



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ



ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ – ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

*Πρόγραμμα Μεταπτυχιακών Σπουδών (ΠΜΣ)
«Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και
Κλινική Βιοπληροφορική»*

Διπλωματική Εργασία

**«Σύγκριση της Γραμμικής Διαχωριστικής Ανάλυσης, της Ανάλυσης
Κύριων Συνιστωσών και της Ανάλυσης Συστάδων»,**

**«Comparison of Linear Discriminant Analysis (LDA), Principal
Component Analysis (PCA) and Cluster Analysis (CA)»**

Κωνσταντίνος Παράσχου

Μπατσίδης Απόστολος,

Επίκουρος Καθηγητής του Τμήματος Μαθηματικών του
Πανεπιστημίου Ιωαννίνων (Επιβλέπων)

Στεφανίδης Ιωάννης,

Καθηγητής του Τμήματος Ιατρικής του Πανεπιστημίου
Θεσσαλίας

Δοξάνη Χρυσούλα,

Επιστημονικός Συνεργάτης του Τμήματος Ιατρικής του
Πανεπιστημίου Θεσσαλίας

ΛΑΡΙΣΑ, ΣΕΠΤΕΜΒΡΙΟΣ 2019

A. ΠΕΡΙΛΗΨΗ

- **Εισαγωγή**

Η πολυμεταβλητή στατιστική ανάλυση αποτελεί ένα πολύτιμο εργαλείο σε πολλές επιστήμες. Οι μέθοδοι της πολυμεταβλητής ανάλυσης αφορούν διαδικασίες και μεθοδολογίες που χρησιμοποιούνται για τη συλλογή, περιγραφή και ανάλυση δεδομένων και εν συνεχεία την εξαγωγή συμπερασμάτων, με τη χρήση πολλών μεταβλητών σε ένα πλήθος ατόμων ή πειραματικών μονάδων. Υπάρχουν πολλές τέτοιες μέθοδοι στη διάθεσή μας, με βάση τις υποθέσεις που γίνονται για την φύση των δεδομένων και τον τύπο των συσχετίσεων που επιθυμούμε να διερευνήσουμε.

- **Στόχοι**

Τρεις ευρέως διαδεδομένες, σε πρακτικές εφαρμογές, τεχνικές της πολυδιάστατης ανάλυσης είναι η Γραμμική Διαχωριστική Ανάλυση (Linear Discriminant Analysis, LDA), η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis, PCA) και η Ανάλυση Συστάδων (Cluster Analysis, CA). Σκοπός της παρούσας Μεταπτυχιακής Διπλωματικής Εργασίας είναι να περιγραφεί και να παρουσιασθεί συνοπτικά καθεμιά από τις προαναφερθείσες μεθοδολογίες και να επισημανθούν οι διαφοροποιήσεις τους και οι όποιες ομοιότητες μέσα από την εφαρμογή τους σε ένα σύνολο δεδομένων.

- **Μέθοδοι**

Αρχικά, θα παρατεθούν βασικά στοιχεία που αφορούν την θεωρία και την μεθοδολογία της πολυμεταβλητής στατιστικής ανάλυσης γενικά, και των τριών υπό μελέτη μεθόδων ειδικότερα. Κατόπιν, με τη βοήθεια κατάλληλων πολυμεταβλητών δεδομένων, θα γίνει εφαρμογή των παραπάνω μεθόδων, κάνοντας χρήση του προγράμματος IBM SPSS Statistics 25®.

- **Αποτελέσματα**

Τα αποτελέσματα που θα προκύψουν θα αναλυθούν εκτενώς, προκειμένου να αποσαφηνιστεί η διαδικασία επεξεργασίας των δεδομένων και να ερμηνευτούν οι νέες συσχετίσεις που προκύπτουν μεταξύ των μεταβλητών.

- **Συμπέρασμα**

Οι μέθοδοι αυτές εφαρμόζονται με διαφορετικό στόχο, αναλύουν με διαφορετικό τρόπο τα υπό θεώρηση δεδομένα και παράγουν αποτελέσματα που απαντούν σε ετερογενή ερευνητικά ερωτήματα. Η ανωτέρω ετερογένεια, θα βοηθήσει να αναδειχθούν τα πλεονεκτήματα και μειονεκτήματα της εφαρμογής των υπό μελέτη μεθόδων ανάλυσης, ενώ με την μεταξύ τους σύγκριση, θα αποσαφηνιστούν οι ενδείξεις χρήσης τους.

A. ABSTRACT

- **Introduction**

Multivariate statistics is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one outcome variable. The application of multivariate statistics is known as multivariate statistical analysis. Multivariate statistical analysis is a very useful tool for many scientific fields, related to the collection, description and analysis of data and drawing conclusions from the multivariate data set extracted from the sample population. There are many multivariate techniques to choose from, based on the nature of the data and the different aims and background of each of them.

- **Aim**

Three widely known multivariate statistical methods, with numerous practical applications, are the Linear Discriminant Analysis (LDA), the Principal Component Analysis (PCA) and the Cluster Analysis (CA). This dissertation aims to describe and to briefly present these multivariate techniques, as well as to point out their differences and similarities, by applying them to a relative dataset.

- **Methods**

To begin with, we will present some key features regarding the theory and the methodology of multivariate statistical analysis in general, and the three aforementioned techniques respectively. Moreover, a medical multivariate dataset will be used to apply these methods and the analysis of the data will be performed with the help of IBM SPSS Statistics 25® program.

- **Results**

The results will be analyzed extensively, in order to clarify the procedure of data processing and to interpret the results.

- **Conclusion**

These methods (LDA, PCA, CA) are used with different goals, they analyze the data differently and the results that they produce give answers to different research hypotheses. The heterogeneity of these techniques, will help us to highlight the advantages and disadvantages of each technique, whereas through comparison with each other, we will clarify the indications based on which they can be used.

B. ΕΙΣΑΓΩΓΗ

Η πολυμεταβλητή στατιστική ανάλυση αφορά διαδικασίες και μεθοδολογίες που χρησιμοποιούνται για την επεξεργασία δεδομένων που περιλαμβάνουν μεγάλο αριθμό μεταβλητών. Αποτελεί ένα πολύτιμο εργαλείο σε πολλές επιστήμες. Αν και αναπτύχθηκε ήδη από το πρώτο μισό του 20ου αιώνα, τις τελευταίες δεκαετίες, η ραγδαία εξέλιξη της τεχνολογίας και των υπολογιστικών προγραμμάτων και συστημάτων, αλλά και η ανάγκη για ταχεία αύξηση της έρευνας, ανέδειξαν την πολυμεταβλητή στατιστική ανάλυση στην αποτελεσματικότερη μέθοδο επίλυσης πρακτικών προβλημάτων σε τομείς όπως η γεωλογία, η μετεωρολογία, η ιατρική, η βιομηχανία, η ψυχολογία, η οικονομία και η γεωπονία.

Οι μέθοδοι της πολυμεταβλητής ανάλυσης επιτυγχάνουν τόσο τη συνοπτική παρουσίαση των δεδομένων όσο και την εύρεση δομών στα δεδομένα και συσχετίσεων μεταξύ των μεταβλητών. Αυτό μας δίνει μια πιο πλούσια και ρεαλιστική εικόνα, από το να εξετάζεται μία μόνο μεταβλητή. Υπάρχουν πολλές τέτοιες μέθοδοι στη διάθεσή μας, με βάση τις υποθέσεις που γίνονται για την φύση των δεδομένων και τον στόχο της μελέτης. Οι τεχνικές αυτές είναι πολύπλοκες και περιλαμβάνουν υψηλού επιπέδου μαθηματικά, που συνήθως απαιτούν την χρήση στατιστικών προγραμμάτων για την υλοποίησή τους και την ανάλυση των δεδομένων. Ακόμη, τα αποτελέσματα που μας δίνουν δεν είναι πάντα εύκολο να ερμηνευτούν και τείνουν να βασίζονται σε υποθέσεις που είναι δύσκολο να αξιολογηθούν, εξεταστούν.

Η επιλογή της κατάλληλης μεθόδου βασίζεται σε δύο βασικά βήματα:

1. Να επιλέξουμε την μέθοδο που ανταποκρίνεται στο ερευνητικό ερώτημα που έχουμε θέσει και στα δεδομένα που έχουμε συλλέξει.
2. Να έχουμε κατά νου τα πλεονεκτήματα, μειονεκτήματα και περιορισμούς υπό τα οποία εφαρμόζονται οι διάφορες μέθοδοι, από τον σχεδιασμό ακόμα της μελέτης, προκειμένου να συγκεντρώσουμε τα κατάλληλα δεδομένα, με τον ενδεδειγμένο τρόπο.

Στην παρούσα εργασία θα περιγραφούν οι βασικές αρχές και εν συνεχεία θα συγκριθούν τρεις μόνο από τις μεθόδους πολυμεταβλητής ανάλυσης, που έχουν ευρύτερη εφαρμογή σε πολλούς επιστημονικούς τομείς: η **Γραμμική Διαχωριστική Ανάλυση** (Linear Discriminant Analysis, LDA), η **Ανάλυση Κύριων Συνιστωσών** (Principal Component Analysis, PCA) και η **Ανάλυση Συστάδων** (Cluster Analysis, CA).

Γ. ΜΕΘΟΔΟΙ

• Πολυμεταβλητή Στατιστική Ανάλυση – Multivariate Statistical Analysis

Κρίνεται σκόπιμο, να γίνουν εξ αρχής ορισμένες τεχνικές επισημάνσεις, οι οποίες είναι πολύ σημαντικές για την κατανόηση των μεθόδων που θα παρουσιαστούν. Στην πολυμεταβλητή ανάλυση οι παρατηρήσεις βρίσκονται σε έναν πολυδιάστατο χώρο μεταβλητών, για παράδειγμα, αν p είναι ο αριθμός των μεταβλητών, σε p -διάστατο χώρο^[1].

Υπάρχουν δύο ειδών απεικονίσεις:

- Τα διαγράμματα διασποράς παρουσιάζουν τις παρατηρήσεις ως σημεία στον p -διάστατο χώρο των μεταβλητών και αυτή είναι η συνηθέστερη απεικόνιση^[1].
- Η απεικόνιση των μεταβλητών στον χώρο των παρατηρήσεων, όπου οι μεταβλητές ορίζονται ως διανύσματα (vectors) στο n -διάστατο χώρο των παρατηρήσεων, αν n είναι ο αριθμός των παρατηρήσεων. Στην περίπτωση αυτή, η απεικόνιση γίνεται λαμβάνοντας υπόψη τις στήλες στη μήτρα δεδομένων, μια μεταβλητή δηλαδή ορίζεται στον διανυσματικό χώρο από τις τιμές όλων των παρατηρήσεων για αυτή. Αυτού του είδους η απεικόνιση είναι πολύ χρήσιμη στην παραγοντική ανάλυση (O'Sullivan & Unwin, 2003)^[1].
- Τέλος, η απόσταση η οποία αποτελεί πολύ σημαντική έννοια στην ανάλυση των συστάδων, αλλά και στη διαχωριστική ανάλυση, είναι η στατιστική απόσταση μεταξύ των παρατηρήσεων. Οι συντεταγμένες δηλαδή των παρατηρήσεων, που χρησιμοποιούνται για να υπολογιστεί η απόσταση μεταξύ τους στον πολυδιάστατο χώρο, είναι οι τιμές των μεταβλητών για κάθε παρατήρηση^[1].

• Γραμμική Διαχωριστική Ανάλυση – Linear Discriminant Analysis (LDA)

Η Γραμμική Διαχωριστική ανάλυση αποτελεί γενίκευση του διαχωριστικού κανόνα του Fisher και είναι η μέθοδος εκείνη που χρησιμοποιείται στη στατιστική για την εύρεση ενός γραμμικού συνδυασμού των χαρακτηριστικών που διαχωρίζουν δύο ή περισσότερες ομάδες αντικειμένων. Ο προκύπτων συνδυασμός μπορεί να χρησιμοποιηθεί ως ένας γραμμικός ταξινομητής ή, πιο συχνά, για τη μείωση των διαστάσεων πριν από την μεταγενέστερη ταξινόμηση. Στο πλαίσιο αυτό, η Γραμμική Διαχωριστική Ανάλυση όχι μόνο κατατάσσει τις παρατηρήσεις σε ομάδες, αλλά προσδιορίζει και την πιθανότητα με την οποία γίνεται η κατάταξη αυτή, υπό την προϋπόθεση ότι τα δεδομένα ακολουθούν πολυμεταβλητή κανονική κατανομή. Βρίσκει ευρεία εφαρμογή μεταξύ άλλων στους τομείς της Βιομετρίας, της Βιοπληροφορικής και της Χημείας.

Η Γραμμική Διαχωριστική Ανάλυση, ως στατιστική τεχνική ταξινόμησης παρατηρήσεων, η οποία όμως προϋποθέτει τον εκ των προτέρων διαχωρισμό των δεδομένων σε δύο ή περισσότερες ομάδες, μας επιτρέπει να μελετήσουμε τις διαφορές μεταξύ των ομάδων αυτών λαμβάνοντας υπόψη ένα σύνολο μεταβλητών που περιγράφουν τα χαρακτηριστικά των ομάδων^[1]. Με βάση το παραπάνω, προκύπτει ότι η Γραμμική Διαχωριστική Ανάλυση σχετίζεται με την Ανάλυση Διακύμανσης και την Παλινδρόμηση, οι οποίες επίσης προσπαθούν να εκφράσουν την εξαρτημένη μεταβλητή ως γραμμικό συνδυασμό των άλλων χαρακτηριστικών ή μετρήσεων. Όμως, η Ανάλυση Διακύμανσης έχει ως εξαρτημένη μεταβλητή, ποσοτική μεταβλητή, και ως ανεξάρτητες, κατηγορικές, ενώ η Γραμμική Διαχωριστική Ανάλυση έχει ποσοτικές ανεξάρτητες μεταβλητές και μια

κατηγορική μεταβλητή που δηλώνεται η ομάδα (class label). Επίσης, η Γραμμική Διαχωριστική Ανάλυση σχετίζεται με την Ανάλυση Κύριων Συνιστωσών (θα περιγραφεί αναλυτικότερα παρακάτω), καθώς και οι δύο τεχνικές αναζητούν γραμμικούς συνδυασμούς των δεδομένων που εξηγούν με τον καλύτερο τρόπο τα δεδομένα. Στο πλαίσιο αυτό, η Γραμμική Διαχωριστική Ανάλυση προσπαθεί να μοντελοποιήσει με τον καλύτερο τρόπο τις διαφορές μεταξύ των ομάδων, ενώ η Ανάλυση Κύριων Συνιστωσών δε λαμβάνει υπόψη καμία διαφοροποίηση μεταξύ ομάδων. Τέλος, η Γραμμική Διαχωριστική Ανάλυση χρησιμοποιείται όταν οι ομάδες είναι γνωστές εκ των προτέρων, ενώ η Ανάλυση κατά Συστάδες (όπως θα αναφερθεί αναλυτικά παρακάτω) προσπαθεί να «εντοπίσει» ομάδες πειραματικών μονάδων.

Οι βασικές προϋποθέσεις που πρέπει να τηρούνται για την εφαρμογή της μεθόδου είναι οι ακόλουθες:

- i. Οι μεταβλητές να ακολουθούν την πολυμεταβλητή κανονική κατανομή. Σε αντίθετη περίπτωση το δείγμα πρέπει να είναι μεγάλο ώστε να είναι ανθεκτική (robust) η διαδικασία.
- ii. Να υπάρχει ομοιογένεια των διακυμάνσεων (homogeneity of variances).
- iii. Να αποφεύγεται η πολυσυγγραμμικότητα (multicollinearity).
- iv. Καμία μεταβλητή δεν πρέπει να είναι γραμμικός συνδυασμός των υπολοίπων διακριτικών μεταβλητών.

Γενικά, η μέθοδος εφαρμόζεται χωρίς ιδιαίτερους ελέγχους αρκεί να μην υπάρχουν ακραίες παρατηρήσεις. Στην περίπτωση που τα δεδομένα δεν ικανοποιούν αυτές τις βασικές υποθέσεις, τα αποτελέσματα δεν απεικονίζουν επακριβώς την πραγματικότητα και πρέπει να γίνονται αποδεκτά με επιφύλαξη^[1]. Επισημαίνεται ότι δεν υπάρχει όριο για τον αριθμό των μεταβλητών, αλλά ο αριθμός των παρατηρήσεων πρέπει να είναι τουλάχιστον δύο ανά ομάδα και γενικά να υπερβαίνει τον αριθμό των μεταβλητών. Τέλος, πρέπει να σημειώσουμε ότι παρόλο που η διακριτική ανάλυση συνδέει δύο ή περισσότερες ομάδες παρατηρήσεων με ένα σύνολο διακριτικών μεταβλητών, δεν περιέχει την έννοια της αιτιότητας^[1].

Αναλυτικότερα ως προς τη μέθοδο ανάλυσης, δημιουργούνται γραμμικοί συνδυασμοί των μεταβλητών, της μορφής

$$Z_1 = \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n,$$

όπου 'α' οι διακριτικοί συντελεστές (coefficients) και 'X' οι αρχικές μεταβλητές, οι οποίοι διαχωρίζουν τα δεδομένα στις ομάδες και μπορούμε να αξιολογήσουμε τη σπουδαιότητα κάθε μεταβλητής στον διαχωρισμό των ομάδων^[11].

Οι γραμμικοί συνδυασμοί που προκύπτουν ονομάζονται κανονικές διαχωριστικές συναρτήσεις (canonical discriminant functions) ή κανονικές μεταβλητές (canonical variables) και ο μέγιστος αριθμός τους ισούται με τον αριθμό των ομάδων μείον ένα ή με τον αριθμό των μεταβλητών, στην περίπτωση που αυτός είναι μικρότερος από τον αριθμό των ομάδων.

Ακόμη, ισχύει ότι η πρώτη κανονική μεταβλητή Z_1 , δίνει το μέγιστο F-ratio (between-class variance/within-class variance), καθώς και ότι οι προκύπτουσες κανονικές μεταβλητές (Z_1, Z_2, \dots, Z_{v-1}) είναι ασυσχέτιστες μεταξύ των ομάδων. Τέλος, κρίνεται αναγκαίο να αναφερθεί ότι οι διακριτικές τιμές (discriminant scores) που προκύπτουν από τις διαχωριστικές συναρτήσεις συνήθως υπολογίζονται σε τυπικές τιμές, οπότε η διακριτική

τιμή μιας παρατήρησης αντιπροσωπεύει τον αριθμό των τυπικών αποκλίσεων κατά τον οποίο η παρατήρηση απέχει από το κέντρο της ομάδας^[1].

- **Ανάλυση Κύριων Συνιστωσών - Principal Component Analysis (PCA)**

Η Ανάλυση Κύριων Συνιστωσών αποτελεί μία από τις πρώτες τεχνικές μείωσης διαστάσεων δεδομένων, αφού περιγράφηκε ήδη από τις αρχές του 1900 από τον K. Pearson. Το 1933, ο Hotelling περιέγραψε πρακτικές υπολογιστικές μεθόδους, καθιερώνοντας τον όρο «κύριες συνιστώσες»^[5].

Η Ανάλυση Κύριων Συνιστωσών αποτελεί την απλούστερη και πλέον διαδεδομένη πολυμεταβλητή ανάλυση και στοχεύει στην αντικατάσταση των αρχικά εμπλεκόμενων p το πλήθος μεταβλητών με νέες μεταβλητές, τις κύριες συνιστώσες. Οι νέες μεταβλητές έχουν την ιδιότητα να είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και παράλληλα να μη συσχετίζονται μεταξύ τους^[4]. Το μεγάλο πλεονέκτημά των κύριων συνιστωσών έγκειται στην ιδιαιτερότητα που διαθέτουν, λόγω του τρόπου κατασκευής τους, να μη συσχετίζονται μεταξύ τους.

Όπως επισημαίνεται, σε πολλές περιπτώσεις δεν είναι απαραίτητο να "κρατηθούν" όλες οι p το πλήθος κύριες συνιστώσες, αλλά ένα υποσύνολο από αυτές και ειδικότερα οι κύριες συνιστώσες που αντιστοιχούν στα πρώτα l ιδιοδιανύσματα. Το υποσύνολο αυτών επιλέγεται με διάφορα κριτήρια όπως π.χ. το scree plot ή το κριτήριο της ιδιοτιμής. Το μεγάλο πλεονέκτημά της παραπάνω διαδικασίας έγκειται στην ιδιαιτερότητα που διαθέτουν, λόγω του τρόπου κατασκευής τους, να εξηγούν πολύ μεγάλο ποσοστό της ολικής μεταβλητότητας που αναπτύσσεται μεταξύ των αρχικών p μεταβλητών, το οποίο τελικά κατανέμεται σε μερικές μόνο νέες μεταβλητές^[5]. Επομένως, επιτυγχάνεται η μείωση των αρχικών μεταβλητών και επιπρόσθετα οι νέες ασυσχέτιστες μεταβλητές παρέχουν ισχυρή πληροφόρηση σχετική με τα χαρακτηριστικά των αρχικών μεταβλητών.

Επιπρόσθετα, οι νέοι, μικρότερης διάστασης, γραμμικοί συνδυασμοί είναι ευκολότερο να ερμηνευτούν και μπορούν να χρησιμεύουν σαν ενδιάμεσο βήμα στην πορεία μιας πολυπλοκότερης ανάλυσης δεδομένων (MANOVA, LDA). Η PCA ασχολείται με την βασική δομή ενός μόνο δείγματος παρατηρήσεων για p μεταβλητές. Δεν απαιτεί τον εκ των προτέρων καθορισμό των μεταβλητών ως εξαρτημένων ή μη, καθώς και την ύπαρξη προκαθορισμένων ομάδων (unsupervised method). Ακόμη, αποτελεί μία ισχυρή μέθοδο για τη διευκρίνιση ύπαρξης ομάδων σε ένα πίνακα δεδομένων, ενώ παράλληλα μπορεί να χρησιμοποιηθεί για να διαπιστώσουμε αν υπάρχουν παράγοντες που πιθανόν να επιδρούν στα δεδομένα.

Η PCA αναλύει πίνακες δεδομένων στους οποίους οι παρατηρήσεις περιγράφονται από πολλές σχετιζόμενες μεταξύ τους ποσοτικές μεταβλητές. Μάλιστα, όσο περισσότερο συσχετισμένες είναι μεταξύ τους οι αρχικές μεταβλητές, είτε θετικά, είτε αρνητικά, τόσο πιο αξιόπιστα αποτελέσματα παίρνουμε από την εφαρμογή της μεθόδου.

Βασικοί στόχοι της Ανάλυσης Κύριων Συνιστωσών είναι:

1. Να εξάγει τις σημαντικότερες πληροφορίες από έναν πίνακα δεδομένων
2. Να μειώσει το μέγεθος των δεδομένων, διατηρώντας τις σημαντικότερες πληροφορίες
3. Να απλοποιήσει την περιγραφή του νέου συστήματος δεδομένων

4. Να αναλύσει τη δομή των παρατηρήσεων και των μεταβλητών

Συνοψίζοντας, λοιπόν, η PCA αντλεί τις βασικές πληροφορίες από έναν πίνακα δεδομένων και τις αναπαριστά σαν ένα νέο σύστημα ορθογώνιων μεταξύ τους μεταβλητών, τις Κύριες Συνιστώσες, και να αναδείξει την ομοιότητα των παρατηρήσεων και των μεταβλητών, ως σημεία σε διάγραμμα.

Από τις p μεταβλητές X_1, X_2, \dots, X_p , δημιουργούνται p συνδυασμοί αυτών, Z_1, Z_2, \dots, Z_p , με τέτοιο τρόπο ώστε να μη συσχετίζονται μεταξύ τους και να είναι μεταξύ τους ορθογώνιες. Η απουσία συσχετισμού μεταξύ των μεταβλητών Z_i προδιαθέτει ότι αυτές μετρούν διαφορετικές «διαστάσεις» των στοιχείων^[5].

Η ανάλυση των δεδομένων ξεκινά με τον σχηματισμό των κύριων συνιστωσών που είναι της μορφής:

$$Z_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \dots + \alpha_{1p}X_p,$$

όπου α_{ip} ειδικός συντελεστής στάθμισης (weight) της p μεταβλητής στην i συνιστώσα και με τον περιορισμό ότι,

$$a_{11}^2 + a_{12}^2 + \dots + a_{1p}^2 = 1$$

ώστε να εξασφαλίζεται η μέγιστη διακύμανση για την Z_1 .

Οι διακυμάνσεις που αναπτύσσονται μεταξύ των μεταβλητών Z_i , διαβαθμίζονται με τέτοιο τρόπο ώστε η πρώτη μεταβλητή Z_1 επιλέγεται να εξηγεί ένα όσο το δυνατόν μέγιστο ποσοστό της ολικής μεταβλητότητας, η Z_2 ένα δεύτερο μέγιστο ποσοστό αυτής κοκ., υπακούοντας στη σχέση:

$$\text{Var}(Z_1) > \text{Var}(Z_2) > \dots > \text{Var}(Z_p).$$

Ο υπολογισμός των ανωτέρω διακυμάνσεων βασίζεται στις ιδιοτιμές (eigenvalues) του πίνακα συνδιακύμανσης (covariance matrix). Από την άλλη πλευρά, ο υπολογισμός των ειδικών συντελεστών στάθμισης, α_{ip} , προκύπτει από τον υπολογισμό των ιδιοδιανυσμάτων (eigenvectors) του πίνακα συνδιακύμανσης.

• **Ανάλυση Συστάδων – Cluster Analysis (CA)**

Η Ανάλυση Συστάδων είναι μία οικογένεια μεθόδων ταξινόμησης, η οποία έχει εφαρμογή σε πολλές επιστήμες. Οι πρώτες εφαρμογές προέρχονται από τις επιστήμες της Βιολογίας και της Ζωολογίας, όπου η μέθοδος ονομάζεται συνήθως αριθμητική ταξινόμηση (numerical taxonomy). Άλλες σημαντικές εφαρμογές είναι στο Marketing, την Αστρονομία, την Κλιματολογία, την Ψυχιατρική, την Αρχαιολογία, τη Βιοπληροφορική και στην τεχνητή νοημοσύνη για την αναγνώριση προτύπων (pattern recognition) (Everitt, Landau, Leese, & Stahl, 2011).

Η CA είναι μια από τις βασικότερες μεθόδους ταξινόμησης δεδομένων. Στόχος της CA είναι ο επιμερισμός ενός συνόλου παρατηρήσεων-πειραματικών μονάδων σε συστάδες. Οι συστάδες συγκροτούνται στη βάση της ομοιότητας των μελών τους. Το γεγονός ότι δεν υπάρχει εκ των προτέρων γνώση σχετικά με την ύπαρξη ομάδων χαρακτηρίζει την CA ως μη επιτηρούμενη μέθοδο^[2].

Υπάρχουν τρία θεμελιώδη βήματα στην πορεία της επεξεργασίας των δεδομένων:

1. Η περιγραφή των δεδομένων και η επιλογή των κατάλληλων χαρακτηριστικών.
2. Η επιλογή του τρόπου καθορισμού του βαθμού ομοιότητας μεταξύ των παρατηρήσεων.
3. Η επιλογή του αλγορίθμου με τον οποίο θα σχηματιστούν οι ομάδες.

Η φύση των παρατηρήσεων παίζει σημαντικό ρόλο στην επιλογή του τρόπου καθορισμού του βαθμού ομοιότητας μεταξύ των παρατηρήσεων. Για τις κατηγορικές μεταβλητές χρησιμοποιούνται τιμές εγγύτητας (proximity values), ενώ για τις ποσοτικές μεταβλητές μήτρες αποστάσεων (distance matrices).

Ένας τρόπος καθορισμού του βαθμού ομοιότητας δύο παρατηρήσεων είναι με τη χρήση της απόστασης τους. Οι παρατηρήσεις θεωρούνται ως σημεία σε έναν πολυδιάστατο χώρο. Η απόσταση τους σε αυτόν τον χώρο αποτελεί το μέτρο της ομοιότητας τους. Εάν όλες οι μεταβλητές είναι αριθμητικές, τότε για τον υπολογισμό της ανομοιότητας χρησιμοποιείται η Ευκλείδεια απόσταση ή κάποια παραλλαγή της, όπως η απόσταση Manhattan ή η απόσταση Minkowski. Για δυαδικές μεταβλητές, εάν αυτές είναι συμμετρικές, τότε η απόσταση των παρατηρήσεων x_a και x_b δίνεται από τον συντελεστή simple matching (simple matching coefficient), ενώ εάν είναι ασύμμετρες χρησιμοποιείται ο συντελεστής Jaccard. Για ονομαστικές μεταβλητές χρησιμοποιείται επίσης ο συντελεστής simple matching, ενώ για διατακτικές μεταβλητές, μετά από κατάλληλο μετασχηματισμό, η ευκλείδεια απόσταση. Σε γενικές γραμμές, όποιο μέτρο απόστασης και αν χρησιμοποιηθεί, οι μικρές αποστάσεις αντιστοιχούν σε παρόμοιες παρατηρήσεις και οι μεγαλύτερες αποστάσεις σε παρατηρήσεις με μεγάλες διαφορές στις τιμές των μεταβλητών.

Οι μέθοδοι Ανάλυσης Συστάδων χωρίζονται στις ακόλουθες κατηγορίες:

- Ιεραρχικές,
- Διαχωριστικές,
- Μέθοδοι βασισμένες στην πυκνότητα,
- Μέθοδοι πλέγματος και
- Μέθοδοι βασισμένες σε μοντέλα

Οι δύο πρώτες μέθοδοι αποτελούν και τις συχνότερα χρησιμοποιούμενες. Οι Ιεραρχικές μέθοδοι δημιουργούν μια ιεραρχία επιπέδων, κάθε ένα από τα οποία περιλαμβάνει ένα σύνολο συστάδων. Η επιλογή του κατάλληλου συνόλου συστάδων εναπόκειται στον χρήστη. Η ιεραρχία των επιπέδων και οι αντίστοιχες συστάδες αναπαρίστανται γραφικά με τη χρήση δένδρογραμμάτων. Οι Ιεραρχικές μέθοδοι υποδιαιρούνται σε συσσωρευτικές (agglomerative), οι οποίες δημιουργούν την ιεραρχία μέσα από μια διαδικασία διαδοχικών συγχωνεύσεων, και σε διαιρετικές (splitting), οι οποίες δημιουργούν την ιεραρχία μέσω διαδοχικών διασπάσεων. Για τη συγχώνευση ή διάσπαση συστάδων απαιτείται καθορισμός της απόστασης τους. Έχουν προταθεί διάφοροι τρόποι μέτρησης της απόστασης των συστάδων. Οι βασικότεροι από αυτούς είναι η μέθοδος της Απλής Σύνδεσης, η μέθοδος της Πλήρους Σύνδεσης, η Σύνδεση Μέσου Όρου και η μέθοδος Ward. Κάθε ενέργεια, η οποία πραγματοποιείται σε ένα στάδιο, είναι μη αναστρέψιμη, δηλαδή από τη στιγμή που ένα αντικείμενο ενταχθεί σε μια ομάδα, θα παραμείνει σε αυτή, και δεν υπάρχει δυνατότητα να ενταχθεί αργότερα σε κάποια άλλη^[2].

Στη Διαχωριστική (Μη Ιεραρχική) μέθοδο, τα αντικείμενα επιμερίζονται σε k συστάδες. Τυπικά, ο αριθμός των συστάδων προκαθορίζεται από τον χρήστη. Στη συνέχεια, εφαρμόζεται μια επαναληπτική διαδικασία, κατά την οποία τα αντικείμενα μετακινούνται από μια συστάδα σε μια άλλη. Η ποιότητα της κάθε λύσης ενδεχόμενων συστάδων μετριέται με τη βοήθεια ενός κριτηρίου. Σε κάθε επανάληψη, και με τη μετακίνηση των σημείων, η τιμή του κριτηρίου μειώνεται. Είναι υπολογιστικά λιγότερο ακριβείς από τις ιεραρχικές μεθόδους, και για τον λόγο αυτό μπορούν να εφαρμοστούν σε μεγαλύτερα σύνολα δεδομένων. Ο πιο γνωστός αλγόριθμος Διαχωριστικής Ανάλυσης Συστάδων είναι ο k -Means^[2].

Ανάλογα με την επιλογή του μέτρου ομοιότητας και της μεθόδου ομαδοποίησης, οι ομάδες που προκύπτουν είναι διαφορετικές. Ο ερευνητής πρέπει να κάνει τις επιλογές αυτές ανάλογα με τη φύση των δεδομένων και το πρόβλημα που εξετάζει. Συχνά απαιτούνται πολλές δοκιμές της ανάλυσης συστάδων, περιλαμβάνοντας διαφορετικές μεταβλητές ή αφαιρώντας κάποιες παρατηρήσεις και χρησιμοποιώντας διαφορετικά μέτρα σύγκρισης, ώστε να εξακριβωθεί η σταθερότητα της ομαδοποίησης. Το τελικό αποτέλεσμα πρέπει να μπορεί να ερμηνευτεί.

Δεν υπάρχει κάποια γενικώς αποδεκτή ως «καλύτερη» μέθοδος. Δυστυχώς, διαφορετικοί αλγόριθμοι δεν παράγουν απαραίτητα τα ίδια αποτελέσματα από τα ίδια δεδομένα και συνήθως υπάρχει έντονο υποκειμενικό κριτήριο στην αξιολόγηση των αποτελεσμάτων, ανεξαρτήτως μεθόδου. Ένας αξιόπιστος τρόπος ελέγχου κάθε αλγόριθμου είναι να πάρουμε δεδομένα με γνωστή κατηγοριοποίηση και να δούμε εάν οι αλγόριθμοι μπορούν να αποδώσουν την ίδια κατηγοριοποίηση.

• Υλοποίηση Μεθόδων

Για την καλύτερη κατανόηση του τρόπου και των ενδείξεων εφαρμογής των μεθόδων που αναλύθηκαν παραπάνω, θα εφαρμόσουμε τις τεχνικές αυτές σε ένα σύνολο ιατρικών πολυμεταβλητών δεδομένων. Τα δεδομένα προέρχονται από την ιστοσελίδα <https://data.world/uci/heart-disease>^[8]. Πιο συγκεκριμένα, επιλέχθηκε η βάση δεδομένων του Cleveland Clinic Foundation, η οποία περιλαμβάνει παρατηρήσεις που προέρχονται από ένα δείγμα 303 ασθενών για 76 μεταβλητές^[12]. Στην ανάλυση περιλαμβάνονται 6 από αυτές, λόγω έλλειψης παρατηρήσεων για τις υπόλοιπες ή παραβίασης των υποθέσεων εφαρμογής των υπό εξέταση τεχνικών. Η τελευταία μεταβλητή “group” αναφέρεται στην παρουσία Καρδιαγγειακής Νόσου στους ασθενείς, παίρνοντας τιμές από 0 (απουσία νόσου) έως 4, ανάλογα με τη βαρύτητα της νόσου, βάσει προηγούμενων μελετών που έγιναν με την ίδια βάση δεδομένων. Για την στατιστική επεξεργασία των δεδομένων θα γίνει χρήση του προγράμματος IBM SPSS Statistics 25[®]. Αναλυτικότερα, οι υπό ανάλυση μεταβλητές είναι οι κάτωθι:

1. Age (years)
2. Resting systolic blood pressure (in mm Hg on admission to the hospital)
3. Serum cholesterol (mg/dl)
4. Maximum heart rate/min
5. ST depression (in mm) induced by exercise relative to rest
6. Presence of heart disease, (0=Absence, 1=Mild, 2=Moderate, 3=High, 4=Severe)

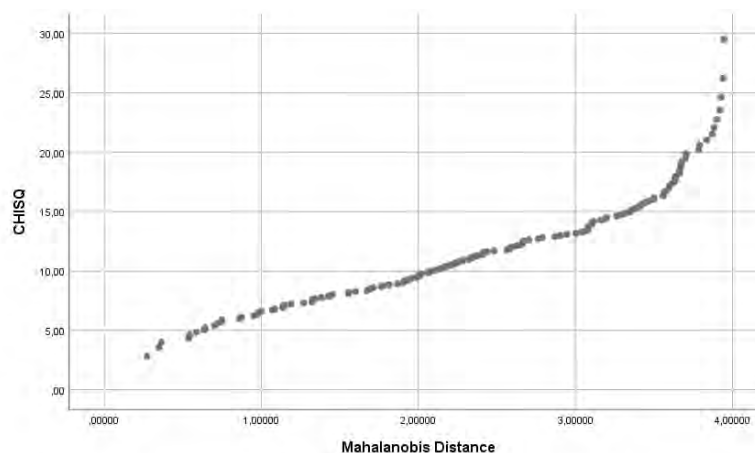
Θα προσαρμόσουμε το ερευνητικό ερώτημα κάθε παραδείγματος στην εκάστοτε τεχνική πολυμεταβλητής ανάλυσης, προκειμένου να αναδείξουμε τις ενδείξεις εφαρμογής τους.

Δ. ΑΠΟΤΕΛΕΣΜΑΤΑ

- **Γραμμική Διαχωριστική Ανάλυση – Linear Discriminant Analysis (LDA)**

Σκοπός του παραδείγματος είναι να ταξινομήσουμε τους ασθενείς σε ομάδες ανάλογα με τη βαρύτητα της Καρδιαγγειακής τους Νόσου. Στην ανάλυση των δεδομένων με την συγκεκριμένη τεχνική χρησιμοποιήθηκαν και οι 5 μεταβλητές, ενώ σαν a priori γνωστές κλάσεις χρησιμοποιήθηκε η ήδη γνωστή ταξινόμηση των ασθενών σε 5 κατηγορίες (0 – 4) Καρδιαγγειακής Νόσου. Πριν από την εφαρμογή της μεθόδου, θα πρέπει να ελέγξουμε εάν παραβιάζονται οι βασικές υποθέσεις της.

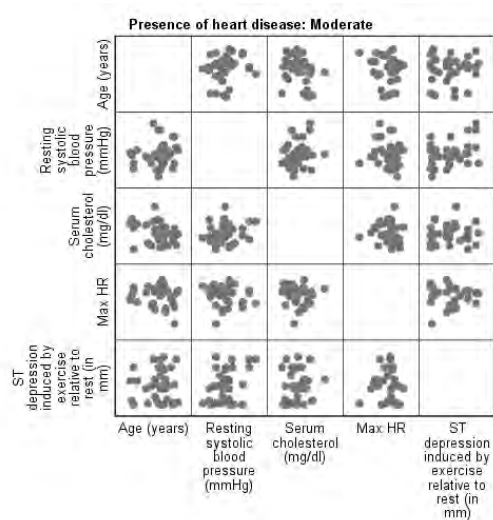
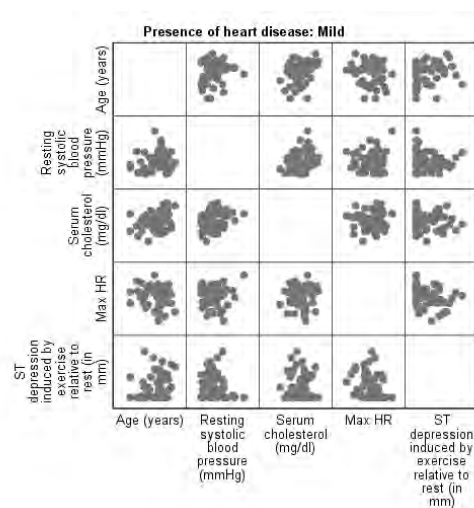
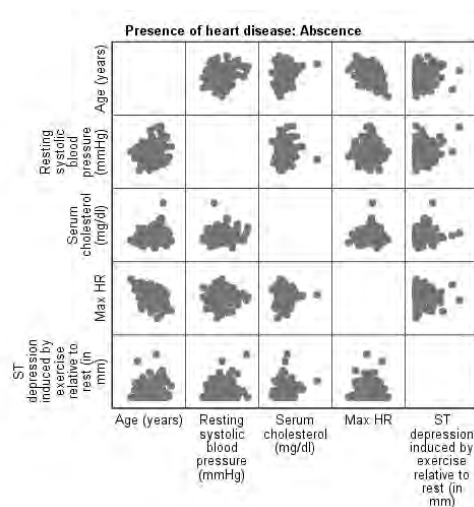
Για να ελέγξουμε εάν οι μεταβλητές ακολουθούν την πολυμεταβλητή κανονική κατανομή θα πρέπει να εφαρμόσουμε το **Mardia's test**. Επειδή όμως το συγκεκριμένο test μας δίνει πληροφορίες σχετικές με την λοξότητα (skewness) και την κύρτωση (kurtosis) και επομένως μόνο έμμεσα μας πληροφορεί για την κανονικότητα των δεδομένων, θα προχωρήσουμε σε γραφικό έλεγχο της υπόθεσης με σχηματισμό του διαγράμματος **Chi-square versus Mahalanobis distance**^[13]. Αν οι μεταβλητές ακολουθούν την πολυμεταβλητή κανονική κατανομή θα σχηματίσουν μια ευθεία γραμμή.

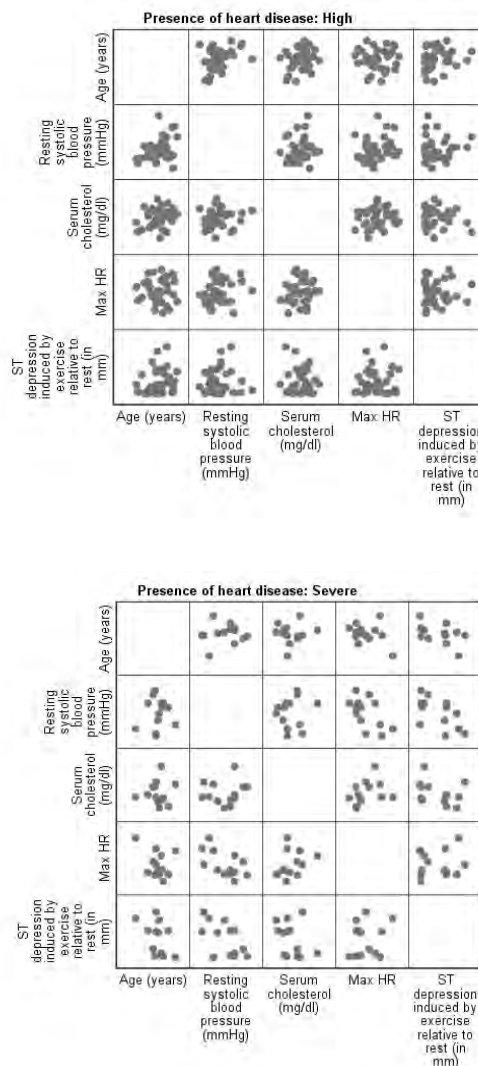


Παρατηρούμε ότι τα δεδομένα προσεγγίζουν σε καλό βαθμό την πολυμεταβλητή κανονική κατανομή. Η μικρή απόκλιση της ευθείας στο δεξιό άκρο, δεν είναι ικανός όρος για να απορρίψουμε την υπόθεση. Σε κάθε περίπτωση, η παραβίαση της υπόθεσης της κανονικότητας, είτε η LDA χρησιμοποιείται ως τεχνική μείωσης των διαστάσεων, είτε για ταξινόμηση παρατηρήσεων, δεν μας αποτρέπει από το να κάνουμε χρήση της μεθόδου και μάλιστα με αξιόπιστα αποτελέσματα^[14].

Για να ελέγξουμε εάν οι μεταβλητές σχετίζονται μεταξύ τους γραμμικά, δημιουργήσαμε τα παρακάτω **scatter-matrix plots**. Η εμφάνιση ελλειπτικών σχημάτων στα επιμέρους κελιά, θα σήμαινε την ύπαρξη γραμμικής συσχέτισης μεταξύ των δύο υπό σύγκριση μεταβλητών σε κάθε γκρουπ. Όπως παρατηρούμε παρακάτω, δεν σχηματίζονται

ελλειπτικά σχήματα, επομένως πληρείται η υπόθεση της μη γραμμικής συσχέτισης μεταξύ των αρχικών μεταβλητών.





Τέλος, για τον έλεγχο της υπόθεσης της ύπαρξης πολυσυγγραμμικότητας, προχωρήσαμε στον σχηματισμό του πίνακα συσχετίσεων, για να υπολογίσουμε τον **συντελεστή Pearson**. Προϋπόθεση για την απόρριψη της ύπαρξης πολυσυγγραμμικότητας, είναι η τιμή του συντελεστή να είναι μικρότερη από 0,8. Όπως παρατηρούμε, ο συντελεστής είναι μικρότερος από 0,8 για όλες τις πιθανές συσχετίσεις και άρα με ασφάλεια μπορούμε να αποκλείσουμε την ύπαρξη πολυσυγγραμμικότητας.

Correlations

		Age (years)	Resting systolic blood pressure (mmHg)	Serum cholesterol (mg/dl)	Max HR	ST depression induced by exercise relative to rest (in mm)
Age (years)	Pearson Correlation	1	,285**	,209**	-,394**	,182**
	Sig. (2-tailed)		,000	,000	,000	,001
	N	303	303	303	303	303
Resting systolic blood pressure (mmHg)	Pearson Correlation	,285**	1	,130*	-,045	,155**
	Sig. (2-tailed)	,000		,023	,432	,007

	N	303	303	303	303	303
Serum cholesterol (mg/dl)	Pearson Correlation	,209**	,130*	1	-,003	,024
	Sig. (2-tailed)	,000	,023		,953	,675
	N	303	303	303	303	303
Max HR	Pearson Correlation	-,394**	-,045	-,003	1	-,270**
	Sig. (2-tailed)	,000	,432	,953		,000
	N	303	303	303	303	303
ST depression induced by exercise relative to rest (in mm)	Pearson Correlation	,182**	,155**	,024	-,270**	1
	Sig. (2-tailed)	,001	,007	,675	,000	
	N	303	303	303	303	303

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

Καταλήγουμε, λοιπόν, στο συμπέρασμα ότι μπορούμε με σχετική αξιοπιστία να εφαρμόσουμε την LDA στα δεδομένα μας. Από την σχετική ανάλυση, προέκυψαν τα παρακάτω αποτελέσματα:

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		303	100,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Total	0	,0
Total		303	100,0

Στον παραπάνω πίνακα βλέπουμε ότι χρησιμοποιήθηκαν και οι 303 παρατηρήσεις.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,405 ^a	93,0	93,0	,537
2	,021 ^a	4,8	97,8	,143
3	,007 ^a	1,6	99,4	,084
4	,002 ^a	,6	100,0	,050

a. First 4 canonical discriminant functions were used in the analysis.

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 4	,690	110,027	20	,000
2 through 4	,970	8,994	12	,703
3 through 4	,990	2,859	6	,826
4	,998	,737	2	,692

Στον πίνακα **Eigenvalues**, παρουσιάζονται οι ιδιοτιμές των διακριτικών συναρτήσεων. Ο μέγιστος αριθμός των διακριτικών συναρτήσεων είναι ίσος με τον αριθμό των ομάδων μείον ένα ($5-1=4$), επομένως εξάγονται τέσσερις συναρτήσεις.

- Η 1^η διακριτική συνάρτηση είναι πιο σημαντική καθώς συγκεντρώνει το 93% της διακύμανσης.
- Ο συντελεστής κανονικής συσχέτισης (**canonical correlation**) εκφράζει τη σχέση μεταξύ των ομάδων και της διακριτικής συνάρτησης και όσο υψηλότερος είναι τόσο καλύτερος ο διαχωρισμός των ομάδων. Για την πρώτη συνάρτηση η τιμή του συντελεστή κανονικής συσχέτισης είναι 0,537.

Ο συντελεστής **Wilks' Lambda**, είναι ένα μέτρο αξιολόγησης του πόσο καλά κάθε διακριτική συνάρτηση διαχωρίζει τις παρατηρήσεις στις ομάδες. Παίρνει τιμές από 0-1 και όσο μικρότερος είναι τόσο καλύτερος διαχωρισμός επιτυγχάνεται. Η 1^η συνάρτηση, έχει το μικρότερο Wilks' Lambda=0,690, με $p\text{-value}<0,001$ και είναι στατιστικά σημαντική, έναντι των υπολοίπων συναρτήσεων που έχουν $p\text{-value}>0,05$ και δεν είναι στατιστικά σημαντικές.

Standardized Canonical Discriminant Function Coefficients

	Function			
	1	2	3	4
Age (years)	-,008	,781	,447	,263
Resting systolic blood pressure (mmHg)	,194	-,082	-,282	,833
Serum cholesterol (mg/dl)	,145	,060	,621	-,359
Max HR	-,734	,764	-,206	-,048
ST depression induced by exercise relative to rest (in mm)	,545	,493	-,524	-,449

Στον πίνακα αυτό παρουσιάζονται οι τυποποιημένοι συντελεστές των κανονικών διακριτικών συναρτήσεων, οι οποίοι αντιστοιχούν σε σχετικά βάρη (weights) και από τους οποίους προκύπτει η σχετική συνεισφορά των αρχικών μεταβλητών στις διακριτικές συναρτήσεις. Το μέγεθος υποδηλώνει τη σπουδαιότητα της μεταβλητής και το πρόσημο την κατεύθυνση της σχέσης.

Παρατηρούμε ότι στην πρώτη συνάρτηση τη μεγαλύτερη συνεισφορά έχει η μεταβλητή «**Max HR**», στην δεύτερη συνάρτηση η μεταβλητή «**Age**», στην τρίτη συνάρτηση η μεταβλητή «**Serum cholesterol**» και στην τέταρτη, η μεταβλητή «**Resting systolic blood pressure**».

Structure Matrix

	Function			
	1	2	3	4
Max HR	-,789*	,438	-,276	-,074
ST depression induced by exercise relative to rest (in mm)	,655*	,467	-,488	-,327
Age (years)	,375	,561*	,511	,380
Serum cholesterol (mg/dl)	,134	,225	,670*	-,207
Resting systolic blood pressure (mmHg)	,247	,197	-,152	,811*

Ο πίνακας αυτός παρουσιάζει τους συντελεστές δομής, οι οποίοι είναι οι συντελεστές συσχέτισης κάθε διακριτικής μεταβλητής με τις διακριτικές συναρτήσεις. Αποτελούν έναν εναλλακτικό τρόπο αξιολόγησης της σχετικής σπουδαιότητας των μεταβλητών. Οι μεταβλητές διατάσσονται σύμφωνα με το απόλυτο μέγεθος των συντελεστών δομής.

Canonical Discriminant Function Coefficients

	Function			
	1	2	3	4
Age (years)	-,001	,089	,051	,030
Resting systolic blood pressure (mmHg)	,011	-,005	-,016	,048
Serum cholesterol (mg/dl)	,003	,001	,012	-,007
Max HR	-,036	,037	-,010	-,002
ST depression induced by exercise relative to rest (in mm)	,053	,048	-,051	-,044
(Constant)	2,780	-10,474	-1,654	-5,461

Στον πίνακα **Canonical Discriminant Function Coefficients** παρουσιάζονται οι συντελεστές των διακριτικών συναρτήσεων οι οποίοι είναι μη τυποποιημένοι (unstandardized). Η πρώτη διακριτική συνάρτηση, που είναι και η σημαντικότερη, παίρνει λοιπόν τη μορφή:

$$Z_1 = 2,780 - 0,001 * \text{Age} + 0,011 * \text{SBP} + 0,003 * \text{Chol} - 0,036 * \text{HR} + 0,053 * \text{STd}$$

Classification Results^a

		Predicted Group Membership					Total
		Presence of heart disease	Absence	Mild	Moderate	High	
Original	Count	Absence	111	21	10	8	164
		Mild	21	10	2	11	55
		Moderate	7	7	2	11	36
		High	5	5	4	14	35

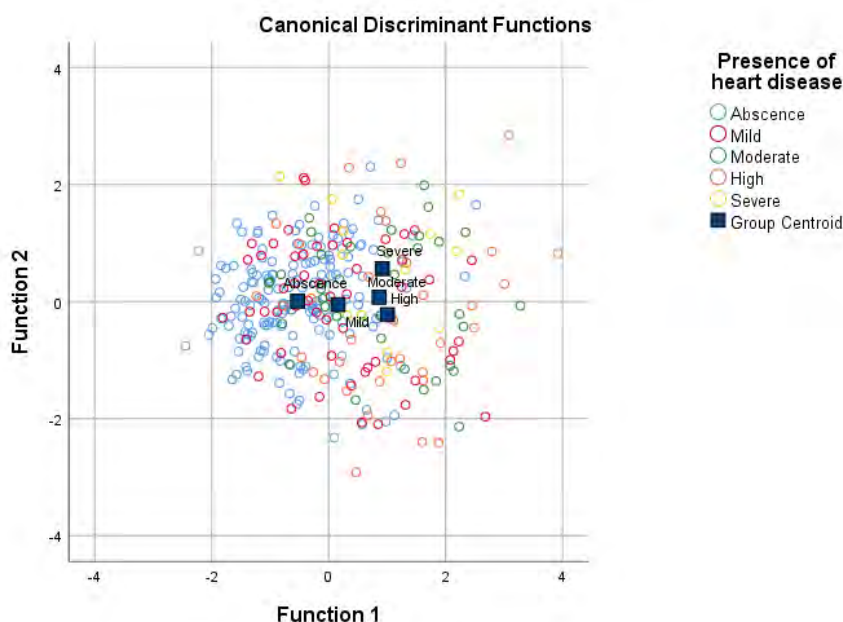
	Severe	1	3	0	3	6	13
%	Absence	67,7	12,8	6,1	4,9	8,5	100,0
	Mild	38,2	18,2	3,6	20,0	20,0	100,0
	Moderate	19,4	19,4	5,6	30,6	25,0	100,0
	High	14,3	14,3	11,4	40,0	20,0	100,0
	Severe	7,7	23,1	,0	23,1	46,2	100,0

a. 47,2% of original grouped cases correctly classified.

Ο παραπάνω πίνακας δείχνει πόσες από τις αρχικές παρατηρήσεις έχουν ταξινομηθεί σωστά, βάσει της ήδη γνωστής ταξινόμησης. Αναλυτικότερα:

- Από τους 164 ασθενείς της ομάδας 0 (Absence of HD), 111 κατατάχθηκαν ορθά (67,7% ορθή ταξινόμηση).
- Από τους 55 ασθενείς της ομάδας 1 (Mild HD), 10 κατατάχθηκαν στη σωστή ομάδα (18,2% ορθή ταξινόμηση).
- Από τους 36 ασθενείς της ομάδας 2 (Moderate HD), 2 κατατάχθηκαν σωστά (5,6% ορθή ταξινόμηση).
- Από τους 35 ασθενείς της ομάδας 3 (High HD), 14 κατατάχθηκαν ορθά (40% ορθή ταξινόμηση).
- Από τους 13 ασθενείς της ομάδας 4 (Severe HD), 6 κατατάχθηκαν σωστά (46,2% ορθή ταξινόμηση).

Επομένως, υπάρχει ένα συνολικό ποσοστό ορθής ταξινόμησης ίσο με 47,2%.



Γενικά, ισχύει ότι όσο πιο κοντά μεταξύ τους βρίσκονται τα κέντρα των διάφορων ομάδων (group centroids), τόσο πιθανότερο είναι να έχουμε δυσταξινόμηση. Από το διάγραμμα παρατηρούμε ότι δεν υπάρχει πλήρης διαχωρισμός των ασθενών στις πέντε ομάδες κινδύνου για Καρδιαγγειακή Νόσο, αλλά υπάρχει επικάλυψη μεταξύ τους. Αξίζει να τονιστεί ότι, η ομάδα 0 που αντιστοιχεί σε απουσία νόσου, διαφοροποιείται σαφώς από τις

ομάδες 1,2,3,4 που αντιστοιχούν σε υψηλή πιθανότητα ύπαρξης Καρδιαγγειακής Νόσου και άρα οι υπό μελέτη μεταβλητές μπορούν με μεγάλη αξιοπιστία να διαχωρίσουν την απουσία από την παρουσία νόσου. Ακόμη, παρατηρούμε ότι τα κέντρα των ομάδων 0 και 1 διαχωρίζονται σαφώς από τα κέντρα των υπολοίπων ομάδων που αντιστοιχούν σε άνω του μετρίου σοβαρότητα Καρδιαγγειακής Νόσου.

Συνοψίζοντας, πραγματοποιήθηκε Γραμμική Διαχωριστική Ανάλυση για να εκτιμηθεί πόσο καλά θα μπορούσε να προβλεφθεί η σοβαρότητα Καρδιαγγειακής Νόσου από ένα σύνολο 5 μεταβλητών. Η ανάλυση αυτή παρήγαγε τέσσερις διακριτικές συναρτήσεις, με την πρώτη να αντιστοιχεί στο 93% της συνολικής διακύμανσης. Τέλος, το 47,2% των περιπτώσεων κατατάχθηκε σωστά στην αρχική τους κατηγορία.

• Ανάλυση Κύριων Συνιστωσών - Principal Component Analysis (PCA)

Οι ασθενείς κατατάσσονται σε κατηγορίες πρόγνωσης κινδύνου για Καρδιαγγειακή Νόσο σε μία δεκαετία, με βάση τον δείκτη Framingham Risk Score. Σκοπός του συγκεκριμένου παραδείγματος, είναι να εκτιμηθεί εάν από τις 5 υπό ανάλυση μεταβλητές της ίδιας βάσης δεδομένων, μπορούμε να εξάγουμε πέντε νέες, ασυσχέτιστες μεταξύ τους, μεταβλητές, οι οποίες θα μπορούν να προβλέψουν ικανοποιητικά τον Καρδιαγγειακό κίνδυνο. Το μέγεθος του δείγματος (303 παρατηρήσεις, ιδανικά >150) είναι ικανοποιητικό για την εφαρμογή της μεθόδου. Η γραμμική συσχέτιση μεταξύ των μεταβλητών έχει ελεγχθεί ως υπόθεση, μέσω των **scatter-matrix plots** για την προηγούμενη ανάλυση. Από την σχετική ανάλυση παρήχθησαν τα παρακάτω αποτελέσματα:

Correlation Matrix

		Age (years)	Resting systolic blood pressure (mmHg)	Serum cholesterol (mg/dl)	Max HR	ST depression induced by exercise relative to rest (in mm)
Correlation	Age (years)	1,000	,285	,209	-,394	,182
	Resting systolic blood pressure (mmHg)	,285	1,000	,130	-,045	,155
	Serum cholesterol (mg/dl)	,209	,130	1,000	-,003	,024
	Max HR	-,394	-,045	-,003	1,000	-,270
	ST depression induced by exercise relative to rest (in mm)	,182	,155	,024	-,270	1,000

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,553
Bartlett's Test of Sphericity	Approx. Chi-Square	124,633
	df	10
	Sig.	,000

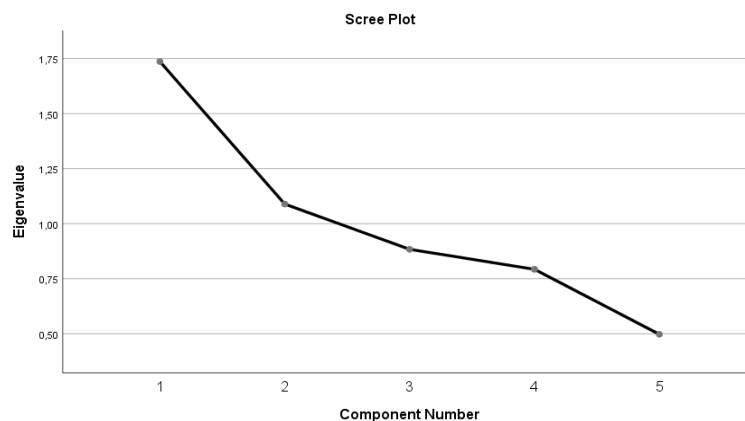
Από τον πίνακα **Correlation Matrix**, παίρνουμε το είδος (θετική-αρνητική) και το ύψος της συσχέτισης μεταξύ των αρχικών μεταβλητών. Όπως παρατηρούμε, η πλειοψηφία των συσχετίσεων είναι θετική και κατά βάση χαμηλή. Αυτό αποτελεί μειονέκτημα για την PCA, αφού θα πρέπει να είμαστε επιφυλακτικοί για την αξιοπιστία των αποτελεσμάτων.

Παρόλα αυτά, από τον επόμενο πίνακα βλέπουμε ότι, η τιμή του **Kaiser-Meyer-Olkin** δείκτη (0,553) είναι ικανοποιητική και άρα υπάρχει ένδειξη για εφαρμογή της PCA, γεγονός που επιβεβαιώνεται και από την στατιστικά σημαντική τιμή του **Bartlett's Test** ($p\text{-value} < 0,001$), η οποία επιβεβαιώνει την ύπαρξη συσχέτισης μεταξύ των αρχικών μεταβλητών.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1,736	34,730	34,730	1,736	34,730	34,730	1,003	20,062	20,062
2	1,089	21,784	56,514	1,089	21,784	56,514	1,002	20,047	40,109
3	,884	17,674	74,188	,884	17,674	74,188	1,002	20,044	60,153
4	,793	15,858	90,046	,793	15,858	90,046	1,002	20,032	80,184
5	,498	9,954	100,000	,498	9,954	100,000	,991	19,816	100,000

Extraction Method: Principal Component Analysis.



Από τον πρώτο πίνακα, παίρνουμε τις **ιδιοτιμές** κάθε Κύριας Συνιστώσας και παρατηρούμε ότι οι 2 πρώτες Κύριες Συνιστώσες ($\text{eigenvalue} > 1$), αντιστοιχούν σε $> 50\%$ της συνολικής διακύμανσης ($\text{Cumulative\%} = 56,514\%$). Η παρατήρηση αυτή επιβεβαιώνεται και από τον **έλεγχο κρημνού (scree test)**, με την εφαρμογή του κρημνογραφήματος (scree plot), μέσω του οποίου βλέπουμε ότι οι σημαντικότερες συνιστώσες βρίσκονται στο άνω σκέλος του αγκώνα. Με άλλα λόγια, οι κυριότερες συνιστώσες είναι εκείνες (σημεία στο κρημνογράφημα) που βρίσκονται στο τμήμα της απότομης πτώσης της ευθείας διαλογής και μέχρι το σημείο που η πτώση συνεχίζεται επιβραδυνόμενη.

Communalities

	Initial	Extraction
Age (years)	1,000	1,000
Resting systolic blood pressure (mmHg)	1,000	1,000
Serum cholesterol (mg/dl)	1,000	1,000
Max HR	1,000	1,000
ST depression induced by exercise relative to rest (in mm)	1,000	1,000

Από τον παραπάνω πίνακα, παρατηρούμε τί ποσοστό της διακύμανσης κάθε μεταβλητής εξηγείται από το μοντέλο των πέντε Κύριων Συνιστωσών, που η ανάλυση μας έχει δώσει. Παρατηρούμε ότι, το συγκεκριμένο μοντέλο ικανοποιεί σε υψηλό βαθμό την πλειοψηφία των μεταβλητών.

Component Matrix^a

	Component				
	1	2	3	4	5
Age (years)	,784	,091	-,205	-,330	,475
Max HR	-,656	,495	,346	,162	,423
Serum cholesterol (mg/dl)	,336	,715	-,422	,422	-,140
Resting systolic blood pressure (mmHg)	,520	,428	,652	-,256	-,237
ST depression induced by exercise relative to rest (in mm)	,555	-,376	,344	,643	,135

Extraction Method: Principal Component Analysis.

Rotated Component Matrix^a

	Component				
	1	2	3	4	5
Resting systolic blood pressure (mmHg)	,987	,059	,073	-,005	,134
Serum cholesterol (mg/dl)	,058	,993	,007	,008	,098
ST depression induced by exercise relative to rest (in mm)	,073	,007	,986	-,128	,072
Max HR	-,005	,009	-,133	,972	-,195
Age (years)	,145	,108	,077	-,203	,959

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Στους δύο παραπάνω πίνακες βλέπουμε τους ειδικούς συντελεστές στάθμισης των αρχικών μεταβλητών που παίζουν καθοριστικό ρόλο στον υπολογισμό των Κύριων Συνιστωσών. Οι συντελεστές αυτοί αποτελούν και δείκτη βαρύτητας της κάθε μεταβλητής, ενώ το πρόσημο τους φανερώνει και το είδος της σχέσης τους με την κύρια συνιστώσα. Ο πίνακας **Component Matrix** μας δίνει τους συντελεστές των αρχικών Κύριων Συνιστωσών, ενώ ο επόμενος **Rotated Component Matrix**, τους συντελεστές που προκύπτουν μετά τον ορθογώνιο μετασχηματισμό.

Έτσι έχουμε: $Z_1 = 0,784 * \text{Age} - 0,656 * \text{HR} + 0,555 * \text{STd} + 0,520 * \text{SBP} + 0,336 * \text{Chol}$,
 $Z_2 = 0,091 * \text{Age} + 0,495 * \text{HR} - 0,376 * \text{STd} + 0,428 * \text{SBP} + 0,715 * \text{Chol}$,
 $Z_3 = -0,205 * \text{Age} + 0,346 * \text{HR} + 0,344 * \text{STd} + 0,652 * \text{SBP} - 0,422 * \text{Chol}$,
 $Z_4 = -0,330 * \text{Age} + 0,162 * \text{HR} + 0,643 * \text{STd} - 0,256 * \text{SBP} + 0,422 * \text{Chol}$,
 $Z_5 = 0,475 * \text{Age} + 0,423 * \text{HR} + 0,135 * \text{STd} - 0,237 * \text{SBP} - 0,140 * \text{Chol}$

και μετά τον ορθογώνιο μετασχηματισμό:

$Z_1 = 0,987 * \text{SBP} + 0,058 * \text{Chol} + 0,073 * \text{STd} - 0,005 * \text{HR} + 0,145 * \text{Age}$,
 $Z_2 = 0,059 * \text{SBP} + 0,993 * \text{Chol} + 0,007 * \text{STd} + 0,009 * \text{HR} + 0,108 * \text{Age}$,

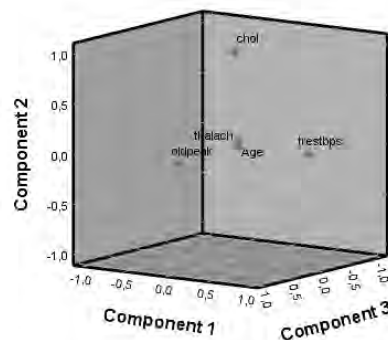
$$Z_3 = 0,073*SBP + 0,007*Chol + 0,986*STd - 0,133*HR + 0,077*Age,$$

$$Z_4 = -0,005*SBP + 0,008*Chol - 0,128*STd + 0,972*HR - 0,203*Age,$$

$$Z_5 = 0,134*SBP + 0,098*Chol + 0,072*STd - 0,195*HR + 0,959*Age$$

Το παρακάτω διάγραμμα αποτελεί την γραφική αναπαράσταση του ορθογώνιου μετασχηματισμού των Κύριων Συνιστωσών. Παρατηρούμε ότι υπάρχει σαφής διαχωρισμός μεταξύ των Κύριων Συνιστωσών και άρα η απουσία της μεταξύ τους συσχέτισης, αποτελεί θετικό γεγονός για την αξιοπιστία των αποτελεσμάτων που παράγουν.

Component Plot in Rotated Space



Συνεπώς, οι παραγόμενες Κύριες Συνιστώσες μπορούν να λειτουργήσουν ικανοποιητικά ως ασυσχέτιστοι παράγοντες πρόγνωσης του κινδύνου Καρδιαγγειακής Νόσου, βάσει του score που θα επιτυγχάνουν.

• Ανάλυση Συστάδων – Cluster Analysis (CA)

Με την τεχνική αυτή θα επιδιώξουμε να ομαδοποιήσουμε τα δεδομένα μας σε όσο το δυνατόν πιο ομοιογενείς ομάδες, επιτυγχάνοντας παράλληλα τον μέγιστο, μεταξύ των ομάδων, διαχωρισμό. Για χάρη του παραδείγματος θα υποθέσουμε ότι δεν υπάρχει a priori γνωστός αριθμός ομάδων ή γενικότερα δυνατότητα διαχωρισμού των παρατηρήσεων μας σε ομάδες. Έτσι, θα εφαρμόσουμε την τεχνική της Ανάλυσης Συστάδων, βασιζόμενοι στον Συσσωρευτικό Ιεραρχικό αλγόριθμο της Μονής Σύνδεσης (**Single linkage algorithm**). Ακόμη, επειδή οι μεταβλητές μας δεν έχουν τις ίδιες μονάδες μετρήσεις, θα τις μετασχηματίσουμε πρώτα σε **z-values**, μέσω του SPSS, και στην συνέχεια θα προχωρήσουμε στον σχηματισμό των ομάδων. Τα δεδομένα που παίρνουμε είναι τα ακόλουθα:

Case Processing Summary^a

Valid		Cases		Total	
		Missing			
N	Percent	N	Percent	N	Percent
303	100,0%	0	0,0%	303	100,0%

a. Squared Euclidean Distance used

Στην ανάλυση συμπεριλήφθηκαν και οι 303 περιπτώσεις και από τον υπολογισμό των τετραγώνων των Ευκλείδειων αποστάσεων, προέκυψε ο πίνακας αποστάσεων εγγύτητας (**Proximity matrix**), ο οποίος για χάρη συντομίας παρουσίασης των δεδομένων παραλείπεται. Με τη βοήθεια αυτού, σχηματίστηκε ο παρακάτω Συσσωρευτικός πίνακας, στον οποίο φαίνεται αναλυτικά η διαδικασία ενοποίησης των κλάσεων βάσει της ομοιότητας των παρατηρήσεων. Παρατηρήσεις που απέχουν την μικρότερη ευκλείδεια απόσταση, θεωρούνται παρόμοιες και ενώνονται σε ενιαίες κλάσεις, μέχρι τον τελικό σχηματισμό μίας.

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	69	149	,015	0	0	11
2	176	248	,044	0	0	15
3	65	78	,051	0	0	38
4	132	136	,055	0	0	26
5	170	262	,061	0	0	220
6	108	120	,063	0	0	12
7	42	157	,071	0	0	20
8	22	68	,076	0	0	52
9	23	47	,087	0	0	115
10	59	158	,101	0	0	55
11	69	143	,111	1	0	13
12	48	108	,116	0	6	40
13	69	75	,124	11	0	46
14	172	221	,128	0	0	62
15	176	291	,134	2	0	155
16	138	216	,140	0	0	98
17	159	184	,140	0	0	110
18	92	160	,146	0	0	38
19	44	51	,150	0	0	21
20	42	91	,161	7	0	48
21	44	222	,169	19	0	31
22	11	76	,173	0	0	123
23	133	135	,176	0	0	59
24	105	247	,178	0	0	42
25	119	181	,196	0	0	74
26	132	183	,204	4	0	29
27	177	230	,204	0	0	34
28	13	276	,208	0	0	78
29	115	132	,210	0	26	40
30	50	93	,212	0	0	56
31	44	210	,214	21	0	39

32	81	123	,214	0	0	66
33	34	71	,221	0	0	39
34	6	177	,230	0	27	75
35	100	219	,235	0	0	65
36	202	297	,235	0	0	212
37	179	273	,235	0	0	79
38	65	92	,242	3	18	65
39	34	44	,242	33	31	57
40	48	115	,243	12	29	54
41	60	94	,246	0	0	61
42	105	286	,252	24	0	51
43	190	268	,259	0	0	256
44	110	191	,262	0	0	47
45	30	56	,270	0	0	129
46	69	117	,273	13	0	70
47	110	204	,279	44	0	71
48	8	42	,279	0	20	52
49	156	288	,281	0	0	139
50	43	88	,283	0	0	72
51	105	129	,284	42	0	56
52	8	22	,284	48	8	66
53	31	114	,286	0	0	74
54	48	269	,288	40	0	61
55	59	116	,290	10	0	70
56	50	105	,293	30	51	80
57	34	280	,297	39	0	63
58	246	249	,301	0	0	259
59	82	133	,304	0	23	102
60	80	245	,308	0	0	174
61	48	60	,308	54	41	63
62	172	186	,315	14	0	73
63	34	48	,318	57	61	64
64	34	194	,320	63	0	67
65	65	100	,327	38	35	67
66	8	81	,328	52	32	68
67	34	65	,331	64	65	76
68	8	58	,331	66	0	77
69	147	207	,334	0	0	109
70	59	69	,337	55	46	114
71	110	167	,340	47	0	86
72	43	127	,342	50	0	88

73	53	172	,343	0	62	95
74	31	119	,346	53	25	85
75	6	254	,351	34	0	116
76	34	37	,356	67	0	78
77	8	33	,360	68	0	100
78	13	34	,361	28	76	80
79	179	274	,362	37	0	136
80	13	50	,363	78	56	81
81	13	21	,367	80	0	83
82	52	87	,369	0	0	169
83	13	251	,371	81	0	84
84	13	201	,373	83	0	87
85	31	134	,386	74	0	94
86	98	110	,386	0	71	90
87	13	161	,389	84	0	92
88	43	266	,390	72	0	120
89	107	146	,392	0	0	96
90	98	124	,394	86	0	94
91	144	240	,398	0	0	178
92	12	13	,399	0	87	95
93	41	225	,400	0	0	168
94	31	98	,402	85	90	109
95	12	53	,403	92	73	97
96	107	237	,403	89	0	146
97	12	287	,408	95	0	101
98	32	138	,412	0	16	110
99	3	25	,413	0	0	111
100	8	57	,413	77	0	104
101	12	213	,415	97	0	103
102	82	224	,420	59	0	173
103	12	141	,427	101	0	106
104	8	79	,428	100	0	106
105	187	235	,428	0	0	190
106	8	12	,433	104	103	107
107	8	174	,437	106	0	113
108	29	97	,439	0	0	142
109	31	147	,440	94	69	114
110	32	159	,441	98	17	113
111	3	66	,445	99	0	232
112	15	199	,446	0	0	159
113	8	32	,452	107	110	115

114	31	59	,456	109	70	120
115	8	23	,458	113	9	116
116	6	8	,461	75	115	121
117	16	109	,470	0	0	128
118	96	209	,483	0	0	153
119	17	171	,488	0	0	130
120	31	43	,493	114	88	122
121	6	185	,502	116	0	122
122	6	31	,509	121	120	124
123	11	38	,521	22	0	124
124	6	11	,523	122	123	125
125	6	175	,524	124	0	127
126	19	101	,526	0	0	202
127	6	63	,526	125	0	129
128	16	89	,528	117	0	166
129	6	30	,531	127	45	130
130	6	17	,534	129	119	131
131	6	211	,534	130	0	132
132	6	113	,536	131	0	133
133	5	6	,537	0	132	140
134	27	203	,542	0	0	147
135	154	238	,543	0	0	181
136	179	233	,545	79	0	200
137	55	122	,551	0	0	158
138	182	243	,552	0	0	143
139	49	156	,557	0	49	162
140	5	46	,558	133	0	144
141	28	206	,558	0	0	228
142	29	295	,558	108	0	237
143	1	182	,562	0	138	151
144	5	70	,562	140	0	145
145	5	155	,569	144	0	147
146	107	292	,577	96	0	196
147	5	27	,578	145	134	149
148	131	214	,587	0	0	167
149	5	121	,591	147	0	152
150	2	301	,592	0	0	298
151	1	265	,592	143	0	152
152	1	5	,596	151	149	157
153	96	217	,601	118	0	180
154	24	300	,605	0	0	164

155	176	212	,610	15	0	163
156	62	83	,618	0	0	194
157	1	231	,623	152	0	158
158	1	55	,624	157	137	161
159	15	189	,626	112	0	209
160	61	130	,627	0	0	251
161	1	74	,630	158	0	162
162	1	49	,631	161	139	163
163	1	176	,632	162	155	164
164	1	24	,634	163	154	167
165	173	242	,638	0	0	197
166	16	118	,641	128	0	185
167	1	131	,643	164	148	171
168	7	41	,643	0	93	190
169	52	261	,651	82	0	182
170	153	282	,653	0	0	234
171	1	169	,657	167	0	173
172	84	193	,657	0	0	241
173	1	82	,662	171	102	175
174	80	229	,665	60	0	186
175	1	125	,671	173	0	176
176	1	4	,688	175	0	178
177	188	267	,696	0	0	235
178	1	144	,698	176	91	180
179	35	226	,699	0	0	205
180	1	96	,700	178	153	181
181	1	154	,700	180	135	182
182	1	52	,708	181	169	183
183	1	9	,723	182	0	184
184	1	95	,730	183	0	187
185	16	223	,735	166	0	193
186	80	180	,735	174	0	189
187	1	20	,743	184	0	188
188	1	195	,749	187	0	191
189	72	80	,750	0	186	192
190	7	187	,757	168	105	214
191	1	227	,759	188	0	192
192	1	72	,760	191	189	193
193	1	16	,763	192	185	195
194	62	250	,772	156	0	207
195	1	103	,776	193	0	196

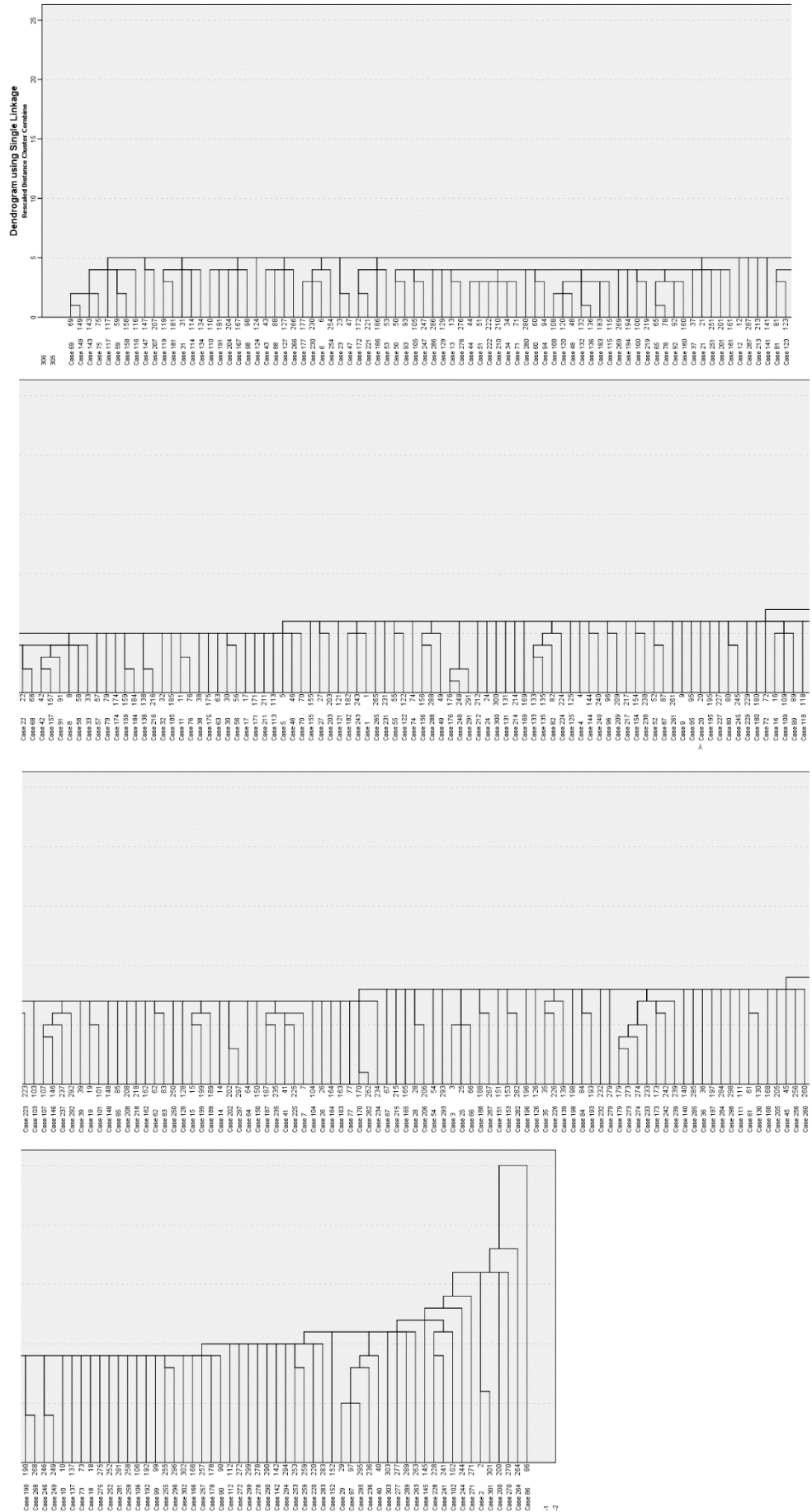
196	1	107	,783	195	146	199
197	173	239	,784	165	0	200
198	85	208	,786	0	0	201
199	1	39	,789	196	0	202
200	173	179	,790	197	136	222
201	85	218	,805	198	0	204
202	1	19	,809	199	126	203
203	1	148	,813	202	0	204
204	1	85	,826	203	201	206
205	35	139	,827	179	0	226
206	1	162	,881	204	0	207
207	1	62	,889	206	194	208
208	1	128	,893	207	0	209
209	1	15	,895	208	159	210
210	1	14	,903	209	0	212
211	64	150	,909	0	0	213
212	1	202	,919	210	36	213
213	1	64	,921	212	211	214
214	1	7	,946	213	190	215
215	1	104	,952	214	0	216
216	1	26	,958	215	0	217
217	1	164	,965	216	0	218
218	1	163	,968	217	0	219
219	1	77	,995	218	0	220
220	1	170	,995	219	5	221
221	1	234	1,000	220	0	224
222	140	173	1,008	0	200	233
223	168	205	1,010	0	0	252
224	1	67	1,019	221	0	225
225	1	215	1,028	224	0	227
226	35	198	1,046	205	0	240
227	1	165	1,049	225	0	228
228	1	28	1,056	227	141	230
229	232	279	1,059	0	0	242
230	1	54	1,061	228	0	231
231	1	293	1,065	230	0	232
232	1	3	1,076	231	111	235
233	140	285	1,079	222	0	243
234	153	196	1,110	170	0	238
235	1	188	1,111	232	177	236
236	1	151	1,121	235	0	238

237	29	236	1,138	142	0	262
238	1	153	1,141	236	234	239
239	1	126	1,148	238	0	240
240	1	35	1,151	239	226	241
241	1	84	1,154	240	172	242
242	1	232	1,162	241	229	243
243	1	140	1,162	242	233	244
244	1	36	1,166	243	0	245
245	1	197	1,174	244	0	246
246	1	284	1,189	245	0	247
247	1	298	1,190	246	0	250
248	255	296	1,203	0	0	271
249	253	259	1,211	0	0	284
250	1	111	1,220	247	0	251
251	1	61	1,222	250	160	252
252	1	168	1,226	251	223	253
253	1	45	1,244	252	0	254
254	1	256	1,250	253	0	255
255	1	260	1,259	254	0	256
256	1	190	1,290	255	43	259
257	18	275	1,290	0	0	264
258	166	257	1,305	0	0	274
259	1	246	1,308	256	58	260
260	1	10	1,323	259	0	261
261	1	137	1,331	260	0	263
262	29	40	1,345	237	0	289
263	1	73	1,379	261	0	264
264	1	18	1,387	263	257	265
265	1	252	1,438	264	0	266
266	1	281	1,449	265	0	267
267	1	258	1,471	266	0	268
268	1	106	1,504	267	0	269
269	1	192	1,506	268	0	270
270	1	99	1,507	269	0	272
271	255	302	1,510	248	0	272
272	1	255	1,516	270	271	274
273	228	241	1,530	0	0	288
274	1	166	1,543	272	258	275
275	1	178	1,543	274	0	276
276	1	90	1,559	275	0	277
277	1	112	1,606	276	0	278

278	1	272	1,612	277	0	279
279	1	299	1,661	278	0	280
280	1	278	1,665	279	0	281
281	1	290	1,686	280	0	282
282	1	142	1,715	281	0	283
283	1	294	1,763	282	0	284
284	1	253	1,806	283	249	285
285	1	220	1,831	284	0	286
286	1	283	1,888	285	0	287
287	1	152	1,950	286	0	289
288	102	228	1,951	0	273	295
289	1	29	1,970	287	262	290
290	1	303	2,008	289	0	291
291	1	277	2,102	290	0	292
292	1	289	2,154	291	0	293
293	1	263	2,214	292	0	294
294	1	145	2,282	293	0	295
295	1	102	2,591	294	288	296
296	1	244	2,786	295	0	297
297	1	271	3,183	296	0	298
298	1	2	4,326	297	150	299
299	1	200	4,388	298	0	300
300	1	270	4,489	299	0	301
301	1	264	5,445	300	0	302
302	1	86	10,672	301	0	0

Στο δένδρογραμμα παρουσιάζεται η σειρά των συνδέσεων καθώς και οι αποστάσεις στις οποίες ενώνονται οι ομάδες. Παρατηρήσεις που συνδέονται μεταξύ τους στην αριστερή πλευρά του δένδρογράμματος είναι παρόμοιες και άρα μπορούν να ταξινομηθούν στην ίδια ομάδα. Παρατηρήσεις που συνδέονται μεταξύ τους στη δεξιά πλευρά του δένδρογράμματος, θεωρούνται πιο ανομοιογενείς.

Στην περίπτωση μας, παρατηρούμε ότι ο βέλτιστος διαχωρισμός επιτυγχάνεται με τον σχηματισμό 5 ομάδων. Πιο συγκεκριμένα, η πρώτη ομάδα θα μπορούσε να περιλαμβάνει από την παρατήρηση 69 ως την 218, η δεύτερη ομάδα από την παρατήρηση 162 ως την 268, η τρίτη από την 246 ως την 142, η τέταρτη από την 294 ως την 145 και η πέμπτη ομάδα από την παρατήρηση 228 ως την 86. Η ταξινόμηση αυτή, συμπίπτει με την ήδη γνωστή ταξινόμηση των ασθενών σε 5 ομάδες κινδύνου για Καρδιαγγειακή Νόσο και άρα θεωρείται αξιόπιστη. Το δένδρογραμμα της ανάλυσης μας παρατίθεται παρακάτω.



Ε. ΣΥΜΠΕΡΑΣΜΑ

Η παρουσίαση των βασικών θεωρητικών στοιχείων των τριών μεθόδων πολυμεταβλητής ανάλυσης, σε συνδυασμό με την εφαρμογή τους σε πρακτικά παραδείγματα ανέδειξαν τα πλεονεκτήματα και μειονεκτήματα της κάθε μεθόδου, καθώς και τις ενδείξεις χρήσης τους.

Η Γραμμική Διαχωριστική Ανάλυση (LDA) είναι μια επιτηρούμενη τεχνική μείωσης διαστάσεων δεδομένων που απαιτεί τον εκ των προτέρων καθορισμό των κλάσεων, γεγονός που την διαφοροποιεί από τις άλλες δύο τεχνικές. Χρησιμοποιείται για τον έλεγχο της ορθής ταξινόμησης πολυμεταβλητών δεδομένων σε ομάδες, καθώς και για την ταξινόμηση νέων παρατηρήσεων σε ήδη γνωστές κλάσεις. Ακόμη, συμβάλλει στην αξιολόγηση της σπουδαιότητας κάθε μεταβλητής στον διαχωρισμό των ομάδων. Χρησιμοποιείται για ταξινόμηση με βάση τα δεδομένα (data classification), σε αντίθεση με την Ανάλυση Κύριων Συνιστωσών (PCA) που κάνει ταξινόμηση βάσει μιας ιδιότητας (feature classification).

Από την άλλη πλευρά, η PCA είναι μια μη επιτηρούμενη τεχνική μείωσης διαστάσεων δεδομένων, η οποία συνήθως χρησιμοποιείται για να επιτύχουμε γραμμικούς συνδυασμούς μικρότερης διάστασης, ασυσχέτιστων μεταξύ τους, στην πορεία μιας πιο σύνθετης ανάλυσης. Μπορεί να χρησιμοποιηθεί ως πρώτο βήμα πριν να εφαρμόσουμε την LDA. Η μέθοδος αυτή, προκαλεί αλλαγή του σχήματος και της θέσης των αρχικών μεταβλητών όταν τους μετασχηματίζει σε μικρότερο διαστατικό χώρο, σε αντίθεση με την LDA που δεν μεταβάλλει την θέση των μεταβλητών, παρά μόνο επιτυγχάνει την μέγιστη διαχωριστικότητα μεταξύ των ομάδων. Βασικό πλεονέκτημα της μεθόδου, αποτελεί η συμβολή της τόσο στην εξακρίβωση της ύπαρξης ομάδων με κοινά χαρακτηριστικά σε ένα σύνολο δεδομένων, όσο και η ανάδειξη του βαθμού επίδρασης των μεταβλητών στο υπό διερεύνηση ερώτημα.

Τέλος, η Ανάλυση Συστάδων (CA) χρησιμοποιείται αποκλειστικά και μόνο για την ανάδειξη του βέλτιστου τρόπου διαχωρισμού ενός συνόλου παρατηρήσεων σε ομάδες. Κατά αντιστοιχία με την PCA, δεν απαιτεί την εκ των προτέρων γνώση της ύπαρξης ή του αριθμού ομάδων (εξαίρεση αποτελούν οι μέθοδοι μη ιεραρχικής ανάλυσης). Ακόμη, η CA ομαδοποιεί παρατηρήσεις σε κλάσεις που έχουν παρόμοια χαρακτηριστικά, ενώ η PCA ομαδοποιεί μεταβλητές σε συνιστώσες, βασιζόμενη στο ότι αυτές έχουν παρόμοιες σχέσεις με λανθάνουσες μεταβλητές.

Συμπερασματικά, πρόκειται για τρεις διαφορετικές μεθόδους, που εφαρμόζονται με διαφορετικό στόχο, αναλύουν με διαφορετικό τρόπο τα υπό θεώρηση δεδομένα και παράγουν αποτελέσματα που απαντούν σε ετερογενή ερευνητικά ερωτήματα.

ΣΤ. ΑΝΑΦΟΡΕΣ

- [1] Ηλιοπούλου, Πολυξένη. 2016. Γεωγραφική Ανάλυση. Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών: 179-213. <http://hdl.handle.net/11419/2060> (Accessed 2019-7-20).
- [2] Κύρκος, Ευστάθιος. 2015. Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων. Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών: 261-287. <http://hdl.handle.net/11419/1226> (Accessed 2019-7-28).
- [3] Νικήτας, Παναγιώτης. 2013. Εισαγωγή στη στατιστική ανάλυση πειραματικών δεδομένων με χρήση EXCEL και SPSS. Θεσσαλονίκη: Εκδόσεις ΣΙΜΩΝΗ.
- [4] Πετρίδης, Δημήτριος. 2015. Ανάλυση Πολυμεταβλητών Τεχνικών, Εφαρμογές Περιπτώσεων. Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών: 126-157.
- [5] Abdi H., Williams L. 2010. Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2: 433-459. doi: 10.1002/wics.10.
- [6] Balakrishnama S., Ganapathiraju A. 1998. Linear Discriminant Analysis: A Brief Tutorial: Department of Electrical and Computer Engineering, Mississippi State University.
- [7] Härdle W., Simar L. 2003. Applied Multivariate Statistical Analysis. Berlin: <http://www.xplore-stat.de> (Accessed 2019-07-17).
- [8] Lichman, Moshe. 2013. UCI Machine Learning Repository [<https://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
- [9] Manly B., Navarro Alberto J. 2017. Multivariate statistical methods: a primer. 4th ed. Boca Raton: CRC Press.
- [10] Shiker, Mustak. 2012. Multivariate Statistical Analysis. British Journal of Science 6 (1): 55-66.
- [11] Tharwat A., Gaber T., Ibrahim A., Hassanien A.E. 2017. Linear discriminant analysis: A detailed tutorial. Ai Communications 30: 169-190. doi: 10.3233/AIC-170729.
- [12] V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
- [13] Arifin, Wan Nor. (2015). The Graphical Assessment of Multivariate Normality Using SPSS. Education in Medicine Journal. 7. 10.5959/eimj.v7i2.361.
- [14] Li, T., Zhu, S. & Ogihara, M. (2006). Using Discriminant Analysis for Multi-Class Classification: An Experimental Investigation. Knowledge and Information Systems 10: 453. <https://doi.org/10.1007/s10115-006-0013-y>