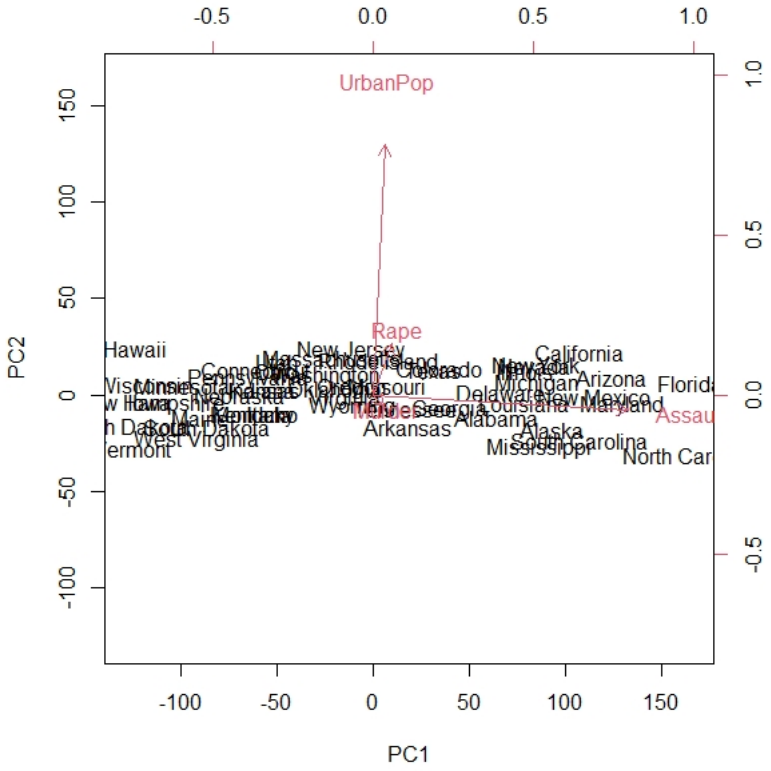
	PC1	**PC2**
Murder	0.04170432	-0.04482166
Assault	0.99522128	-0.05876003
UrbanPop	0.04633575	0.97685748
Rape	0.07515550	0.20071807

Πίνακας 4.5: *Δύο πρώτες κύριες συνιστώσες του δείγματος USArrests, του οποίου οι μεταβλητές δεν έχουν τεθεί υπό κλίμακα*

Παρατηρείται ότι οι τιμές του πίνακα 4.5 αποκλίνουν σε μεγάλο βαθμό από αυτές του πίνακα 4.4, όπου οι μεταβλητές είχαν τεθεί υπό κλίμακα.

> biplot (pca.res , scale =0)
Το αποτέλεσμα της εντολής φαίνεται στην επόμενη σελίδα.

Με μία πρώτη ματιά είναι πασιφανές ότι τα γραφήματα 4.3 και 4.4 αποκλίνουν με ειδοποιό διαφορά ότι η πρώτη κύρια συνιστώσα εστιάζει αποκλειστικά στη μεταβλητή Assault, ενώ η δεύτερη επικεντρώνεται κυρίως στη UrbanPop και



Σχήμα 4.4: *Διάγραμμα Biplot στο δείγμα USArrests,το οποίο δεν έχει τεθεί υπό κλίμακα .*

λιγότερο στη Rape. Η συσχέτιση μεταξύ των μεταβλητών της δεύτερης κύριας συνιστώσας είναι μεγάλη. Επομένως, διαπιστώνεται ότι η υψηλή διασπορά της μεταβλητής Assault επηρεάζει σε πολύ μεγάλο βαθμό την κατασκευή των κύριων συνιστωσών καθώς και την εξαγωγή στατιστικών συμπερασμάτων, η οποία κατά πάσα πιθανότητα θα είναι εσφαλμένη και σίγουρα θα διαφέρει από αυτήν, που θα έχει γίνει από το δείγμα, του οποίου οι μεταβλητές έχουν τεθεί υπό κλίμακα. Για αυτόν τον λόγο, προτείνεται ως πρώτο στάδιο σε προβλήματα ανάλυσης κύριων συνιστωσών η κλιμάκωση των μεταβλητών.

4.1.2 Χρήση Ανάλυσης Κύριων Συνιστωσών κατά τη συμπλήρωση δεδομένων

Το δείγμα δεδομένων USArrests έχει πλήρως καταγεγραμμένα και συμπληρωμένα όλα τα στοιχεία του. Ωστόσο, στην περίπτωση, όπου κάποια παρατήρηση στο δείγμα αυτό είχε παραλειφθεί να καταγραφεί είτε λόγω αμέλειας είτε λόγω μη δυνατότητας παρατήρησης της συγκεκριμένης μεταβλητής, τότε με τη βοήθεια της ανάλυσης κύριων συνιστωσών θα μπορούσε να προβλεφθεί αυτή η παρατήρηση και να συμπληρωθεί πλήρως ο πίνακα δεδομένων USArrests, όπως έχει αναφερθεί στο κεφάλαιο 2.8.

Αν μελετηθεί μία εφαρμογή στο δείγμα USArrests με ελλιπείς πληροφορίες με χρήση της R, τότε χρειάζεται προφανώς να διαγραφούν κάποιες παρατηρήσεις στο αρχικό δείγμα USArrests. Στο συγκεκριμένο παράδειγμα αφαιρούνται οι παρατηρήσεις της μεταβλητής Murder για τις πολιτείες Colorado και Illinois , της μεταβλητής UrbanPop για την πολιτεία Arizona , καθώς επίσης και της μεταβλητής Rape για την πολιτεία Montana .

> *install.packages("missMDA")*
> *USArrests[3,3]<-NA*
> *USArrests[13,1]<-NA*
> *USArrests[26,4]<-NA*
> *USArrests[6,1]<-NA*

Έπειτα, υπολογίζεται ο αναγκαίος αριθμός συνιστωσών για το καινούργιο δείγμα.

> *nopc<- estim_ncpPCA(USArrests,method.cv = "Kfold", verbose = FALSE)*

> *nopc$ncp*
2

Οπότε, γίνεται αντιληπτό ότι χρειάζονται 2 κύριες συνιστώσες για τη συμπλήρωση των τεσσάρων ελλιπών δεδομένων. Σε αυτό το στάδιο με βάση τον αριθμό των συνιστωσών, υπολογίζονται όλα τα ελλιπή δεδομένα ως εξής:

> *res.comp <- imputePCA(USArrests, ncp = nopc$ncp)*
> *res.comp$completeObs[3,3]*
75.63599
> *res.comp$completeObs[13,1]*
9.408646

> res.comp$completeObs[26,4]
14.87607
> res.comp$completeObs[6,1]
10.81897
Έτσι, με τις παραπάνω εντολές υπολογίστηκαν προσεγγιστικά όλες οι παρατηρήσεις, οι οποίες είχαν αφαιρεθεί χειροκίνητα και εκούσια. Αν θα ήθελε κάποιος να συγκρίνει τις τιμές αυτές με τις αντίστοιχες αρχικές τους πριν αφαιρεθούν, τότε θα παρατηρούσε ότι η απόκλιση των τιμών αυτών ίσως είναι αμελητέα. Συγκεκριμένα, η μεταβλητή Murder της πολιτείας Colorado ισούται με 7.9, ενώ στην παραπάνω διαδικασία υπολογίστηκε 10.8, η μεταβλητή Murder της πολιτείας Illinois, η οποία ισούται με 10.4, έχει πλέον πάρει την τιμή 9.4, η μεταβλητή UrbanPop της πολιτείας Arizona με τιμή 80 τώρα είναι ίση με 75.6. Πρόκειται, λοιπόν, για προσσεγγιστικές τιμές αρκετά παραπλήσιες σε αυτές του αρχικού δείγματος USArrests και, ως εκ τούτου, γίνεται πασιφανής η αξία και η συνεισφορά της μεθόδου της ανάλυσης κυρίων συνιστωσών στην εκτίμηση τιμών και παρατηρήσεων σε ελλιπή δεδομένα.

4.1.3 Εφαρμογές Ανάλυσης Συστάδων στο δείγμα δεδομένων USArrests

Μέθοδος K-Means

Όπως έχει προλεχθεί στο προηγούμενο κεφάλαιο (Κεφάλαιο 3), η μέθοδος K-Means προϋποθέτει την a priori εκτίμηση του αριθμού των συστάδων, που δύνανται να περιγράψουν κατάλληλα το δείγμα. Για αυτόν το λόγο, η εντολή kmeans() της R απαιτεί ως input τον αριθμό αυτό. Επίσης, ο αλγόριθμος δε θα τρέξει μία μόνο φορά, αλλά nstart φορές, έχοντας οριστεί στην εντολή, ξεκινώντας κάθε φορά από μία τυχαία διαμέριση των σημείων σε συστάδες. Στο τέλος της διαδικασίας, δίνει το καλύτερο αποτέλεσμα των nstart αναλύσεων.

Έτσι, για το δείγμα USArrests, κρίνεται αναγκαία η εύρεση του βέλτιστου αριθμού των συστάδων, που δύνανται να περιγράψουν καταλληλότερα το δείγμα. Ως εκ τούτου, πρέπει τα δεδομένα USArrests να τυποποιηθούν έτσι, ώστε κάθε μεταβλητή να έχει μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση, και έπειτα είναι απαραίτητο να βρεθεί ο αριθμός συστάδων για το δείγμα αυτό με τη βοήθεια της εντολής fviz_nbclust() κάνοντας χρήση του πακέτου *factoextra* .

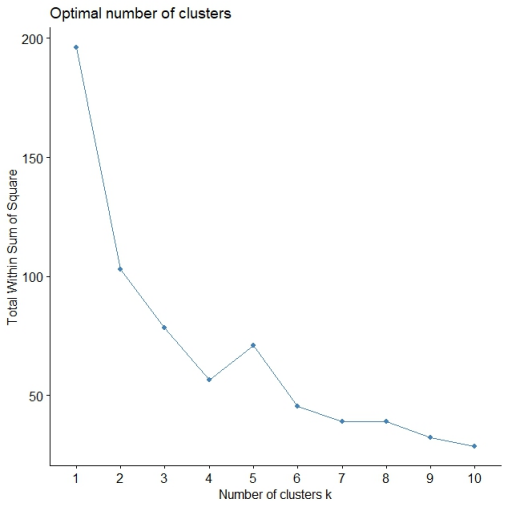
Με αυτόν τον τρόπο, οι εντολές έχουν ως εξής:

> library(factoextra)
> library(cluster)

> USArrestsScaling <- USArrests
> USArrestsScaling <- na.omit(USArrestsScaling)
> USArrestsScaling <- scale(USArrestsScaling)

Έτσι, με τις παραπάνω εντολές τα δεδομένα USArrests τέθηκαν υπό κλίμακα έχοντας παραλήψει πρώτα τις γραμμές του πίνακα δεδομένων USArrests με μη καταγεγραμμένες παρατηρήσεις.

> fviz_nbclust(USArrestsScaling, kmeans, method = "wss")



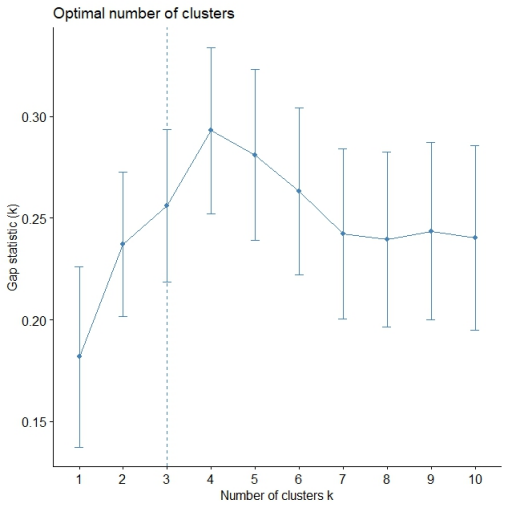
Σχήμα 4.5: *Διάγραμμα του αριθμού συστάδων έναντι του συνολικού αθροίσματος τετραγώνων για τα δεδομένα USArrests.*

Εξετάζοντας το σχήμα 4.5, παρατηρείται ότι ο βέλτιστος αριθμός συστάδων για το δείγμα θα μπορούσε να είναι 4, δεδομένου ότι στο γράφημα εμφανίζεται σε αυτό το σημείο ένας βραχίονας. Ωστόσο, για την επίρρωση του παραπάνω ισχυρισμού καθώς και για την εξάλλειψη οποιασδήποτε αμφιβολίας σχετικά με την επιλογή των 4 συστάδων για τα δεδομένα αυτά, γίνεται

χρήση της εντολής clusGap() , η οποία υποστηρίζεται από το πακέτο δεδομένων cluster. Έτσι, δια μέσου της εν λόγω εντολής, υπολογίζεται σχετικά εύκολα το gap statistic για κάθε αριθμό συστάδας, γεγονός που συνδράμει στην επιλογή του αριθμού συστάδων για την περιγραφή του δείγματος, σε συνδυασμό με το γράφημα των αριθμών συστάδων έναντι του gap statistic μέσω της εντολής fviz_gap_stat() .

> *gap_statistic <- clusGap(USArrestsScaling, FUN = kmeans, nstart = 25, K.max = 10, B=50)*
Clustering k = 1,2,..., K.max (= 10): .. done Bootstrapping, b = 1,2,..., B (= 50) [one "." per sample]: 50

> *fviz_gap_stat(gap_statistic)*
Το output της εντολής εμφανίζεται στο σχήμα 4.6.



Σχήμα 4.6: *Διάγραμμα του αριθμού συστάδων έναντι του gap statistic για τα δεδομένα USArrests.*

Γίνεται ξεκάθαρο, λοιπόν, ότι οι συστάδες, οι οποίες περιγράφουν καταλληλότερα τα δεδομένα, είναι 4, αφού και σύμφωνα με το γράφημα 4.6 παρατηρείται ότι το σημείο 4 του άξονα των αριθμών συστάδων έχει το υψηλότερο gap statistic. Με αυτόν τον τρόπο, όλα τα στοιχεία οδηγούν στην επιλογή k= 4.

Έπειτα, εφαρμόζεται η μέθοδος k-means μέσω της εντολής kmeans() για k= 4, εξασφαλίζοντας πρώτα ότι παράγονται οι ίδιες τυχαίες μεταβλητές κάθε φορά που τρέχει ο κώδικας, δημιουργώντας με αυτόν τον τρόπο αναπαράξιμα αποτελέσματα. Ως εκ τούτου ο κώδικας έχει ως εξής:

> set.seed(1)
> km.res <- kmeans(USArrestsScaling, centers=4, nstart = 25)
> km.res
K-means clustering with 4 clusters of sizes 13, 13, 16, 8
Cluster means:

	Murder	Assault	UrbanPop	Rape
1	-0.9615407	-1.1066010	-0.9301069	-0.96676331
2	0.6950701	1.0394414	0.7226370	1.27693964
3	-0.4894375	-0.3826001	0.5758298	-0.26165379
4	1.4118898	0.8743346	-0.8145211	0.01927104

Within cluster sum of squares by cluster:
11.952463 19.922437 16.212213 8.316061
(between_SS / total_SS = 71.2%)

Available components:
"cluster" "centers" "totss" "withinss" "tot.withinss"
"betweenss" "size" "iter" "ifault"

Έτσι, στην παραπάνω εντολή *kmeans(USArrestsScaling, centers=4, nstart = 25)* έχουν δηλωθεί ο αριθμός των συστάδων (4) καθώς και ο αριθμός των επαναλήψεων, κατά τις οποίες θα τρέξει ο κώδικας, δίνοντας το καλύτερο αποτέλεσμα 25. Αυτός ο αριθμός είναι σημαντικό να δοθεί, και μάλιστα να είναι αρκετά υψηλός (nstart= 25 ή nstart= 50) έτσι, ώστε ο εν λόγω αλγόριθμος να αποδώσει το καλύτερο αποτέλεσμα χωρίς το φόβο για τυχόν υπολογισμό τοπικού βελτίστου.

Το output της εντολής αυτής περιέχει αρχικά τον αριθμό των πολιτειών, που ανήκουν σε κάθε συστάδα, το άθροισμα τετραγώνων και τη μέση τιμή κάθε συστάδας. Επίσης, εμφανίζεται αναλύτικα στον πίνακα 4.6 η αντιστοιχία κάθε πολιτείας στην κάθε συστάδα. Με αυτόν τον τρόπο, για παράδειγμα, είναι

Clustering vectors:			
Alabama	Alaska	Arizona	Arkansas
4	2	2	4
California	Colorado	Connecticut	Delaware
2	2	3	3
Florida	Georgia	Hawaii	Idaho
2	4	3	1
Illinois	Indiana	Iowa	Kansas
2	3	1	3
Kentucky	Louisiana	Maine	Maryland
1	4	1	2
Massachusetts	Michigan	Minnesota	Mississippi
3	2	1	4
Missouri	Montana	Nebraska	Nevada
2	1	1	2
New Hampshire	New Jersey	New Mexico	New York
1	3	2	2
North Carolina	North Dakota	Ohio	Oklahoma
4	1	3	3
Oregon	Pennsylvania	Rhode Island	South Carolina
3	3	3	4
South Dakota	Tennessee	Texas	Utah
1	4	2	3
Vermont	Virginia	Washington	West Virginia
1	3	3	1
Wisconsin	Wyoming		
1	3		

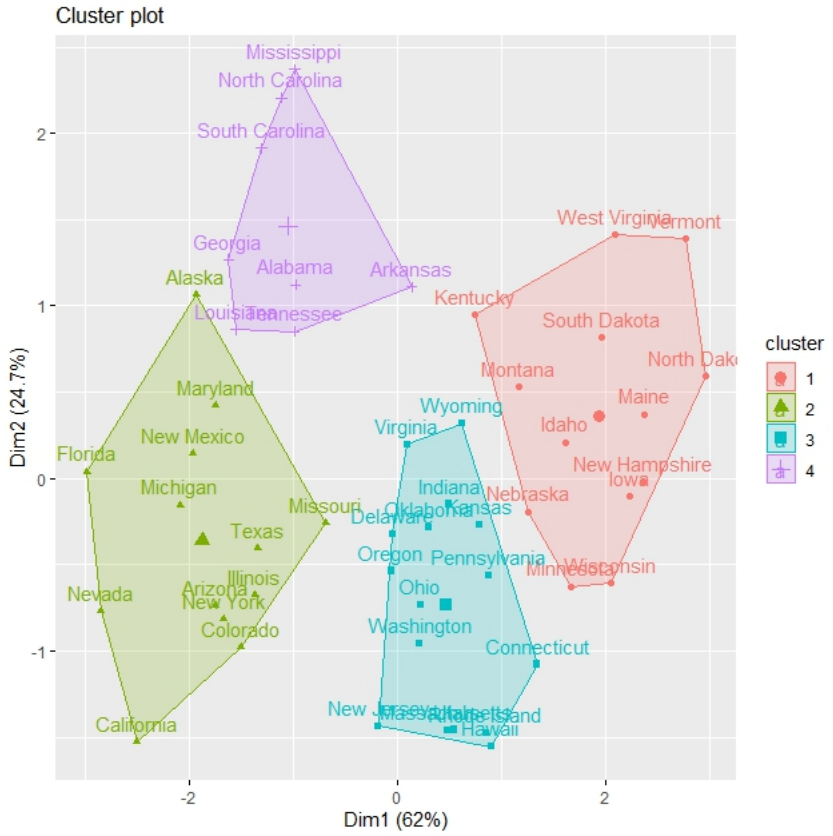
Πίνακας 4.6: *Output εντολής km.res <- kmeans(USArrestsScaling, centers=4, nstart = 25)*

φανερό ότι η πολιτεία Alaska ανήκει στη δεύτερη συστάδα μαζί με την Arizona, California, Colorado, Florida, Illinois, Maryland, Michigan, Missouri, Nevada, New Mexico, New York, Texas κ.ο.κ.

Παράλληλα, η αντιστοιχία της κάθε πολιτείας στην κάθε συστάδα μπορεί να γίνει γνωστή και μέσω του γραφήματος 4.7, το οποίο παράγεται από

την παρακάτω εντολή στην R:

> *fviz_cluster(km.res, data = USArrestsScaling)*



Σχήμα 4.7: *Διάγραμμα αντιστοιχίας κάθε πολιτείας σε κάθε συστάδα για τα δεδομένα USArrests με τη μέθοδο K-means.*

Έτσι, γίνεται πασιφανές ότι οι πολιτείες με κόκκινο χρώμα ανήκουν στην πρώτη συστάδα, αυτές με πράσινο στη δεύτερη, αυτές με μπλε στην τρίτη και τέλος αυτές με μωβ στην τέταρτη.

Έπειτα, θεωρείται σκόπιμη η εύρεση της μέσης τιμής της κάθε μεταβλητής για την εκάστοτε συστάδα. Αυτή η εύρεση επιτυγχάνεται μέσω της

παρακάτω εντολής στην R και το αποτέλεσμα της εμφανίζεται στον πίνακα 4.7.

> aggregate(USArrests, by=list(cluster=km$cluster), mean)

cluster	Murder	Assault	UrbanPop	Rape
1	3.60000	78.53846	52.07692	12.17692
2	10.81538	257.38462	76.00000	33.19231
3	5.65625	138.87500	73.87500	18.78125
4	13.93750	243.62500	53.75000	21.41250

Πίνακας 4.7: *Output εντολής aggregate(USArrests, by=list(cluster=km$cluster), mean).*

Αν κάποιος θα ήθελε να ερμηνεύσει τον πίνακα 4.7, τότε θα ισχυριζόταν ότι, ανά 100.000 πολίτες, το ποσοστό των φόνων (Murders) ανέρχεται κατά μέσο όρο στο 3.6 για τις πολιτείες, οι οποίες ανήκουν στην πρώτη συστάδα, δηλαδή στις Idaho, Iowa, Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, South Dakota, Vermont, West Virginia και Wisconsin. Αντίστοιχα, η μέση τιμή των βιαιοπραγιών Assaults για τις εν λόγω πολιτείες είναι 78.5, ενώ των βιασμών Rapes είναι 12.2.Παράλληλα, σε αυτές τις πολιτείες το μέσο ποσοστό των κατοίκων ανέρχεται σε αυτό των 52.1.

Σε ό,τι αφορά τις πολιτείες, οι οποίες απαρτίζουν τη δεύτερη συστάδα και οι οποίες είναι οι Alaska, Arizona, California, Colorado, Florida, Illinois, Maryland, Michigan, Missouri, Nevada, New Mexico, New York και Texas, το μέσο ποσοστό των φόνων είναι 10.8, των βιαιοπραγιών 257.4 και των βιασμών 33.2, ανά 100.000 πολίτες. Επίσης, 76.0 των πολιτών κατοικούν σε αυτές τις πολιτείες.

Επιπρόσθετα, αναφορικά με τις πολιτείες της τρίτης συστάδας, ήτοι Connecticut, Delaware, Hawaii, Indiana, Kansas, Massachusetts, New Jersey, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Utah, Virginia, Washington και Wyoming , και ανά 100.000 πολίτες, οι φόνοι καταλαμβάνουν το μέσο ποσοστό των 5.7, οι βιαιοπραγίες καταλαμβάνουν αυτό των 138.9, οι κάτοικοι σε αυτές τις πολιτείες ανέρχονται σε αυτό των 73.9 και οι βιασμοί αποτελούν το 18.9 αντίστοιχα.

Τέλος, όταν πρόκειται για τις πολιτείες, οι οποίες αποτελούν μέρη της τέταρτης

συστάδας, οι φόνοι έχουν μέση τιμή 13.9, οι βιαιοπραγίες 243.6, οι βιασμοί 21.4 ανά 100.000 πολίτες. Ακόμα, 53.8 ζούν σε αυτές τις πολιτείες (Alabama, Arkansas, Georgia, Louisiana, Mississippi, North Carolina, South Carolina και Tennessee).

Μέθοδος Ιεραρχικής Ανάλυσης

Ένας εναλλακτικός τρόπος εύρεσης συστάδων ανάμεσα στα δεδομένα είναι η μέθοδος ιεραρχικής ανάλυσης. Όπως έχει ειπωθεί στο τρίτο κεφάλαιο, είναι απαραίτητη η μέτρηση της διαφορετικότητας των ζευγών, καθώς θεωρείται χρήσιμη η επιλογή του τρόπου ζεύξης στα δεδομένα. Εξαιτίας του παραπάνω ισχυρισμού, λοιπόν, επιλέγεται για τα δεδομένα USArrests η Ευκλείδεια απόσταση ως μέσο για τον υπολογισμό της διαφορετικότητας των ζευγών, και εφαρμόζονται και αναλύονται οι πιο βασικοί τρόποι ζεύξης (δηλαδή η μονή, ολοκληρωμένη και η μέση). Με αυτόν τον τρόπο, η εφαρμογή στην R της μεθόδου ιεραρχικής ανάλυσης στα δεδομένα USArrests, τα οποία έχουν τεθεί υπό κλίμακα ως USArrestsScaling από την προηγούμενη υποενότητα, έχει ως εξής:

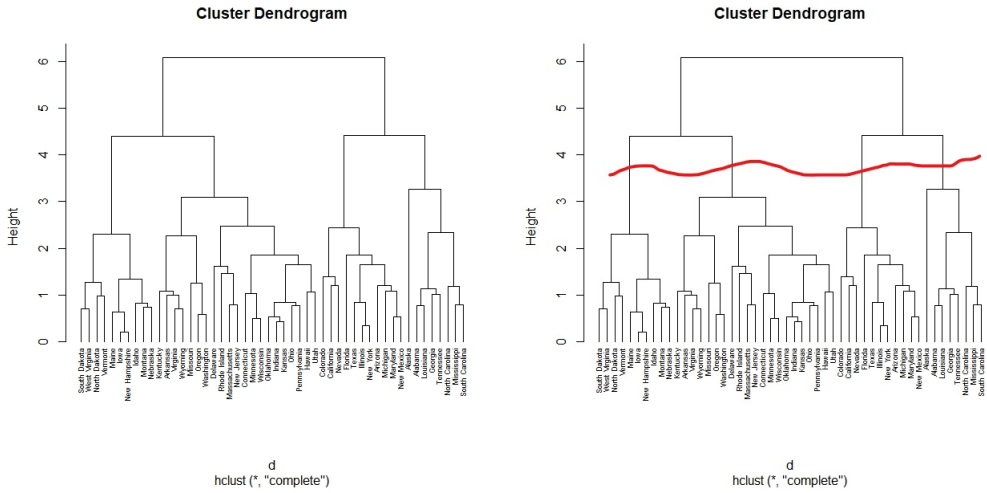
> d <- dist(USArrestsScaling, method = "euclidean")

Έτσι, με την παραπάνω εντολή γίνεται η μέτρηση της Ευκλείδιας απόστασης μεταξύ των παρατηρήσεων, προκειμένου να γίνει η μέτρηση της διαφορετικότητάς τους. Έπειτα, με βάση την επιλογή της ολοκληρωμένης ζεύξης για αρχή, οι εντολές και το αντίστοιχο δενδρόγραμμα, το οποίο δίνεται σαν αποτέλεσμα, έχουν ως εξής:

> hc.complete.us <- hclust(d, method = "complete")
> plot(hc.complete.us, cex = 0.6, hang = -1)

Το δενδρόγραμμα από την παραπάνω εντολή εμφανίζεται στο αριστερό σχήμα 4.8.

Έτσι, αναλύοντας το εν λόγω δενδρόγραμμα και έχοντας ως στόχο την εύρεση του αριθμού των συστάδων, οι οποίες αντιπροσωπεύουν το δείγμα USArrests, θεωρείται αναγκαία η σχεδίαση μίας γραμμής, η οποία θα σχίζει το δενδρόγραμμα και η οποία δεν είναι απόρροια κάποιου αντικειμενικού κανόνα, αλλά εναπόκειται σε υποκειμενικές επιλογές. Δεδομένου αυτού του γεγονότος, η γραμμή αυτή θα μπορούσε να σχηματιστεί όπως απεικονίζεται στο δεξί σχήμα 4.8. Παρατηρείται, λοιπόν, ότι η κόκκινη γραμμή τέμνει τέσσερεις



Σχήμα 4.8: *Δενδρόγραμμα για το δείγμα USArrests με ολοκληρωμένη ζεύξη*

ευθείες του δενδρογράμματος- γεγονός, το οποίο συνηγορεί στον ισχυρισμό ότι οι συστάδες, που συνιστούν τα δεδομένα, θα μπορούσαν να είναι τέσσερεις. Ως εκ τούτου, εξετάζοντας προσεκτικά το δεξί σχήμα 4.8 γίνεται αντιληπτό ότι οι πολιτείες, οι οποίες αποτελούν τα άκρα του πρώτου από τα τέσσερα κλαδιά, τα οποία τέμνει η κόκκινη γραμμή, ανήκουν στην πρώτη συστάδα. Αντίστοιχα, οι πολιτείες, οι οποίες περιλαμβάνονται στο δεύτερο κλαδί διαμορφώνουν τη δεύτερη συστάδα. Αυτές, οι οποίες είναι ΄φύλλα΄ του τρίτου κλαδιού αντιπροσωπεύουν την τρίτη συστάδα, και τέλος αυτές του τέταρτου κλαδιού, συνιστούν την τέταρτη συστάδα. Συνεπώς, οι Idaho, Iowa, Maine, Montana, Nebraska, New Hampshire, North Dakota, South Dakota, Vermont και West Virginia αποτελούν την πρώτη συστάδα, οι Connecticut, Delaware, Hawaii, Indiana, Kansas, Massachusetts, New Jersey, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Utah, Virginia, Washington, Kentucky, Arkansas, Missouri, Minnesota, Wisconsin και Wyoming τη δεύτερη, οι Arizona, California, Colorado, Florida, Illinois, Maryland, Michigan, Nevada, New Mexico, New York και Texas την τρίτη και οι Alaska, Alabama, Georgia, Louisiana, Mississippi, North Carolina, South Carolina και Tennessee την τέταρτη κατά αντιστοιχία.

Τα παραπάνω συμπεράσματα, θα μπορούσαν να επιβεβαιωθούν και από το αποτέλεσμα της εντολής > *cutree(hc.complete.us, k = 4)*, όπου k= 4

δηλώνει τον αριθμό των συστάδων. Το output της εντολής αυτής εμφανίζεται στον πίνακα 4.8. Παράλληλα, συνδυάζοντας την ανάλυση K-means, η οποία έχει προηγηθεί με αυτά του πίνακα 4.8, παρατηρείται μία αρκετή ομοιότητα στα αποτελέσματα των πολιτειών, οι οποίες ανήκουν σε κοινές συστάδες.

Alabama	Alaska	Arizona	Arkansas
4	4	3	2
California	Colorado	Connecticut	Delaware
3	3	2	2
Florida	Georgia	Hawaii	Idaho
3	4	2	1
Illinois	Indiana	Iowa	Kansas
3	2	1	2
Kentucky	Louisiana	Maine	Maryland
2	4	1	3
Massachusetts	Michigan	Minnesota	Mississippi
2	3	2	4
Missouri	Montana	Nebraska	Nevada
2	1	1	3
New Hampshire	New Jersey	New Mexico	New York
1	2	3	3
North Carolina	North Dakota	Ohio	Oklahoma
4	1	2	2
Oregon	Pennsylvania	Rhode Island	South Carolina
2	2	2	4
South Dakota	Tennessee	Texas	Utah
1	4	3	2
Vermont	Virginia	Washington	West Virginia
1	2	2	1
Wisconsin	Wyoming		
2	2		

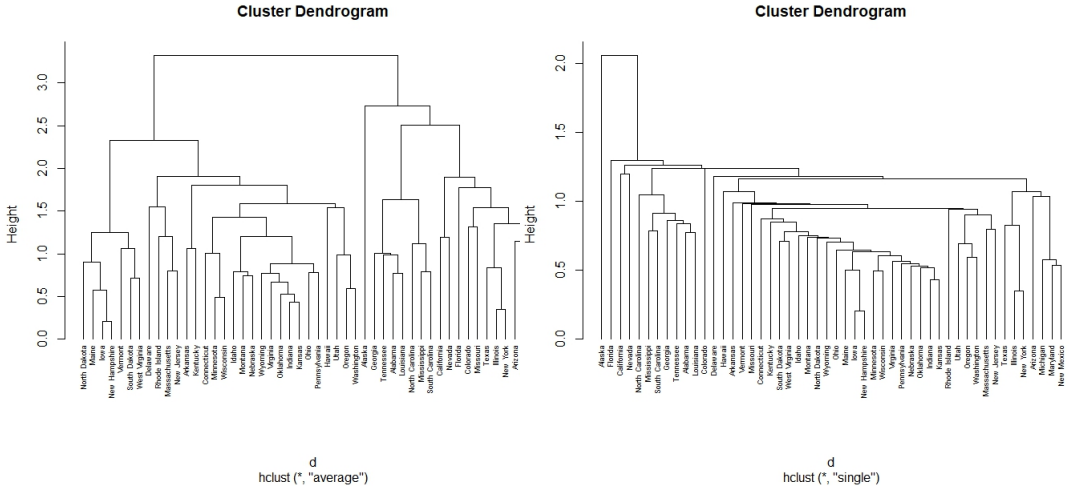
Πίνακας 4.8: *Output εντολής cutree(hc.complete.us, k = 4)*

Έπειτα, γίνεται η ιεραρχική ανάλυση συστάδων στο δείγμα USArrests με τη μέση και μονή ζεύξη, και στη συνέχεια σχεδιάζονται τα δενδρογράμματα, τα οποία προέρχονται από την αντίστοιχη ζεύξη και τα οποία εμφανίζονται στο

σχήμα 4.9 . Έτσι, οι εντολές στην R είναι οι παρακάτω:

> hc.average.us <- hclust(d, method = "average")
> hc.single.us <- hclust(d, method = "single")

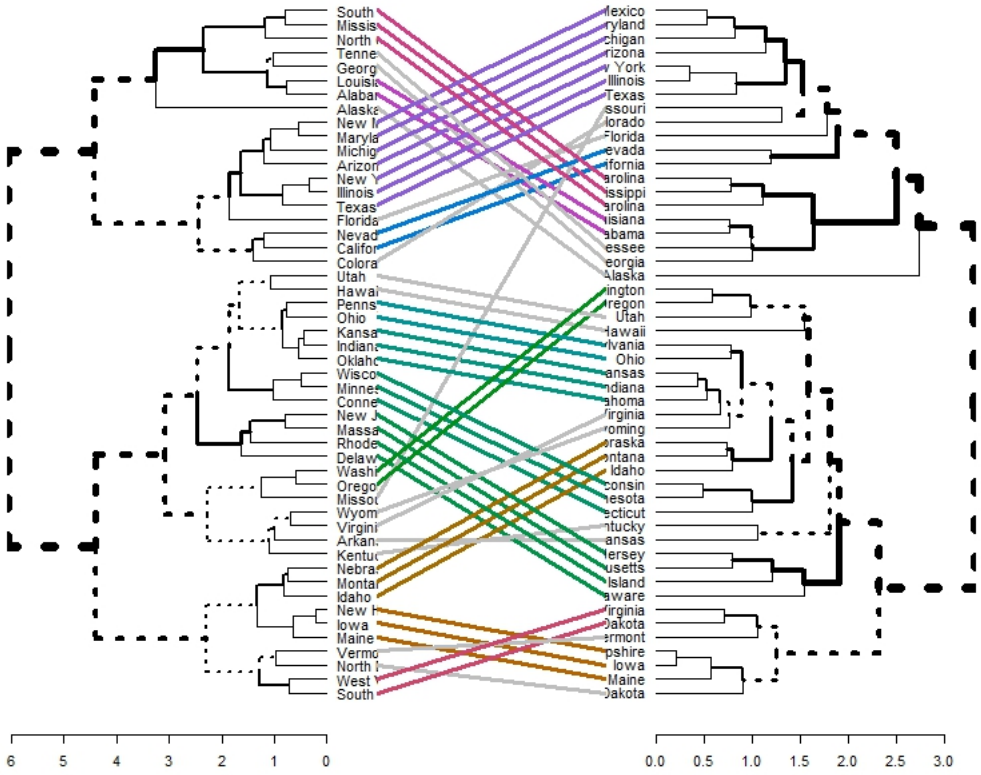
> plot(hc.average.us, cex = 0.6, hang = -1)
> plot(hc.single.us, cex = 0.6, hang = -1)



Σχήμα 4.9: *Δενδρογράμματα για το δείγμα USArrests με μέση και μονή ζεύξη*

Με μία πρώτη ματιά στο αριστερό δενδρόγραμμα του σχήματος 4.9, συνειδητοποιείται ότι οι τρεις συστάδες θα μπορούσαν να είναι αντιπροσωπευτικές όσο η ζεύξη είναι μέση. Αντίθετα, είναι φανερό από το δεξί δενδρόγραμμα του εν λόγω σχήματος ότι η μονή ζεύξη επιφέρει 2 συστάδες. Μάλιστα, παρατηρείται ότι η μία συστάδα αποτελείται από μία μόνο πολιτεία, ενώ η άλλη από όλες τις υπόλοιπες του δείγματος USArrests- γεγονός, το οποίο υποδηλώνει τη μη καταλληλότητα της μονής ζεύξης στην ανάλυση συστάδων του δείγματος USArrests.

Λαμβάνοντας όλα τα παραπάνω υπόψη, θα μπορούσε να είναι ωφέλιμη η σύγκριση των δενδρογραμμάτων, στα οποία το ένα υπακούει στην ολοκληρωμένη ζεύξη, ενώ το άλλο στη μέση. Πρόκειται, δηλαδή, για ζεύξεις, οι οποίες φαίνονται να είναι ορθές για τις παρατηρήσεις του δείγματος. Αυτή η σύγκριση επιτυγχάνεται μέσω της εντολής *tanglegram()* , η οποία είναι διαθέσιμη στο



πακέτο *dendextend* . Επιπρόσθετα, με τη βοήθεια της εντολής αυτής εμφανίζονται τα δύο δενδρογράμματα, το ένα δίπλα στο άλλο, και οι παρατηρήσεις του ενός δενδρογράμματος συνδέονται με τις ίδιες παρατηρήσεις ίδιου χρώματος, οι οποίες υπάρχουν στο άλλο δενδρόγραμμα. Επίσης, στα δενδρογράμματα σχηματίζονται διακεκομμένες γραμμές επισημαίνοντας με αυτό τον τρόπο ότι σε αυτά τα κλαδιά με τις διακεκομμένες υπάρχει απόκλιση σε σχέση με το αντίστοιχο δενδρόγραμμα του άλλου τρόπου ζεύξης. Οι εντολές στην R για την επίτευξη της παραπάνω διαδικασίας έχουν ως εξής:

```
> library(dendextend)
> dend1 <- as.dendrogram (hc.complete.us)
> dend2 <- as.dendrogram (hc.average.us)
> tanglegram(dend1, dend2)
```

Είναι, λοιπόν, αντιληπτό ότι οι διαφορές μεταξύ των δύο τρόπων ζεύξης είναι μικρές. Με μία πρώτη ματιά θα μπορούσε να βγει σαν πρόχειρο συμπέρασμα ότι η μία (αυτή που βρίσκεται στο πάνω μέρος του δεξιού δενδρογράμματος) από τις 3 συστάδες της μέσης ζεύξης περιλαμβάνει τις πολιτείες των δύο συστάδων της ολοκληρωμένης ζεύξης καθώς και περιέχει και την πολιτεία Missouri, η οποία ανήκει σε μία εντελώς διαφορετική συστάδα της ολοκληρωμένης ζεύξης. Επιπλέον, φαίνεται ότι η τρίτη συστάδα της μέσης ζεύξης (αυτή που βρίσκεται στο κάτω μέρος του αριστερού δενδρογράμματος) να ταυτίζεται σχεδόν με την τέταρτη συστάδα της ολοκληρωμένης ζεύξης, με εξαίρεση τις πολιτείες Nebraska, Montana και Idaho, οι οποίες ανήκουν στη δεύτερη συστάδα της μέσης ζεύξης.

4.2 Δεδομένα heptathlon

Τα δεδομένα heptathlon (Everitt και Hothorn, 2010) έχουν σχέση με συγκεντρωμένους βαθμούς και αποτελέσματα γυναικών αθλητριών, οι οποίες διαγωνίστηκαν στο αγώνισμα του επτάθλου στους Ολυμπιακούς Αγώνες το 1988 στην πρωτεύουσα της Νότιας Κορέας, Σεούλ. Στο δείγμα αυτό έχουν καταγραφεί οι επιδόσεις 25 αθλητριών από όλον τον κόσμο και στα 7 αγωνίσματα του επτάθλου, δηλαδή στο στίβο στα 100 μέτρα μετ΄ εμποδίων (hurdles), στα 100 μέτρα άλμα εις ύψος (highjump), στη σφαιροβολία (shot), στα 200 μέτρα στίβο (run200m), στα 200 μέτρα άλμα εις μήκος (longjump), στα 200 μέτρα ακοντισμό (javelin) και στα 800 μέτρα στίβο (run800m). Τέλος, έχει σημειωθεί η συνολική βαθμολογία για κάθε αθλήτρια του επτάθλου (score).
Για παράδειγμα, αν θα επιθυμούσε κάποιος να ρίξει μία ματιά στον πίνακα δεδομένων για τις επιδόσεις των πρώτων τριών καταγεγραμμένων αθλητριών στα 6 πρώτα αθλήματα και τα αντίστοιχα αποτελέσματά τους, θα εκτελούσε τις παρακάτω εντολές στην R.

```
> data("heptathlon", package = "HSAUR2")
> heptathlon[1:3,1:6]
```

Στον πίνακα 4.10 φαίνονται οι δειγματικές μέσες τιμές και ο δειγματικός μέσος των επιδόσεων για κάθε άθλημα του επτάθλου καθώς και για τις βαθμολογίες των αθλητριών. Εμφανίζονται, δηλαδή, οι δειγματικές ελάχιστες, μέσες και μέγιστες τιμές, οι δειγματικές διάμεσοι καθώς και οι τιμές του πρώτου και τρίτου τεταρτημορίου για κάθε άθλημα και για τη σχετική βαθμολογία.

	hurdles	highjump	shot	run200m
Joyner-Kersee (USA)	12.69	1.86	15.80	22.56
John (GDR)	12.85	1.80	16.23	23.65
Behmer (GDR)	13.20	1.83	14.20	23.10
	longjump	**javelin**		
Joyner-Kersee (USA)	7.27	45.66		
John (GDR)	6.71	42.56		
Behmer (GDR)	6.68	44.54		

Πίνακας 4.9: *Αποτέλεσμα της εντολής heptathlon[1:3,1:6]*

	hurdles	highjump	shot	run200m
Min.	12.69	1.500	10.00	22.56
1st Q.	13.47	1.770	12.32	23.92
Med.	13.75	1.800	12.88	24.83
Mean	13.84	1.782	13.12	24.65
3rd Q.	14.07	1.830	14.20	25.23
Max.	16.42	1.860	16.23	26.61
	longjump	**javelin**	**run800m**	**score**
Min.	4.880	35.68	124.2	4566
1st Q.	6.050	39.06	132.2	5746
Med.	6.250	40.28	134.7	6137
Mean	6.152	41.48	136.1	6091
3rd Q.	6.370	44.54	138.5	6351
Max.	7.270	47.50	163.4	7291

Πίνακας 4.10: *Σύνοψη δεδομένων heptathlon έχοντας εφαρμόσει την εντολή στην R > summary(heptathlon)*

> score <- which(colnames(heptathlon) == "score")
> round(cor(heptathlon[,-score]), 2)

> plot(heptathlon[,-score])

Το διάγραμμα συσχετίσεων για τα αθλήματα του δείγματος απεικονίζεται στο σχήμα 4.10.

	hurdles	highjump	shot	run200m	longjump	javelin
hurdles	1.00	-0.81	-0.65	0.77	-0.91	-0.01
highjump	-0.81	1.00	0.44	-0.49	0.78	0.00
shot	-0.65	0.44	1.00	-0.68	0.74	0.27
run200m	0.77	-0.49	-0.68	1.00	-0.82	-0.33
longjump	-0.91	0.78	0.74	-0.82	1.00	0.07
javelin	-0.01	0.00	0.27	-0.33	0.07	1.00
run800m	0.78	-0.59	-0.42	0.62	-0.70	0.02

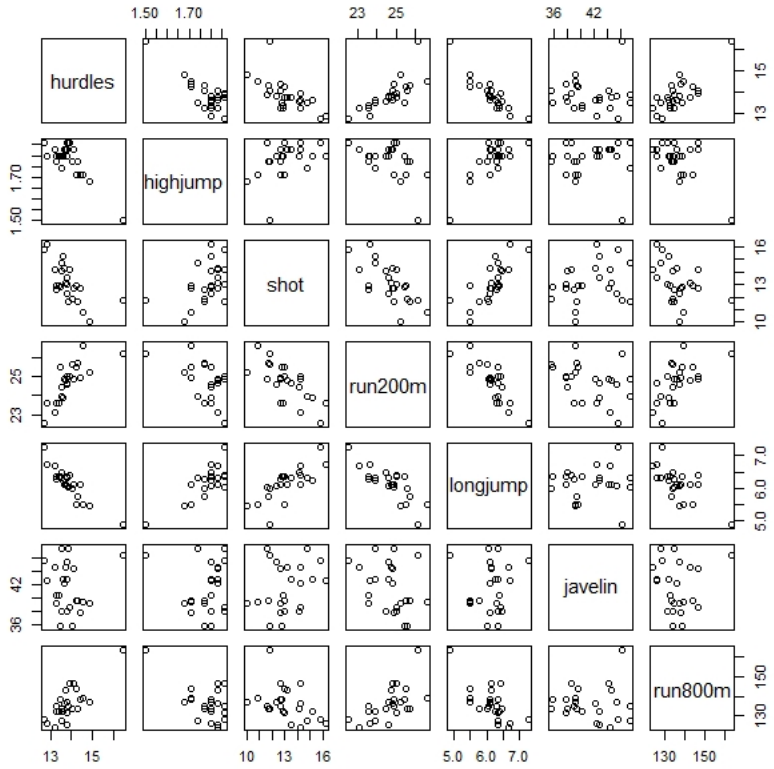
	run800m
hurdles	0.78
highjump	-0.59
shot	-0.42
run200m	0.62
longjump	-0.70
javelin	0.02
run800m	1.00

Πίνακας 4.11: *Πίνακας συσχετίσεων για τις μεταβλητές του δείγματος heptathlon εκτός από τη μεταβλητή score*

Από τον πίνακα 4.11 καθίσταται εμφανές ότι οι μεταβλητές και κατ΄ επέκταση τα αγωνίσματα έχουν υψηλή συσχέτιση μεταξύ τους με μόνη εξαίρεση το άθλημα του ακοντισμού, για το οποίο διαπιστώνεται ότι έχει πολύ χαμηλή έως μηδαμινή σε κάποιες περιπτώσεις συσχέτιση με όλα τα υπόλοιπα αθλήματα. Επιπλέον, στο ίδιο πόρισμα καταλήγει κανείς και με τη βοήθεια του διαγράμματος συσχετίσεων 4.10, στο οποίο παρατηρείται ότι όλα τα αθλήματα εκτός του ακοντισμού μεταξύ τους σχηματίζουν περίπου μία νοητή γραμμή. Αυτό το γεγονός - της μηδενικής συσχέτισης του ακοντισμού με τις υπόλοιπες μεταβλητές του δείγματος- εγείρει σύγχυση.

> min(heptathlon$score)
4566

> which(heptathlon$score== 4566)
25



Σχήμα 4.10: *Διάγραμμα συσχετίσεων στο δείγμα heptathlon για τις κατηγορίες των αθλημάτων.*

> heptathlon[25,]

	hurdles	highjump	shot	run200m	longjump
Launa (PNG)	16.42	1.5	11.78	26.16	4.88
	javelin	run800m	score		
Launa (PNG)	46.38	163.43	4566		

> which(heptathlon$javelin < 46.38)
1 2 3 4 6 7 8 9 10 11 12 14 15 16 17 18 19 20 21 22 23 24

Παράλληλα, λαμβάνοντας υπόψη τα παραπάνω αποτελέσματα των

εντολών, αν παρατηρήσει κάποιος προσεκτικά τις επιδόσεις των αθλητριών στο δείγμα και ειδικά της αθλήτριας Launa, τότε θα διαπιστώσει ότι αυτή είναι η αθλήτρια με τη μικρότερη βαθμολογία, με βάση την επίδοση της στα αθλήματα (4566). Ωστόσο, η αθλήτρια αυτή έχει υψηλή επίδοση στο αγώνισμα του ακοντισμού, δεδομένου ότι οι 22 άλλες αθλήτριες έχουν χειρότερη επίδοση από αυτήν σε αυτό το άθλημα. Αυτή η αντίφαση οδηγεί στη σκέψη της δοκιμής της παράλειψης αυτής της αθλήτριας από τη μελέτη και την ανάλυση του δείγματος με προοπτική την υποχώρηση της μηδενικής συσχέτισης. Με αυτόν τον τρόπο, βρίσκονται οι συσχετίσεις και το διάγραμμα συσχετίσεων εκ νέου χωρίς να λαμβάνεται υπ' όψιν η αθλήτρια Launa με τη χαμηλότερη επίδοση.

> heptathlon <- heptathlon[-grep("PNG", rownames(heptathlon)),]
> round(cor(heptathlon[,-score]), 2)

	hurdles	highjump	shot	run200m	longjump	javelin
hurdles	1.00	-0.58	-0.77	0.83	-0.89	-0.33
highjump	-0.58	1.00	0.46	-0.39	0.66	0.35
shot	-0.77	0.46	1.00	-0.67	0.78	0.34
run200m	0.83	-0.39	-0.67	1.00	-0.81	-0.47
longjump	-0.89	0.66	0.78	-0.81	1.00	0.29
javelin	-0.33	0.35	0.34	-0.47	0.29	1.00
run800m	0.56	-0.15	-0.41	0.57	-0.52	-0.26

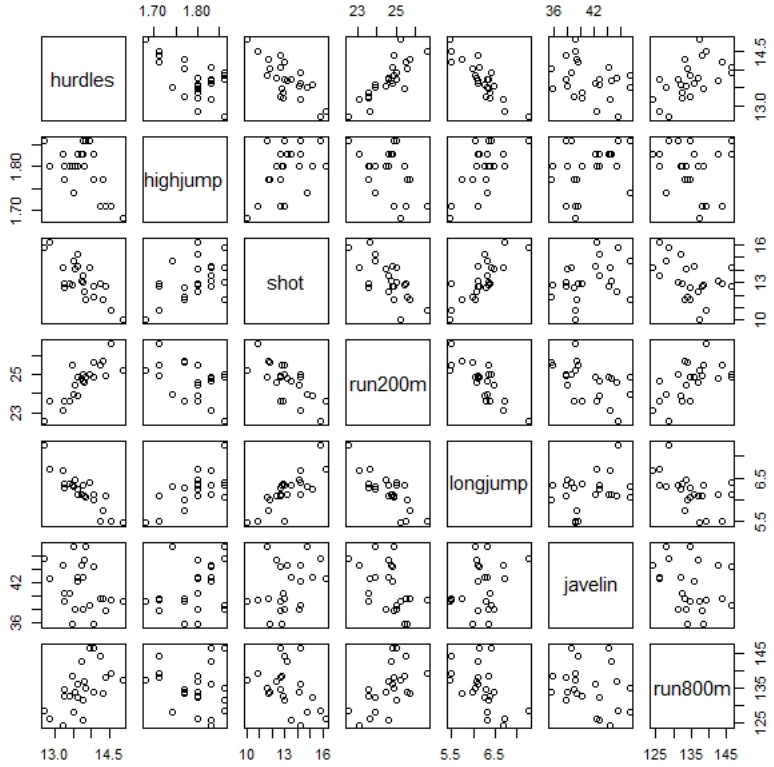
	run800m
hurdles	0.56
highjump	-0.15
shot	-0.41
run200m	0.57
longjump	-0.52
javelin	-0.26
run800m	1.00

Πίνακας 4.12: *Πίνακας συσχετίσεων για τις μεταβλητές του δείγματος heptathlon εκτός από τη μεταβλητή score έχοντας αφαιρέσει από το δείγμα τις επιδόσεις της αθλήτριας Launa (PNG)*

Είναι διακριτό ότι ο τελευταίος πίνακας συσχετίσεων 4.12 περιέχει μη μηδενικές τιμές σε αντίθεση με τον πίνακα 4.11. Έτσι, παρατηρείται μία βελτίωση στην ανάλυση των δεδομένων, την οποία αναμένει κάποιος να δει και στο διάγραμ-

μα συσχετίσεων, στο οποίο θα έχουν αφαιρεθεί οι πληροφορίες της αθλήτριας Launa.

> *plot(heptathlon[,-score])*



Σχήμα 4.11: *Διάγραμμα συσχετίσεων στο δείγμα heptathlon χωρίς την αθλήτρια Launa.*

> *apply(heptathlon[,-score]), 2, var)*

Σημειώνεται ότι οι δειγματικές διασπορές κάθε αθλήματος του επτάθλου διαφέρουν κατά πολύ μεγάλο βαθμό. Ωστόσο, δεν μπορεί να γίνει σύγκριση

hurdles	**highjump**	**shot**	**run200m**	**longjump**	**javelin**
0.2647761	0.0027375	2.2414580	0.8775375	0.1613303	12.0319275
run800m					
37.7886580					

μεταξύ των μεταβλητών, δεδομένου ότι διαφέρουν οι μονάδες μέτρησής τους, εφόσον μερικά αθλήματα από αυτά βασίζονται στον χρόνο επίδοσης, ενώ άλλα σε μονάδες μήκους. Για αυτόν το λόγο, απαιτείται η κατάταξη των δεδομένων σε κλίμακα.

4.2.1 Εφαρμογές Ανάλυσης Κύριων Συνιστωσών στο δείγμα δεδομένων heptathlon

Λαμβάνοντας υπόψιν την αναγκαιότητα κατάταξης των δεδομένων σε κλίμακα καθιστώντας κάθε μεταβλητή να έχει μέση τιμή 0 και τυπική απόκλιση 1, γίνεται εφικτή η εύρεση των κύριων συνιστωσών και των φορτίσεών τους για το δείγμα heptathlon.

> pca.res <- prcomp(heptathlon[, -score], scale = TRUE)
> pca.res$rotation

Στον πίνακα 4.13 εμφανίζονται οι φορτίσεις των 7 κύριων συνιστωσών του δείγματος. Είναι πασιφανές ότι πρέπει να μειωθεί ο αριθμός των συνιστωσών και να επιλεχθούν οι κατά το δυνατόν λιγότερες, οι οποίες θα εκφράζουν επαρκώς το δείγμα heptathlon.

> pca.var =pca.res$sdev ^ 2
> pca.var

4.32364217 0.89899445 0.82974172 0.46675769 0.29832218 0.11387578 0.06866602

Με την παραπάνω εντολή παρατηρείται η διασπορά, την οποία έχει ως τιμή κάθε κύρια συνιστώσα. Αναλυτικότερα, η πρώτη κύρια συνιστώσα έχει διασπορά περίπου 4.4, η δεύτερη έχει πολύ μικρότερη με τιμή σχεδόν 0.9, η τρίτη έχει ακόμα λίγο μικρότερη περίπου 0.8, η τέταρτη έχει την επίσης μικρότερη και φτάνει στην τιμή 0.5 κατά προσέγγιση, η πέμπτη έχει τιμή περίπου 0.3, η έκτη 0.1 και η έβδομη έχει την ελάχιστη τιμή, η οποία τείνει προς το μηδέν.

	PC1	**PC2**	**PC3**	**PC4**
hurdles	0.4503876	-0.05772161	-0.1739345	0.04840598
highjump	-0.3145115	-0.65133162	0.2088272	0.55694554
shot	-0.4024884	-0.02202088	0.1534709	-0.54826705
run200m	0.4270860	-0.18502783	0.1301287	0.23095946
longjump	-0.4509639	-0.02492486	0.2697589	0.01468275
javelin	-0.2423079	-0.32572229	-0.8806995	-0.06024757
run800m	0.3029068	-0.65650503	0.1930020	-0.57418128
	PC5	**PC6**	**PC7**	
hurdles	0.19889364	-0.84665086	0.06961672	
highjump	0.07076358	-0.09007544	0.33155910	
shot	0.67166466	-0.09886359	0.22904298	
run200m	0.61781764	0.33279359	-0.46971934	
longjump	-0.12151793	-0.38294411	-0.74940781	
javelin	0.07874396	0.07193437	-0.21108138	
run800m	-0.31880178	0.05217664	-0.07718616	

Πίνακας 4.13: *Φορτίσεις κύριων συνιστωσών του δείγματος heptathlon*

Συμπερασματικά, με μία πρώτη ματιά θα μπορούσε ίσως να εικαστεί ότι οι δύο ή τρεις πρώτες κύριες συνιστώσες εκφράζουν ένα αρκετά υψηλό βαθμό διασποράς και μεταβλητότητας στο δείγμα. Ο έλεγχος για τον παραπάνω ισχυρισμό γίνεται φυσικά μέσω της εύρεσης της αναλογίας εξηγούμενης διασποράς κάθε συνιστώσας στο συγκεκριμένο δείγμα, της γραφικής αναπαράστασης της, καθώς επίσης της συσσωρευμένης αναλογίας της από κάθε συνιστώσα. Με τη βοήθεια της R ο έλεγχος έχει ως εξής:
> *summary(pca.res)*

Μετατρέποντας προσεγγιστικά τις τιμές των αναλογιών διασποράς σε ποσοστά η διασπορά της πρώτης κύριας συνιστώσας εκφράζει ένα αρκετά μεγάλο ποσοστό του δείγματος ,περίπου το 61.8%, της δεύτερης το 12.8% του δείγματος , της τρίτης το 11.9%, της τέταρτης μόνο το 6.7%, της πέμπτης μόνο το 4.3%, της έκτης ελάχιστα το 1.6% και τέλος της έβδομης σχεδόν το μηδενικό ποσοστό της τάξης 0.1%. Σε ό,τι αφορά στις ποσοστιαίες συσσωρευμένες αναλογίες διασπορών, η πρώτη κύρια συνιστώσα κατέχει περίπου το ποσοστό των 61.8%, η δεύτερη συνυπολογισμένη φέρει περίπου το 74.6% του δείγματος, η τρίτη συνυπολογισμένη συμβάλλει στην εξήγηση του 86.5% του δείγματος, η τέταρτη με τη σειρά της αυξάνει το ποσοστό στο ύψος του 93.1% και η πέμ-

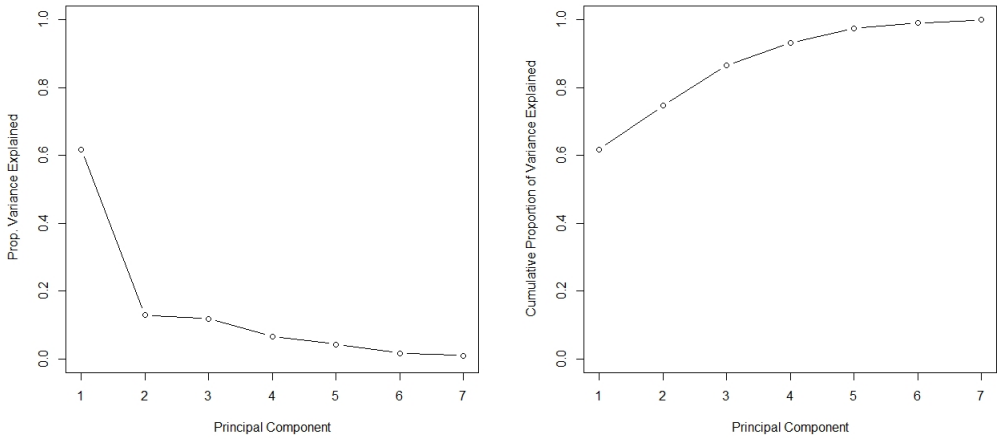
Importance of components					
	PC1	PC2	PC3	PC4	PC5
Standard deviation	2.0793	0.9482	0.9109	0.68320	0.54619
Proportion of Variance	0.6177	0.1284	0.1185	0.06668	0.04262
Cumulative Proportion	0.6177	0.7461	0.8646	0.93131	0.97392
	PC6	PC7			
Standard deviation	0.33745	0.26204			
Proportion of Variance	0.01627	0.00981			
Cumulative Proportion	0.99019	1.00000			

Πίνακας 4.14: *Τυπικές αποκλίσεις, αναλογία εξηγούμενης διασποράς και συσσωρευμένη αναλογία εξηγούμενης διασποράς των κύριων συνιστωσών του δείγματος heptathlon*

πτη αντίστοιχα στο 97.4%. Τέλος, η συνεισφορά στην εξηγούμενη διασπορά του δείγματος των άλλων δύο συνιστωσών είναι πολύ μικρή έως και αμελητέα. Είναι προφανές, λοιπόν, ότι πάλι υπάρχει η ένδειξη ότι οι δύο πρώτες κύριες συνιστώσες εκφράζουν το δείγμα κατά ένα πολύ σημαντικό ποσοστό καθώς και η διαπίστωση της ισχυρής συνεισφοράς και της τρίτης κύριας συνιστώσας
Έπειτα, ακολουθείται ο έλεγχος των γραφικών παραστάσεων των παραπάνω αναλογιών, ο οποίος θα διασαφηνίσει στο έπακρο τον αριθμό των απαιτούμενων κύριων συνιστωσών, οι οποίες χρειάζονται, για να περιγράψουν επαρκώς το δείγμα.

> var.ratio=pca.var/sum(pca.var)
> plot(var.ratio , xlab=" Principal Component ", ylab=" Prop. Variance Explained" , ylim=c(0,1) ,type="b")
> plot(cumsum (var.ratio), xlab=" Principal Component ", ylab =" Cumulative Proportion of Variance Explained ", ylim=c(0,1) ,type="b")

Οι γραφικές παραστάσεις, οι οποίες απεικονίζονται στο σχήμα 4.12, συνδράμουν στην απόφαση του επιθυμητού αριθμού κυρίων συνιστωσών. Στα αριστερά του σχήματος υπάρχει η γραφική παράσταση αναλογίας διασποράς των συνιστωσών, ενώ στα δεξιά διακρίνεται η γραφική παράσταση της συσσωρευμένης αναλογίας τους. Ειδικότερα, στο αριστερό γράφημα επαληθεύεται ότι η πρώτη κύρια συνιστώσα εκφράζει το μεγαλύτερο μέρος της διασποράς, και η δεύτερη και η τρίτη κύρια συνιστώσα εκφράζουν μικρότερο ποσοστό από την



Σχήμα 4.12: *Διάγραμματα αναλογιών - συσσωρευμένων αναλογιών διασποράς κύριων συνιστωσών στο δείγμα heptathlon*

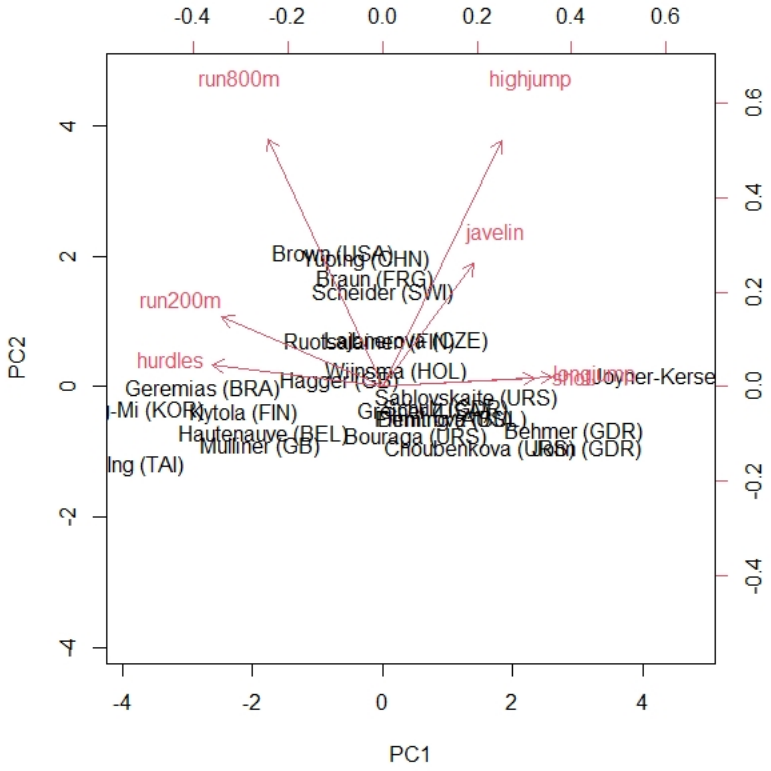
πρώτη αλλά ευρύτερα ένα αρκετά ικανοποιητικό ποσοστό του αρχικού δείγματος heptathlon. Στο δεξιό γράφημα γίνεται αντιληπτό ξανά, όπως είχε υποδείξει και ο πίνακας 4.14, ότι οι πρώτες δύο συνιστώσες μαζί εκφράζουν με τη πρέπουσα επάρκεια τα δεδομένα του αρχικού προβλήματος έχοντας συσσωρευμένη διασπορά πάνω από 0.7.

Συνεπώς, μπορεί κάποιος να αποφασίσει ότι οι πρώτες δύο κύριες συνιστώσες είναι αυτές, που περιγράφουν κατάλληλα το δείγμα, εφόσον εκφράζουν περίπου το 75% της διασποράς του αρχικού δείγματος. Πρόκειται για ένα αρκετά ικανοποιητικό ποσοστό μεγαλύτερο του 70%. Επιπλέον, με βάση τη γραφική παράσταση της αναλογίας διασπορών είναι εμφανές ότι στο δεύτερο σημείο (δεύτερη κύρια συνιστώσα) σημειώνεται ένας βραχίονας, αλλάζει δηλαδή απότομα η κλίση της ευθείας. Συνυπολογίζοντας το γεγονός ότι περίπου η τρίτη κύρια συνιστώσα έχει διασπορά 0.8, η τέταρτη 0.5, η πέμπτη 0.3, η έκτη 0.1 και η έβδομη 0.1, παρατηρούμε ότι αυτές οι τιμές είναι μικρότερες από τη μέση διασπορά του δείγματος, η οποία ισούται με 1. Όλοι οι παραπάνω ισχυρισμοί οδηγούν στην απόφαση ότι χρειαζόμαστε μόνο τις δύο πρώτες κύριες συνιστώσες για την περιγραφή, την επεξήγηση και την οπτικοποίηση του αρχικού δείγματος χάνοντας αμελητέα στατιστική πληροφορία και χρησιμοποιώντας τις λιγότερες μεταβλητές.

Με αυτόν τον τρόπο, έχοντας επιλέξει πλέον τις δύο πρώτες κύριες συνιστώσες και παρατηρώντας στον πίνακα 4.13 τις τιμές των φορτίσεων των πρώτων δύο κύριων συνιστωσών, εύκολα διαπιστώνεται ότι η πρώτη κύρια συνιστώσα κατέχει μεγαλύτερες κατά απόλυτη τιμή τις φορτίσεις, οι οποίες προσδιορίζουν τις μεταβλητές hurdles, shot, run200m και longjump, ενώ η δεύτερη κύρια συνιστώσα έχει τις φορτίσεις των υπόλοιπων μεταβλητών πιο αυξημένες, ήτοι των μεταβλητών javelin, run800m και highjump. Η παραπάνω παρατήρηση λανθάνει την ένδειξη ότι η πρώτη κύρια συνιστώσα «δίνει το βάρος της» στις μεταβλητές hurdles, shot, run200m και longjump σε αντίθεση με τη δεύτερη, η οποία φαίνεται να επικεντρώνεται στις μεταβλητές javelin, run800m και highjump.
Έπειτα, μεγάλο ενδιαφέρον έχει η ανάλυση των δεδομένων του δείγματος heptathlon με τη βοήθεια των κύριων συνιστωσών, οι οποίες έχουν επιλεχθεί, και της γραφικής παράστασης biplot.
Οι εντολές στην R έχουν ως εξής:

> pca.res$rotation=-pca.res$rotation
> pca.res$x=-pca.res$x
> biplot (pca.res , scale =0)

Μελετώντας προσεκτικά το γράφημα 4.13, αρχικά, παρατηρείται εμφανώς ότι η γωνία μεταξύ των διανυσμάτων highjump και javelin είναι αρκετά μικρή. Το ίδιο ισχύει και για τις γωνίες των διανυσμάτων hurdles και run200m και των longjump και shot, όπου η τελευταία αυτή γωνία φαίνεται να είναι μηδενική. Εύκολα, λοιπόν, προκύπτει ότι η συσχέτιση των αγωνισμάτων highjump και javelin είναι μεγαλύτερη, εφόσον η γωνία μεταξύ τους είναι μικρή. Με το ίδιο σκεπτικό, τα αγωνίσματα hurdles και run200m έχουν υψηλή συσχέτιση μεταξύ τους, και τέλος τα αγωνίσματα longjump και shot είναι πολύ υψηλά συσχετισμένα μεταξύ τους. Ωστόσο, η γωνία των διανυσμάτων αυτών των τελευταίων αγωνισμάτων (longjump και shot) με τα διανύσματα run800m, hurdles και run200m είναι αρκετά μεγάλη καθιστώντας εμφανή την αρκετά χαμηλή συσχέτιση των μεταβλητών αυτών μεταξύ τους. Ακριβώς στην ίδια περίπτωση ανήκουν και οι μεταβλητές javelin, hurdles και run200m, στις οποίες είναι διακριτή η χαμηλή συσχέτιση, καθώς το ίδιο επίσης ισχύει και για τις μεταβλητές highjump, hurdles και run200m. Εκτός από τα παραπάνω, αποκτάται και η πληροφορία από τη γραφική παράσταση ότι η γωνία των διανυσμάτων hurdles, shot, run200m και longjump με τον άξονα PC1 είναι μικρή σε αντίθεση με τη γωνία, που σχηματίζουν με τον άλλον άξονα PC2. Οπότε, η συσχέτιση,



Σχήμα 4.13: *Διάγραμμα Biplot heptathlon με δύο κύριες συνιστώσες.*

που έχουν οι μεταβλητές αυτές με την πρώτη κύρια συνιστώσα, είναι μεγάλη. Το αντίθετο χαρακτηρίζει τις μεταβλητές javelin, run800m και highjump, οι οποίες είναι υψηλά συσχετισμένες με τη δεύτερη κύρια συνιστώσα. Αυτή η πληροφορία έρχεται, για να επικυρώσει τα σχόλια, που αναφέρθηκαν πιο πάνω για τον πίνακα 4.13.

Και άλλα εξίσου θεμελιώδη συμπεράσματα μπορούν να εξαχθούν από την εν λόγω γραφική παράσταση. Αν κάποιος προσέξει την τιμή, την οποία λαμβάνει η αθλήτρια Joyner-Kersee από τις Η.Π.Α. κατά τον άξονα της πρώτης κύριας συνιστώσας, τότε θα παρατηρήσει ότι αυτή η τιμή είναι η μεγαλύτερη σε αυτόν τον άξονα, σε αντιπαράθεση με τις τιμές, τις οποίες λαμβάνουν οι αθλήτριες Jeong-Mi από τη Νότια Κορέα και Hui-Ing από την Ταϊλάνδη στον

άξονα αυτόν, οι οποίες είναι οι πιο χαμηλές. Με γνώμονα αυτήν την παρατήρηση, καθίσταται εμφανές ότι η αθλήτρια Joyner-Kersee έχει την καλύτερη επίδοση όσον αφορά στα αγωνίσματα, τα οποία εκφράζονται από την πρώτη κύρια συνιστώσα, ήτοι hurdles, shot, run200m και longjump, ενώ οι αθλήτριες Jeong-Mi και Hui-Ing έχουν την χειρότερη επίδοση σε αυτές τις κατηγορίες αγωνισμάτων.
Παράλληλα, οι αθλήτριες Brown από τις Η.Π.Α. και Yuping από την Κίνα είναι αυτές, που έχουν αξιοσημείωτες και καλύτερες επιδόσεις στα αγωνίσματα javelin, run800m και highjump, δεδομένου ότι οι τιμές της δεύτερης κύριας συνιστώσας, οι οποίες σχετίζονται με τα προαναφερόμενα αγωνίσματα, είναι οι μεγαλύτερες στο δείγμα. Εκ διαμέτρου αντίθετα, οι Mulliner από το Ηνωμένο Βασίλειο, Choubenkova από την Πρώην Σοβιετική Ένωση και Hui-Ing έχουν τις χειρότερες επιδόσεις στα αγωνίσματα javelin, run800m και highjump.
Οι αθλήτριες (όπως Hagger ,Wijnsma κ.λ.π.), οι οποίες βρίσκονται κοντά στην τιμή του μηδενός και από τους δύο άξονες, κατέχουν τις μέσες τιμές όλων των μεταβλητών του δείγματος (και των επτά αγωνισμάτων) και επομένως αυτές οι αθλήτριες έχουν τις μέσες επιδόσεις στα αγωνίσματα.

4.2.2 Εφαρμογές Ανάλυσης Συστάδων στο δείγμα δεδομένων heptathlon

Μέθοδος K-Means

Όπως έχει προλεχθεί κατά την εφαρμογή της μεθόδου K-Means στα δεδομένα heptathlon, κρίνεται αναγκαία αρχικά η εύρεση του βέλτιστου αριθμού των συστάδων (k), που δύνανται να περιγράψουν καταλληλότερα το δείγμα. Ως εκ τούτου, πρέπει τα δεδομένα heptathlon να τυποποιηθούν, και έπειτα να βρεθεί ο αριθμός k για το δείγμα αυτό κάνοντας χρήση του πακέτου *factoextra* καθώς και του *cluster* .

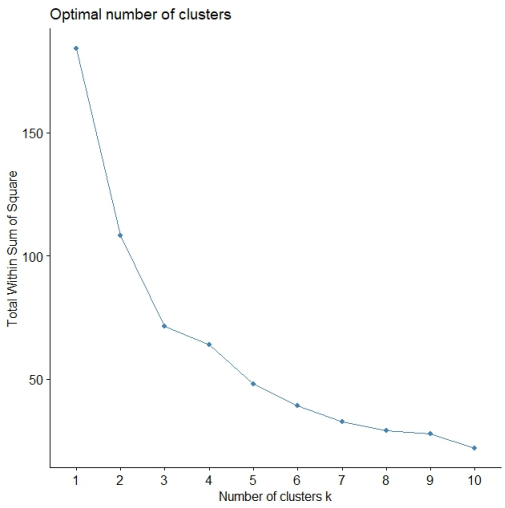
Με αυτόν τον τρόπο, οι εντολές έχουν ως εξής:

> library(factoextra)
> library(cluster)

> heptathlon <- heptathlon[-grep("PNG", rownames(heptathlon)),]
> heptathlonScaling <- heptathlon

> heptathlonScaling <- scale(heptathlonScaling)

> fviz_nbclust(heptathlonScaling, kmeans, method = "wss")
Το output της εντολής εμφανίζεται στο σχήμα 4.14.



Σχήμα 4.14: *Διάγραμμα του αριθμού συστάδων έναντι του συνολικού αθροίσματος τετραγώνων για τα δεδομένα heptathlon.*

Εξετάζοντας το σχήμα 4.14, παρατηρείται ότι ο βέλτιστος αριθμός συστάδων για το δείγμα είναι 3, δεδομένου ότι στο γράφημα εμφανίζεται σε αυτό το σημείο ένας βραχίονας.

Έτσι, εφόσον έχει γίνει η εκτίμηση του κατάλληλου αριθμού των συστάδων, εφαρμόζεται η μέθοδος K-means. Συνεπώς, οι εντολές έχουν ως εξής:

> set.seed(1)
> km.res <- kmeans(heptathlonScaling, centers=3, nstart = 25)
> km.res
K-means clustering with 3 clusters of sizes 4, 14, 6

Clustering vectors:
Within cluster sum of squares by cluster:
12.25929 44.42681 14.84330
(between_SS / total_SS = 61.1%)

Cluster means:

hurdles	highjump	shot	run200m	longjump
-1.3020733	0.2628002	1.38541010	-1.36239460	1.3433854
-0.1769876	0.4880575	0.03880389	-0.07148426	0.1358946
1.2810198	-1.3140008	-1.01414914	1.07505966	-1.2126776

javelin	run800m	score
1.08878230	-1.3389460	1.5414842
0.03301628	0.1933889	0.1041548
-0.80289286	0.4413899	-1.2706841

Joyner-Kersee (USA)	John (GDR)	Behmer (GDR)
1	1	1
Sablovskaite (URS)	Choubenkova (URS)	Schulz (GDR)
2	1	2
Fleming (AUS)	Greiner (USA)	Lajbnerova (CZE)
2	2	2
Bouraga (URS)	Wijnsma (HOL)	Dimitrova (BUL)
2	2	2
Scheider (SWI)	Braun (FRG)	Ruotsalainen (FIN)
2	2	2
Yuping (CHN)	Hagger (GB)	Brown (USA)
2	2	2
Mulliner (GB)	Hautenauve (BEL)	Kytola (FIN)
3	3	3
Geremias (BRA)	Hui-Ing (TAI)	Jeong-Mi (KOR)
3	3	3

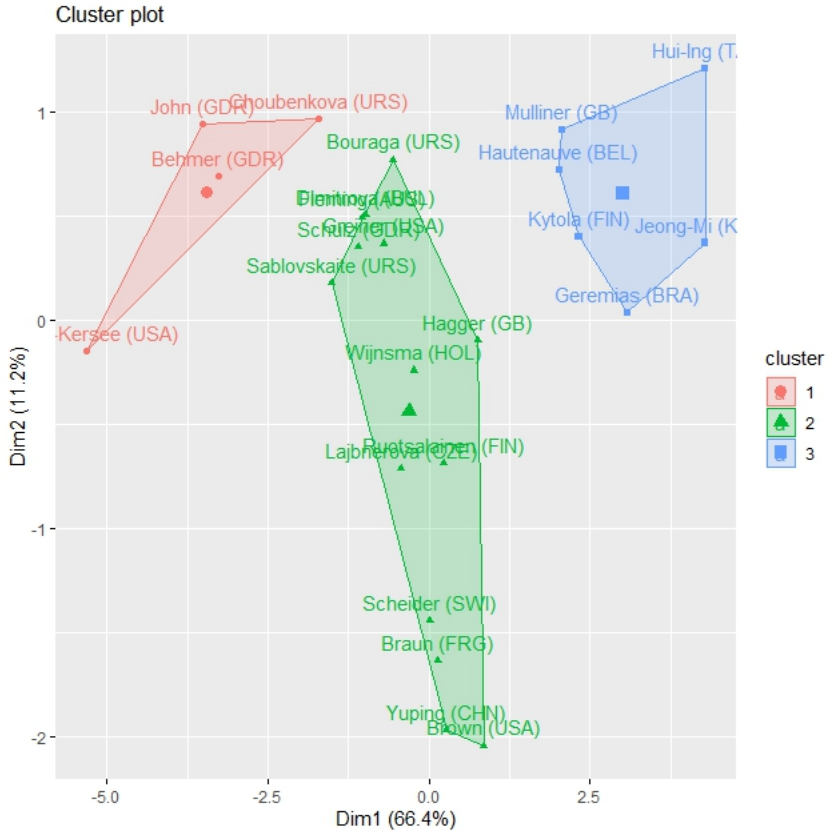
Available components:
"cluster" "centers" "totss" "withinss" "tot.withinss"
"betweenss" "size" "iter" "ifault"

Έτσι, με τη βοήθεια της παραπάνω εντολής παρέχονται οι πληροφορίες

που αφορούν στον αριθμό των αθλητριών, που ανήκουν σε κάθε συστάδα α-
πό τις τρεις, με βάση τις οποίες έχει εκτελεσθεί η μέθοδος, στο άθροισμα

τετραγώνων και στη μέση τιμή κάθε συστάδας. Εκτός από τα παραπάνω, παρέχεται αναλυτικά η αντιστοιχία κάθε αθλήτριας στην εκάστοτε συστάδα. Ως εκ τούτου, θα μπορούσε εύκολα να γίνει γνωστό σε κάποιον ότι οι αθλήτριες Joyner-Kersee, John, Behmer και Choubenkova απαρτίζουν την πρώτη συστάδα, οι Sablovskaite, Schulz, Fleming, Greiner, Lajbnerova, Bouraga, Wijnsma, Dimitrova, Scheider, Braun, Ruotsalainen, Yuping, Hagger και Brown τη δεύτερη και τέλος οι Mulliner, Hautenauve, Kytola, Geremias, Hui-Ing και Jeong-Mi την τρίτη αντίστοιχα. Παράλληλα, οι παραπάνω ισχυρισμοί επιβεβαιώνονται μέσω του γραφήματος 4.15, το οποίο αποκτάται εύκολα από την εντολή:

> *fviz_cluster(km.res, data = heptathlonScaling)*



Σχήμα 4.15: *Διάγραμμα αντιστοιχίας κάθε αθλήτριας σε κάθε συστάδα για τα δεδομένα heptathlon με τη μέθοδο K-means.*

Έπειτα, αν κάποιος θα ήθελε να προχωρήσει σε ανάλυση δεδομένων, που

αφορούν στις επιδόσεις των αθλητριών, που ανήκουν σε κάθε συστάδα ξεχωριστά, τότε θα μπορούσε να επιλέξει την εντολή της R *aggregate(heptathlon, by=list(cluster=km$cluster), mean)* . Με αυτόν τον τρόπο θα ερμήνευε τα αποτελέσματα της εν λόγω εντολής, τα οποία είναι γνωστά από τον πίνακα 4.15. Έτσι, με μία πρώτη ματιά θα μπορούσε εύκολα να καταλήξει στο συμπέρασμα ότι οι αθλήτριες της τρίτης συστάδας έχουν τη χαμηλότερη συνολική επίδοση (score), σε αντίθεση με τις αθλήτριες της πρώτης συστάδας, οι οποίες φαίνονται να έχουν την υψηλότερη συνολική βαθμολογία. Έτσι, οι αθλήτριες Joyner-Kersee, John, Behmer και Choubenkova έχουν κερδίσει τη μέση βαθμολογία 6.897, οι Sablovskaite, Schulz, Fleming, Greiner, Lajbnerova, Bouraga, Wijnsma, Dimitrova, Scheider, Braun, Ruotsalainen, Yuping, Hagger και Brown έχουν λάβει μέσο σκόρ 6.204 και τέλος οι Mulliner, Hautenauve, Kytola, Geremias, Hui-Ing και Jeong-Mi έχουν βαθμολογηθεί κατά μέσο όρο με 5.542. Επιπλέον, οι αθλήτριες, που αποτελούν μέρος της πρώτης συστάδας, φαίνεται να έχουν τις καλύτερες επιδόσεις στα αθλήματα: (1)σφαιροβολία, της οποίας η μέση τιμή τείνει να είναι 15.2 μέτρα, (2) άλμα εις μήκος, στο οποίο οι αθλήτριες κατάφεραν μέσο μήκος, το οποίο ανέρχεται στα 6.7 μέτρα, (3) ακοντισμός, του οποίου η μέση τιμή για τις αθλήτριες αυτές είναι 45.1 μέτρα, (4) στίβος μετ΄ εμποδίων, με μέσο χρόνο τερματισμού 13.1 δευτερόλεπτα, (5) στίβος 200 μέτρων, στον οποίο κατάφεραν να κάνουν μέσο χρόνο 23.3 δευτερόλεπτα και (6) στίβος 800 μέτρων με μέσο χρόνο 126.7 δευτερόλεπτα. Σε ό,τι αφορά τις αθλήτριες της δεύτερης συστάδας, οι επιδόσεις τους ήταν υψηλότερες στο άθλημα άλμα εις ύψος, στο οποίο πέτυχαν άλμα με μέση τιμή 1.82 μέτρα, ενώ στα υπόλοιπα αθλήματα κινήθηκαν σε μέσες επιδόσεις. Τέλος, οι αθλήτριες της τρίτης συστάδας φαίνεται να έχουν σε όλα τα αθλήματα τη χαμηλότερη επίδοση σε σχέση με τις υπόλοιπες αθλήτριες διαφορετικών συστάδων.

Μέθοδος Ιεραρχικής Ανάλυσης

Εκτός από τη μέθοδο K-means, η μέθοδος ιεραρχικής ανάλυσης κρίνεται εξίσου χρήσιμη για την ανάλυση συστάδων ενός δείγματος με μεγάλο όγκο δεδομένων. Για αυτόν το λόγο, υπολογίζεται πρώτα η διαφορετικότητα των παρατηρήσεων μέσω της απόστασης Manhattan. Η εντολή στην R είναι η εξής:
> dhept <- dist(heptathlonScaling, method = "manhattan")

Έπειτα, γίνεται η ζεύξη των παρατηρήσεων με την ολοκληρωμένη, μέση και μονή ζεύξη, όπως εφαρμόστηκε και στα δεδομένα USArrests, ώστε να γίνει η

cluster	hurdles	highjump	shot	run200m
1	13.06250	1.807500	15.24750	23.31000
2	13.64143	1.819286	13.23143	24.51929
3	14.39167	1.725000	11.65500	25.59333

longjump	javelin	run800m	score
6.745000	45.05500	126.6825	6896.500
6.260000	41.39286	136.1021	6204.286
5.718333	38.49333	137.6267	5542.167

Πίνακας 4.15: *Output εντολής aggregate(heptathlon,by=list(cluster=km$cluster),mean), όπου εμφανίζονται οι μέσες τιμές των μεταβλητών για κάθε συστάδα, στις οποίες ανήκουν οι αθλήτριες.*

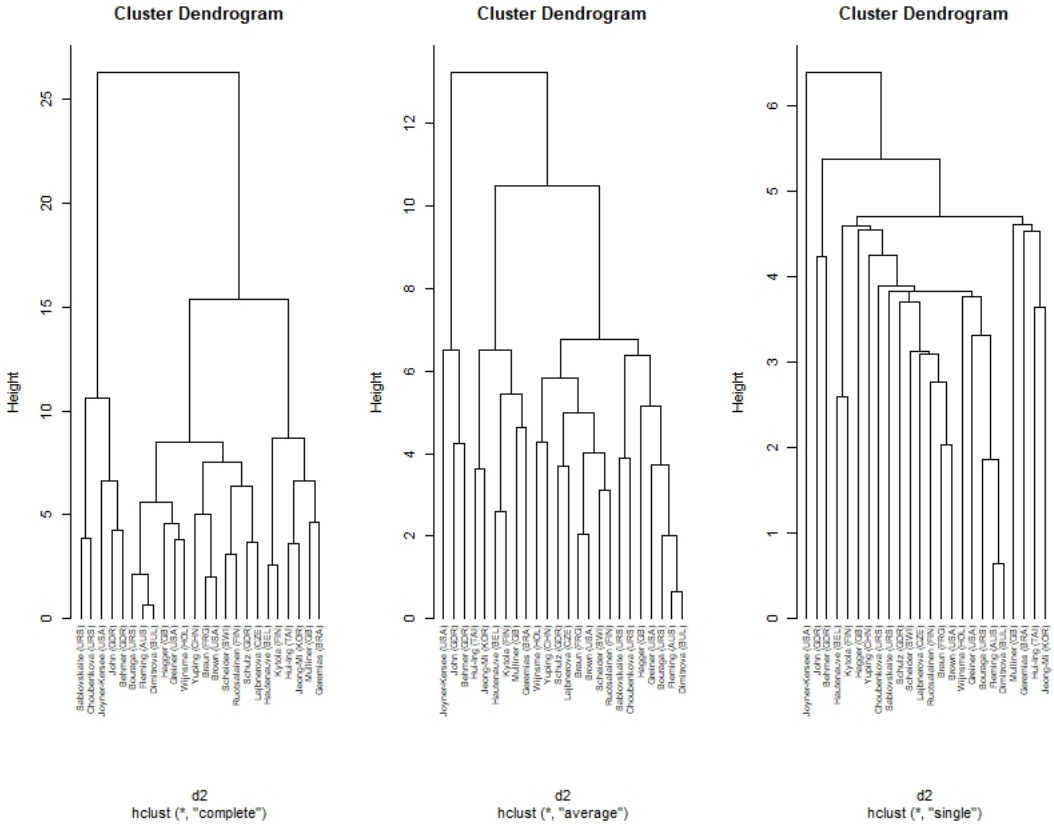
επιλογή της καταλληλότερης μεθόδου για το δείγμα heptathlon. Έτσι, ακολουθείται η παρακάτω διαδικασία:

```
> HC1 <- hclust(dhept, method = "complete" )
> HC2 <- hclust(dhept, method = "average" )
> HC3 <- hclust(dhept, method = "single" )
```

Με αυτόν τον τρόπο, έχοντας υπολογίσει τη διαφορετικότητα των παρατηρήσεων και έχοντάς τες συνδέσει, πραγματοποιείται η κατασκευή των δενδρογραμμάτων.

```
> par(mfrow = c(1, 3))
> plot(HC1, cex = 0.6, hang = -1)
> plot(HC2, cex = 0.6, hang = -1)
> plot(HC3, cex = 0.6, hang = -1)
```

Το output των παραπάνω εντολών εμφανίζεται στο σχήμα 4.16. Παρατηρώντας τα δενδρογράμματα γίνεται ορατή μία αρκετά μεγάλη ομοιότητα ανάμεσα στην ολοκληρωμένη και στη μέση ζεύξη, καθώς και θα ήταν αποδεκτή η ερμηνεία ότι και οι τρεις τρόποι ζεύξεις υποδεικνύουν μέσω των δενδρογραμμάτων τους ότι τρεις συστάδες δύνανται να περιγράψουν κατάλληλα το δείγμα heptathlon. Έτσι, αν κάποιος ήθελε να του γίνουν γνωστές οι αθλήτριες, οι οποίες έχουν ταξινομηθεί σε κάθε συστάδα με τη λόγου χάρη ολοκληρωμένη ζεύξη, τότε θυμίζεται ότι η εντολή, η οποία θα εξυπηρετούσε την ανάγκη αυτή είναι η παρακάτω:



Σχήμα 4.16: *Δενδρογράμματα των δεδομένων heptathlon με ολοκληρωμένη, μέση και μονή ζεύξη.*

> cutree(HC1, k = 3)

Η ταξινόμηση φαίνεται στον πίνακα 4.16. Είναι πασιφανές, λοιπόν, ότι οι αθλήτριες Joyner-Kersee, John, Behmer, Sablovskaite και Choubenkova απαρτίζουν την πρώτη συστάδα, οι Schulz, Fleming, Greiner, Lajbnerova, Bouraga, Wijnsma, Dimitrova, Scheider, Braun, Ruotsalainen, Yuping, Hagger και Brown τη δεύτερη και τέλος οι Mulliner, Hautenauve, Kytola, Geremias, Hui-Ing και Jeong-Mi την τρίτη. Παράλληλα, αξίζει να σημειωθεί ότι αυτή η ταξινόμηση είναι σε συμφωνία με αυτήν της μεθόδου K-means, η οποία εφαρμόστηκε προηγουμένως, με τη μόνη διαφορά ότι η αθλήτρια Sablovskaite ανήκε στη δεύτερη συστάδα με βάση τη μέθοδο K-means.

Joyner-Kersee (USA)	John (GDR)	Behmer (GDR)
1	1	1
Sablovskaite (URS)	Choubenkova (URS)	Schulz (GDR)
1	1	2
Fleming (AUS)	Greiner (USA)	Lajbnerova (CZE)
2	2	2
Bouraga (URS)	Wijnsma (HOL)	Dimitrova (BUL)
2	2	2
Scheider (SWI)	Braun (FRG)	Ruotsalainen (FIN)
2	2	2
Yuping (CHN)	Hagger (GB)	Brown (USA)
2	2	2
Mulliner (GB)	Hautenauve (BEL)	Kytola (FIN)
3	3	3
Geremias (BRA)	Hui-Ing (TAI)	Jeong-Mi (KOR)
3	3	3

Πίνακας 4.16: *Output της εντολής cutree(HC1, k = 3)* .

Επιπρόσθετα, εάν πραγματοποιηθεί η εντολή *cutree* για τη μέση ζεύξη των παρατηρήσεων, τότε θα γινόταν αντιληπτό ότι η μόνη απόκλιση σε ό,τι αφορά την ταξινόμηση των αθλητριών με αυτή με την ολοκληρωμένη ζεύξη, θα ήταν ότι οι αθλήτριες Sablovskaite και Choubenkova είναι στη δεύτερη συστάδα και όχι στην πρώτη. Ακόμα, η μονή ζεύξη θα προκαλούσε σημαντική ανομοιότητα στην ταξινόμηση των αθλητριών σε συστάδες, μιας και η Joyner-Kersee θα αποτελούσε μόνη της μία συστάδα, οι John και Behmer θα αποτελούσαν τη δεύτερη, ενώ όλες οι υπόλοιπες αθλήτριες θα αποτελούσαν την τρίτη συστάδα. Έτσι, αυτό το γεγονός φαίνεται να είναι σε αντιδιαστολή με τα συμπεράσματα της μέσης και της ολοκληρωμένης ζεύξης.

4.3 Συμπεράσματα

Έχοντας εφαρμόσει τις δύο μεθόδους, ήτοι την Ανάλυση Κύριων Συνιστωσών καθώς και την Ανάλυση Συστάδων, σε δύο διαφορετικά δείγματα δεδομένων (USArrests και heptathlon), κρίνεται σκόπιμη η εξαγωγή ορισμένων

γενικών συμπερασμάτων ως προς τη χρησιμότητα και τους «καρπούς», τους οποίους απέφερε η κάθε μία ξεχωριστά.

Αρχικά, αναφορικά με την εφαρμογή της Ανάλυσης Κύριων Συνιστωσών στα προαναφερθέντα δεδομένα, πραγματοποίειται η μείωση των μεταβλητών (συνήθως συσχετισμένες) του κάθε δείγματος με σκοπό την καλύτερη περιγραφή του δείγματος. Έτσι, στο δείγμα USArrests φάνηκε ότι δύο μεταβλητές - συνιστώσες, οι οποίες είναι ασυσχέτιστες, δύνανται να περιγράψουν ένα αρκετά υψηλό ποσοστό (87%) της μεταβλητότητας του αρχικού δείγματος. Ειδικότερα, βρέθηκε ότι η πρώτη κύρια συνιστώσα δίνει το βάρος της στις μεταβλητές του αρχικού δείγματος Murder, Assault και Rape, σε αντίθεση με τη δεύτερη, η οποία εκφράζει περισσότερο τη μεταβλητή UrbanPop. Επιπλέον, με εργαλείο το διάγραμμα Biplot έγινε εφικτός ένας πρώτος διαχωρισμός των πολιτειών ως προς τις δύο συνιστώσες. Συγκεκριμένα, έγινε ορατό ποιά πολιτεία έχει υψηλή ή χαμηλή τιμή στην πρώτη κύρια συνιστώσα, και άρα στην εγκληματικότητα, και ποιά έχει υψηλή ή χαμηλή τιμή στη δεύτερη κύρια συνιστώσα, και άρα σε ποιά διαμένει μεγάλος ή μικρός αριθμός πολιτών. Αντίστοιχα, η ίδια διαδικασία ακολουθήθηκε και στο δείγμα heptathlon, στο οποίο έγινε αντιληπτό ότι δύο συνιστώσες εκφράζουν υψηλό ποσοστό διασποράς του αρχικού δείγματος. Μάλιστα, η πρώτη κύρια συνιστώσα περιλαμβάνει τις μεταβλητές- αθλήματα hurdles, shot, run200m και longjump, ενώ η δεύτερη επικεντρώνεται στις μεταβλητές javelin, run800m και highjump. Και σε αυτήν την περίπτωση, εύκολα έγινε ο διαχωρισμός της κάθε αθλήτριας σε ποιά συνιστώσα (αθλήματα) έχει υψηλότερη ή χαμηλότερη επίδοση.

Όσον αφορά στα αποτελέσματα της χρήσης της Ανάλυσης Συστάδων στα δεδομένα, ομαδοποιούνται οι παρατηρήσεις του κάθε δείγματος με βάση τη διαφορετικότητα τους, με στόχο την εύρεση κοινών γνωρισμάτων ανάμεσα στις παρατηρήσεις. Αναλυτικότερα, για το δείγμα USArrests αξιοποιώντας τη μη ιεραρχική μέθοδο K-means καθορίστηκε εκ των προτέρων ο αριθμός συστάδων, οι οποίες δύνανται να εμπεριέχουν τις πολιτείες, στην τιμή 4. Ως εκ τούτου, οπτικοποιείται σε ποιά συστάδα από τις τέσσερεις ανήκει η κάθε πολιτεία του δείγματος, και συνεπώς γίνεται εύλογο το συμπέρασμα ποιές πολιτείες εμφανίζουν κοινά χαρακτηριστικά ως προς την εγκληματικότητα και το πλήθος των πολιτών, που κατοικούν σε αυτές. Επιπρόσθετα, με τον ίδιο τρόπο εφαρμόσθηκε η ανάλυση στο δείγμα heptathlon, όπου επιλέχθηκαν a priori 3 συστάδες, και επομένως γίνονται γνωστές οι αποκλίσεις και οι συγκλίσεις μεταξύ των επιδόσεων των αθλητριών. Παράλληλα, κάνοντας χρήση της ιεραρχικής μεθόδου

στο δείγμα USArrests, υπολογίζοντας τη διαφορετικότητα των πολιτειών με τη βοήθεια της Ευκλείδειας απόστασης και έχοντας επιλέξει την ολοκληρωμένη ή τη μέση ζεύξη προκύπτουν τα δενδρογράμματα, από τα οποία αποκτώνται οι κύριες συστάδες, καθώς και οι υποομάδες. Συγκεκριμένα, γίνεται εμφανές ότι με την ολοκληρωμένη ζεύξη οι κύριες συστάδες είναι τέσσερεις, και κατά συνέπεια πραγματοποιείται η ομαδοποίηση των πολιτειών σε 4 συστάδες. Κατά αντιστοιχία, όταν εφαρμόζεται η μέση ζεύξη κρίνεται ότι τρεις συστάδες μπορούν να περιλάβουν τις πολιτείες. Επίσης, παρατηρώντας τα φύλλα του δενδρογράμματος λαμβάνεται η πληροφορία ποιές πολιτείες εμφανίζουν τις περισσότερες ομοιότητες μεταξύ τους. Παρόμοια μεθοδολογία πραγματοποιήθηκε και στο δείγμα heptathlon με τη συνδρομή της απόστασης Μανχάταν και της ολοκληρωμένης και μέσης ζεύξης, όπου κάθε αθλήτρια ταξινομήθηκε σε μία από τις τρεις συστάδες με βάση τις παρόμοιες επιδόσεις στα αθλήματα του επτάθλου.

Λαμβάνοντας όλα τα παραπάνω υπ΄ όψιν, θεωρείται ότι οι δύο τεχνικές (Ανάλυση Κύριων Συνιστωσών και Ανάλυση Συστάδων) έχουν συμπληρωματική δράση και όχι ανταγωνιστική. Επιπρόσθετα, γίνεται προφανές ότι η επιλογή της ιεραρχικής ή της K-means μεθόδου, όσον αφορά στην Ανάλυση Συστάδων, καθώς και του τρόπου ζεύξης (μονή, μέση, ολοκληρωμένη) εξαρτάται από το αντικείμενο της μελέτης. Εκτός από τα παραπάνω, καθίσταται πασιφανές ότι η ερμηνεία των αποτελεσμάτων των δύο τεχνικών εναπόκειται σε μεγάλο βαθμό στην κρίση και στην υποκειμενικότητα του στατιστικού- ερευνητή.

Βιβλιογραφία

Anderson, E. The irises of the Gaspe peninsula. *Bulletin of the American Iris Society*, **59**: 2–5, 1935.

Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39B**: 188, 1977.

Dubes, R. C., Jain, A. K. *Algorithms for Clustering Data*. Prentice Hall, New Jersey, 1988.

Everitt, B., Hothorn, T. *A Handbook of Statistical Analysis Using R*. 2nd Edition, Taylor and Francis Group, LLC, New York, 2010.

Everitt, B., Hothorn, T. *An Introduction to Applied Multivariate Analysis with R*. Springer, New York, 2011.

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**: 179–188, 1936.

Flynn, P.J., Jain, A.K., Murty, M.N. Data clustering: a review. *ACM Computing Surveys*, **31**: 264–323, 1999.

Fraser, W.R., Gorman, K.B., Williams, T.D. Ecological Sexual Dismorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis), *Plus One* , 2014.

Hastie, T., James, G., Tibshirani, R., Witten, D. *An Introduction to Statistical Learning with Applications in R*. Springer, New York,

2013.

Hastie, T., James, G., Tibshirani, R., Witten, D. *An Introduction to Statistical Learning with Applications in R*. 2nd Edition, Springer, New York, 2021.

Hastie, T., Mazdumar R., Tibshirani, R. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of Machine Learning Research*, **11**: 2287-2322, 2010.

Hochreiter, S. *Basic Methods of Data Analysis*. Institute of Bioinformatics, Johannes Kepler University Linz, Austria, 2014.

Husson, F., Josse, J. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software*, **70(1)**: 1-31, 2016.

Jolliffe, I.T. *Principal Component Analysis*. 2nd Edition, Springer, New York, 2002.

Jolliffe, I.T., Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, **374**: 20150202, 2016.

McNeil, D.R. *Interactive Data Analysis*. Wiley, New York, 1977.

Sharma, S. *Applied Multivariate Techniques*. John Wiley and Sons Inc., New York, 1996.

Ευρετήριο