

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΙΑ ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΑΝΑΛΥΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΑΠΟ ΤΗΝ ΕΚΠΑΙΔΕΥΣΗ

Αντώνιος Γεωρ. Σαρρής

ΕΡΓΑΣΙΑ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου ΑΘηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος
Συμπληρωματικής Ειδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time)

Αθήνα
Φεβρουάριος 2012

ΑΦΙΕΡΩΣΗ

Στη σύζυγό μου Μαρίνα
και στα παιδιά μου Γιώργο και Βιβή

ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ την επιβλέπουσα καθηγήτριά μου Δρ Ιουλία Παπαγεωργίου για την καθοδήγησή της και την πολύτιμη βοήθειά της σε κάθε δυσκολία που μου προέκυψε.

Ευχαριστώ τον κο Σωκράτη Μπατσιούδη Διευθυντή του 1ου Γενικού Λυκείου Ασπροπύργου και προϊστάμενο μου, για τα στατιστικά στοιχεία που αφορούν στους μαθητές του εν λόγω Λυκείου και τα οποία πρόθυμα μου παρείχε.

Ευχαριστώ την κα Γυφτάκη Βασιλική καθηγήτρια Φιλόλογο για τις ιστορικές πληροφορίες.

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Είμαι πτυχιούχος του Μαθηματικού τμήματος της Σχολής Θετικών Επιστημών του Πανεπιστημίου Κρήτης από το 1982.

Μετά την εκπλήρωση των στρατιωτικών μου υποχρεώσεων υπηρέτησα από το (1985-1995) ως μόνιμος δημόσιος υπάλληλων στο τότε Υπουργείο Δημόσιας Τάξης με εξειδίκευση στην πληροφορική. Στο διάστημα αυτό παρακολούθησα σεμινάρια για θέματα πληροφορικής και Δημόσιας διοίκησης. Παράλληλα επιμορφώθηκα στο Περιφερειακό Επιμορφωτικό Κέντρο Λαμίας σε εκπαιδευτικά θέματα αφού η εκπαίδευση ήταν πάντα ο στόχος μου.

Το 1995 διορίστηκα ως εκπαιδευτικός στη Δευτεροβάθμια Εκπάιδευση στο Γυμνάσιο Σπαρτύλα Κέρκυρας. Στη συνέχεια δίδαξα σε σχολεία της Βοιωτίας και της Αττικής. Η τελευταία οργανική θέση στην οποία υπηρετώ είναι το 1^{ο} Γενικό Λύκειο Ασπροπύργου.

Τα πρώτα βήματα μου στη Στατιστική έγιναν στα φοιτητικά μου χρόνια με εργασία που πραγματοποίησα πάνω σε δεδομένα πωλήσεων αυτοκινήτων τα οποία πρόθυμα μου παρείχε γερμανική αυτοκινητοβιομηχανία από την οποία τα ζήτησα. Ως εκπαιδευτικός διαπίστωσα την όλο αυξανόμενη εισαγωγή της Στατιστικής σαν ιδιαίτερο τομέα των Μαθηματικών και ειδικότερα στην οικονομική κατεύθυνση του λυκείου. Αυτός ήταν και ο λόγος για τον οποίο επέλεξα το συγκεκριμένο Μεταπτυχιακό Πρόγραμμα.

ABSTRACT

Antonios Sarris

AN INTRODUCTION TO MULTIRIVATIVE ANALYSIS WITH DATA FROM EDUCATION

January 2012

The aim of this study is to focus on Multivariate Analysis texts and more specifically the Discriminant Analysis and Correspondence Analysis. For this purpose, it was considered advisable to begin with the development of Cluster Analysis as the later is necessary to understand the distinctions. Along with the theory, examples are studied as application of the theory.

In application of the above, two examples are analyzed with real data. The data is drawn from the first Senior School of Aspropyrgos. This school, due to the recent repatriation of the former USSR Pontians and the increase of economic migrants consists of a number of students which could be categorized into Greek children, Repatriates and foreigners.

The research is focused on answering two questions. The first question relates to the performance of pupils per class and subject. As regards the second question, the selection of pupils per class and study direction is investigated. After applying the Discriminant and Correspondence to the questions above respectively, an evaluation is expected as to whether the results can eventually be used for more general conclusions

ΠΕΡΙΛΗΨΗ

Αντώνιος Σαρρής

ΜΙΑ ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΑΝΑΛΥΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΑΠΟ ΤΗΝ ΕΚΠΑΙΔΕΥΣΗ

Ιανουάριος 2012

Η παρούσα μελέτη έχει ως αντικείμενο την Πολυμεταβλητή Ανάλυση και ειδικότερα την Διαχωριστική Ανάλυση και την Correspondence Analysis. Θεωρήθηκε σκόπιμο να προηγηθεί μια ανάπτυξη της Cluster Analysis καθώς αυτή είναι απαραίτητη για την κατανόηση της Διαχωριστικής. Παράλληλα με την θεωρία παρατίθενται παραδείγματα στα οποία εφαρμόζεται η θεωρία.
Σε εφαρμογή των παραπάνω αναλύονται δύο παραδείγματα με πραγματικά δεδομένα. Τα δεδομένα αντλούνται από το 1^{ο} Γενικό Λύκειο Ασπροπύργου. Το Λύκειο αυτό λόγω της πρόσφατης παλινόστησης των Ποντίων της πρώην ΕΣΣΔ αλλά και της αύξησης των οικονομικών μεταναστών αποτελείται από ένα πλήθος μαθητών που θα μπορούσε να κατηγοριοποιηθεί σε Ελληνόπαιδες, Παλινοστούντες και Αλλοδαπούς.
Η προσπάθεια της έρευνας εστιάζεται στο να απαντηθούν δύο ερωτήματα. Το πρώτο ερώτημα αφορά στην επίδοση των μαθητών ανά κατηγορία και μάθημα. Στο δεύτερο ερώτημα διερευνάται η επιλογή των μαθητών ανά κατηγορία και κατεύθυνση σπουδών. Αφού εφαρμοσθεί η Διαχωριστική και η Correspondence στα παραπάνω ερωτήματα αντίστοιχα, αναμένεται και η αξιολόγηση του κατά πόσον τελικά μπορεί με αξιοπιστία να χρησιμοποιηθούν τα αποτελέσματα για γενικότερα συμπεράσματα.

ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

	Σελ.
ΕΙΣΑΓΩΓΗ	1
ΚΕΦΑΛΑΙΟ 1,ΠΕΡΙΓΡΑΦΗ	3
1. Cluster Ανάλυση για μεικτού τύπου δεδομένα.	3
1.1.Το Πρόβλημα.	3
1.2 Η εγγύτητα μεταξύ των αντικειμένων	4
1.3. Ομοιότητα των αντικειμένων με δυαδική δομή	5
1.4.Μέτρα απόστασης για συνεχείς μεταβλητές	8
1.5. Αλγόριθμοι Cluster	10
1.5.1 Ιεραρχικοί Αλγόριθμοι, Agglomerative Τεχνικές	11
ΚΕΦΑΛΑΙΟ 2, ΠΕΡΙΓΡΑΦΗ.	13
2.Διαχωριστική ανάλυση (Discriminant Analysis)	13
2.1 Κανόνες κατανομής για γνωστές κατανομές	14
2.2.Η γραμμική διαχωριστική ανάλυση (LDA) και η γραμμική συνάρτηση διαχωρισμού Fisher	15
2.2.1 Η γραμμική διαχωριστική ανάλυση (LDA)	15
2.2.2 Η γραμμική διαχωριστική ανάλυση (LDA) για δύο κλάσεις.	16
2.2.3 Η γραμμική συνάρτηση διαχωρισμού Fisher	17
2.2.4 Η τετραγωνική διαχωριστική ανάλυση (Quadratic Discriminant Analysis)	19
2.3. Κανόνας μέγιστης πιθανοφάνειας (ML).	19
2.3.1. Παραδείγματα εκτιμήσεων μέγιστης πιθανοφάνειας.	21
2.4. Ο Κανόνας διαχωρισμού Bayes.	28
2.4.1 Πιθανότητα λανθασμένης ταξινόμησης με ίσους πίνακες συνδιακύμανσης για τον κανόνα ML (J=2).	29
2.4.2. Ταξινόμηση με διαφορετικούς πίνακες συνδιακύμανσης	30
2.5 Οι διαχωριστικοί κανόνες στην πράξη.	30
2.5.1 Εκτίμηση πιθανοτήτων εσφαλμένης ταξινόμησης	31
ΚΕΦΑΛΑΙΟ 3, ΠΕΡΙΓΡΑΦΗ	35
3 Correspondence Analysis	35
3.1 Κίνητρα	36
3.2 Αποσύνθεση του X^2 (X^2- Decomposition)	38
3.3 Correspondence Analysis in Practice	43
3.4 Πρακτικοί υπολογισμοί-Ορισμός inertia	48
3.5 Biplots	49
ΚΕΦΑΛΑΙΟ 4	57
Εφαρμογή της διαχωριστικής ανάλυσης και της correspondence analysis σε δεδομένα που προκύπτουν από την εκπαιδευση	57
Συμπεράσματα της διαχωριστικής ανάλυσης.	67
Πίνακες, διαγράμματα και συμπεράσματα για την correspondence analysis	67
Συμπεράσματα της correspondence analysis.	75
ΠΑΡΑΤΗΜΑ Α	77
Βασικές έννοιες Γραμμικής Άλγεβρας και Αναλυτικής Γεωμετρίας	77
ΒΙΒΛΙΟΓΡΑΦΙΑ	81
ΔΙΑΔΥΚΤΙΑΚΟΙ ΤΟΠΟΙ	82

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Σελ.

Πίνακας 1.1 Μέτρα ομοιότητας. 6
Πίνακας 1.2 Τιμές μεταβλητών τύπων αυτοκινήτων. 7
Πίνακας 2.1 Κόστος ¨κακής" αξιολόγησης πιστοληπτικής ικανότητας 23
Πίνακας 2.2 Ταξινόμηση ελβετικών χαρτονομισμάτων 32
Πίνακας 3.1 Δεδομένα για την ανάγνωση εφημερίδων 45
Πίνακας 3.2. Ιδιοτιμές και αποτελέσματα Correspondence Analysis των δεδομένων 46
Πίνακας 3.3 Απόλυτη συνεισφορά των παραγόντων της γραμμής r_k. 47
Πίνακας 3.4 Πίνακας συνάφειας συνηθειών καπνίσματος. 51
Πίνακας 3.5 Παρατηρούμενες συχνότητες καπνίσματος ανά κατηγορία υπαλλήλων. 52
Πίνακας 3.6 Παρατηρούμενες συχνότητες ανά κατηγορία καπνίσματος και συνολικές ανά γραμμή. 52
Πίνακας 3.7 Αποτελέσματα Correspondence Analysis. 53
Πίνακας 3.7.1 Αποτελέσματα ανά κατηγορία υπαλλήλων. 54
Πίνακας 3.7.2 Αποτελέσματα ανά κατηγορία καπνίσματος. 54
Πίνακας 4.1 Πλήθος δείγματος. 59
Πίνακας 4.2 Μέσοι.ανά κατηγορία και μάθημα. 59
Πίνακας 4.3 Πίνακας ελέγχου ισότητας μέσων. 60
Πίνακας 4.4 Πίνακας βαθμών και λογάριθμοι των οριζουσών των πινάκων διακύμανσης. 60
Πίνακας 4.5 Πίνακας ελέγχου ισότητας διακυμάνσεων 60
Πίνακας 4.6 Πίνακας ιδιοτιμών των δύο πρώτων διαχωριστικών συναρτήσεων. 61
Πίνακας 4.7 Πίνακας Wilks' Lambda. 61
Πίνακας 4.8 Συντελεστές συναρτήσεων 63
Πίνακας 4.9 Συντελεστές συναρτήσεων 63
Πίνακας 4.10 Μέσος των Scores 63
Πίνακας 4.11 Πίνακας αποτελεσμάτων ομαδοποίησης 63
Πίνακας 4.12 Περιγραφικά στοιχεία Βαθμού Μαθηματικών και DS σε DF_1 DF_2 65
Πίνακας 4.13 Test ομοιογένειας διακυμάνσεων 65
Πίνακας 4.14 Πίνακας ANOVA για Βαθμό Μαθηματικών 65
Πίνακας 4.15. Πίνακας συνάφειας ανά κατηγορία και κατεύθυνση σπουδών. 67
Πίνακας 4.16 Πίνακας ποσοστών ανά γραμμή 68
Πίνακας 4.17 Πίνακας ποσοστών ανά στήλη 68
Πίνακας 4.18 Συνοπτικός πίνακας Correspondence Analysis 68
Πίνακας 4.19 Πίνακας αξιολόγησης ανά γραμμή 69
Πίνακας 4.20 Πίνακας αξιολόγησης ανά στήλη 69
Πίνακας 4.21 Πίνακας τυπικής απόκλισης ανά γραμμή 70
Πίνακας 4.22 Πίνακας τυπικής απόκλισης ανά στήλη 70

ΚΑΤΑΛΟΓΟΣ ΓΡΑΦΗΜΑΤΩΝ

Σελ.

Γράφημα 2.1 Περίπτωση 2 κανονικών πληθυσμών και ML μέθοδος εύρεσης διαχωριστικής ανάλυσης. 26

Γράφημα 3.1 Γραφική παράσταση των κατηγοριών των εφημερίδων στις δυο πρώτες διαστάσεις. 46

Διάγραμμα 3.1 Διαγραμματική απεικόνιση Π.3.7.1 σε D1 54

Διάγραμμα 3.2 Διαγραμματική απεικόνιση Π.3.7.1 σε D2 54

Διάγραμμα 3.3 Διαγραμματική απεικόνιση Π.3.7.2 σε D1 55

Διάγραμμα 3.4 Διαγραμματική απεικόνιση Π.3.7.2 σε D2 55

Διάγραμμα 3.5 Διαγραμματική απεικόνιση Κατηγορίας Υπαλλήλων σε D1 D2 55

Διάγραμμα 3.6 Διαγραμματική απεικόνιση Κατηγορίας Καπνίσματος σε D1 D2 55

Διάγραμμα 4.1. Canonical Discriminant Functions 62

Διάγραμμα 4.2. Διαγραμματική απεικόνιση των scores DF_1 DF_2. 64

Διάγραμμα 4.3 Μέσοι ανά κατηγορία 66

Διάγραμμα 4.4 Μέσοι DF_1 ανά κατηγορία 66

Διάγραμμα 4.5 Μέσοι DF_2 ανά κατηγορία 66

Διάγραμμα 4.6 Διαγραμματική απεικόνιση D1 ανά κατηγορία 71

Διάγραμμα 4.7 Διαγραμματική απεικόνιση D2 ανά κατηγορία 71

Διάγραμμα 4.8 Διαγραμματική απεικόνιση D1 ανά κατεύθυνση σπουδών 72

Διάγραμμα 4.9 Διαγραμματική απεικόνιση D2 ανά κατεύθυνση σπουδών 72

Διάγραμμα 4.10 Διαγραμματική απεικόνιση των Scores ανά κατηγορία 73

Διάγραμμα 4.11 Διαγραμματική απεικόνιση των Scores ανά κατεύθυνση 73

Διάγραμμα 4.12. Biplot 74

"πρὸς γὰρ τὸ τελευταῖον ἐκβὰν ἕκαστον τῶν πρὶν ὑπαρξάντων κρίνεται"
Δημοσθένης 384-322 π.Χ Ὀλυνθιακός Α΄

ΕΙΣΑΓΩΓΗ

Το μεγάλο κύμα μετανάστευσης-επαναπατρισμού που δέχθηκε η Ελλάδα τα τελευταία 20 περίπου χρόνια δημιούργησε ανάγκες για οικονομικές, πολιτιστικές, κοινωνικές αλλά και εκπαιδευτικές προσαρμογές στην Ελληνική κοινωνία.

Υπάρχουν σχολεία στα οποία φοιτούν μαθητές Ελληνικής καταγωγής αλλά και παλιννοστούντες Ελληνικής καταγωγής, κυρίως από χώρες της πρώην Σοβιετικής Ένωσης.

Τα περισσότερα Δημόσια Ελληνικά σχολεία όλων των βαθμίδων έχουν μεικτό πληθυσμό.

Σε ένα τέτοιο πληθυσμό θα ήταν ενδιαφέρον να εξετάσουμε τις διαφορές όπως:

- Στις επιδόσεις
- Στις επιλογές κατεύθυνσης σπουδών
- Στα αποτελέσματα των πανελλαδικών εξετάσεων, αλλά και σε κοινωνικές διαφορές όπως:
    - Το επάγγελμα των γονιών
    - Τις ημέρες των διακοπών τους κλπ.

Τα δεδομένα έχουν προέλθει από τα αρχεία του $1^{ου}$ Γενικού Λυκείου Ασπροπύργου, δεδομένου ότι ο Ασπρόπυργος είναι μια περιοχή της Αττικής με πολύ μεγάλο αριθμό παλιννοστούντων κατοίκων.

Οι στατιστικές μέθοδοι που θα χρησιμοποιηθούν θα αφορούν κυρίως σε τεχνικές της Πολυμεταβλητής Στατιστικής Ανάλυσης (Multivariate Techniques) και αναλυτικά είναι:

A. **Cluster Analysis** για μεικτού τύπου δεδομένα
B. **Discriminant Analysis (Διαχωριστική Ανάλυση)**
C. **Correspondense Analysis ή** X^2 **decomposition.**

Η παρούσα μελέτη περιέχει 4 Κεφάλαια.

Στο $1^{ο}$ Κεφάλαιο περιέχεται η αναλυτική περιγραφή της **Cluster Analysis** για μεικτού τύπου δεδομένα καθώς και τα σχετικά παραδείγματα ώστε να γίνει πιο

κατανοητή η θεωρία. Με την Cluster Analysis αναπτύσσονται εργαλεία και μέθοδοι με σκοπό τη διερεύνηση ύπαρξης ή μη φυσικών ομάδων ή τάξεων ατόμων σε ένα μεγάλο αριθμό ατόμων.

Η **Διαχωριστική Ανάλυση (Discriminant Analysis)** περιγράφεται αναλυτικά στο 2^{ο} Κεφάλαιο με τα σχετικά παραδείγματα. Σκοπός της Διαχωριστικής ανάλυσης είναι η ταξινόμηση ή μη μιας ή διαφόρων παρατηρήσεων σε ήδη γνωστές ομάδες (clusters) του πληθυσμού. Σε αντίθεση με την Cluster στην οποία διερευνάτε η.ύπαρξη ομάδων εδώ οι ομάδες είναι εκ των προτέρων γνωστές. Με την Διαχωριστική Ανάλυση σε άλλο Κεφάλαιο της έρευνας μελετάται η δυνατότητα διαχωρισμού των μαθητών του 1^{ου} ΓΕΛ Ασπροπύργου τρείς κατηγορίες σε σχέση με την βαθμολογία τους.

Στο 3^{ο} Κεφάλαιο περιέχεται η αναλυτική περιγραφή καθώς και τα σχετικά παραδείγματα της **Correspondence Analysis.** Η Correspondence Analysis χρησιμοποιείται όταν διερευνάται η σχέση μεταξύ των γραμμών και των στηλών ενός πίνακα συνάφειας. Πίνακας συνάφειας είναι ένας πίνακας συχνοτήτων δύο εισόδων όπου αναφέρονται οι κοινές συχνότητες δύο ποιοτικών μεταβλητών. Στην συγκεκριμένη έρευνα η μέθοδος αυτή εφαρμόζεται για να μελετηθεί η σχέση των κατηγοριών των μαθητών με την κατεύθυνση σπουδών τους.

Στο 4^{ο} Κεφάλαιο γίνεται εφαρμογή της Διαχωριστικής και της Correspondence Analysis σε δεδομένα που προκύπτουν από την εκπαίδευση. Σε κάθε περίπτωση αναφέρονται τα σχετικά συμπεράσματα.

Μαθηματικοί ορισμοί που απαιτούνται για την κατανόηση της μελέτης παρατίθενται σε σχετικό παράρτημα.

ΚΕΦΑΛΑΙΟ 1

ΠΕΡΙΓΡΑΦΗ

1. Cluster Ανάλυση για μεικτού τύπου δεδομένα.

Σύμφωνα με τους Hardle and Simar (κεφ.11**,** 2003), κατά την εξέταση ομάδων δεδομένων σε ένα πολυμεταβλητό σύνολο δεδομένων, δύο περιπτώσεις μπορούν να προκύψουν:

1. να δούμε αν υπάρχουν κάποιες φυσικές ομάδες ή τάξεις ατόμων (clusters), και
2. να ομαδοποιήσουμε τα άτομα σύμφωνα με ένα σύνολο από τις υπάρχουσες ομάδες (clusters).

Η Cluster ανάλυση αναπτύσσει εργαλεία και μεθόδους για την πρώτη περίπτωση, δηλαδή, δίνεται ένας πίνακας δεδομένων που περιέχει πολυμεταβλητές μετρήσεις σε ένα μεγάλο αριθμό ατόμων (ή αντικείμενων) και ο στόχος είναι να δημιουργήσουμε κάποιες φυσικές υποομάδες ή ομάδες ατόμων. Αυτό γίνεται με την ομαδοποίηση των ατόμων που είναι "παρόμοια" σύμφωνα με ορισμένα χαρακτηριστικά ή μεταβλητές στη γλώσσα της Στατιστικής.

Η Cluster ανάλυση εφαρμόζεται σε πολλούς τομείς, όπως οι φυσικές επιστήμες, οι ιατρικές επιστήμες, η οικονομία, το μάρκετινγκ, κλπ. Στο μάρκετινγκ, για παράδειγμα, είναι χρήσιμο να οικοδομήσουμε και να περιγράψουμε τα διαφορετικά τμήματα της αγοράς από μια έρευνα σχετικά με τους πιθανούς καταναλωτές. Μια ασφαλιστική εταιρεία, από την άλλη πλευρά, μπορεί να ενδιαφέρεται για τη διάκριση μεταξύ των κατηγοριών των δυνητικών πελατών, έτσι ώστε να μπορεί να αντλήσει βέλτιστες τιμές για τις υπηρεσίες της.

1.1 Το Πρόβλημα :

Ιδανικά σε μια καλή ομαδοποίηση θα προκύψουν clusters που να είναι όσο το δυνατόν πιο ομοιογενή στο εσωτερικό τους και όσο το δυνατό με μεγαλύτερη ανομοιογένεια μεταξύ τους .

Θα περιγράψουμε την ιεραρχική ομαδοποίηση (hierarchical clustering), μια μεγάλη κατηγορία ομαδοποίησης, ίσως η πιο δημοφιλής, όπου η ιεραρχία στην κατάταξη των στοιχείων σε ομάδες παραμένει αμετάβλητη στην πορεία της ομαδοποίησης.

Η ιεραρχική Cluster ανάλυση αποτελείται από εκτέλεση διαδοχικών βημάτων και τα δύο βασικά στάδια πριν την εφαρμογή της είναι:

1. **<u>Επιλογή του μέτρου εγγύτητας:</u>**

   Ελέγχει κάθε ζεύγος των παρατηρήσεων (αντικείμενα) για την «ομοιότητα» των αξιών τους.

   Ένα μέτρο ομοιότητας ορίζεται ως η μέτρηση της «εγγύτητας» των αντικειμένων. Όσο "πιο κοντά" είναι, τόσο πιο πιθανό να καταταγούν σύμφωνα με την τεχνική στην ίδια ομάδα.

2. **<u>Επιλογή του αλγορίθμου για την δημιουργία του cluster</u>:**

   Είναι το κριτήριο που η κάθε μέθοδος χρησιμοποιεί για να αποφανθεί εάν δύο ομάδες είναι κοντά και μπορούν να ενωθούν σε μια ή όχι.

1.2 Η εγγύτητα μεταξύ των αντικειμένων

Το σημείο εκκίνησης της cluster ανάλυσης είναι ένας πίνακας δεδομένων X (n × p) με n μετρήσεις (παρατηρήσεις) και p μεταβλητές. Η εγγύτητα (ομοιότητα) ανάμεσα σε παρατηρήσεις περιγράφεται από ένα τετράγωνο πίνακα D, διάστασης όσο το πλήθος των παρατηρήσεων, που δίνει την εγγύτητα για κάθε δυνατό ζεύγος. Σχηματικά είναι:

$$D = \begin{pmatrix} d_{11} & d_{12} & \dots & \dots & \dots & d_{1n} \\ \vdots & d_{22} & & & & \vdots \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & \dots & \dots & d_{nn} \end{pmatrix}$$

και ο υπολογισμός του θα φανεί χρήσιμος ως ένα κριτήριο του ποιες παρατηρήσεις θα μπορούσαν να ενωθούν.

Ο πίνακας D ισοδύναμα μπορεί να περιέχει και μέτρα ανομοιότητας μεταξύ των αντικειμένων n αντί για εγγύτητα. Εάν οι τιμές d_{ij} είναι αποστάσεις, τότε είναι μέτρο ανομοιότητας. Όσο μεγαλύτερη είναι η απόσταση, τόσο λιγότερο όμοιες είναι οι παρατηρήσεις.

Εάν οι τιμές d_{ij} είναι μέτρο εγγύτητας, τότε ισχύει το αντίθετο, δηλαδή, όσο μεγαλύτερη είναι η αξία της εγγύτητας, τόσο πιο όμοιες είναι οι παρατηρήσεις.

Προφανώς $d_{ii}=0 \ \forall i$ εάν D είναι πίνακας απόστασης και $d_{ii}=1 \ \forall i$ εάν D είναι πίνακας ομοιότητας και χρησιμοποιούμε [0,1] κλίμακα.

Ένας πίνακας αποστάσεων, για παράδειγμα, θα μπορούσε να οριστεί από το $L_2-norm: d_{ij}=\left\|x_i-x_j\right\|_2$ όπου x_i και x_j δηλώνουν τις γραμμές του πίνακα δεδομένων X.

Απόσταση και ομοιότητα είναι φυσικά αλληλένδετες έννοιες..

Εάν το d_{ij} είναι απόσταση τότε το $d'_{ij}=\max_j\left\{d_{ij}\right\}-d_{ij}$ είναι ένα μέτρο εγγύτητας.

Η φύση των παρατηρήσεων παίζει σημαντικό ρόλο στην επιλογή του μέτρου εγγύτητας.

Στις Nominal (ονομαστικές) κατηγορικές μεταβλητές (όπως δυαδικές) συνήθως χρησιμοποιούνται τιμές της εγγύτητας, ενώ στις μετρικές κατηγορικές μεταβλητές πίνακες αποστάσεων μιας και στην περίπτωση αυτή από πρακτικής καθαρά άποψης είναι ευκολότερος ο υπολογισμός της απόστασης με τη γνωστή έννοια που γνωρίζουμε από τη Γεωμετρία.

Παρουσιάζουμε κατ 'αρχήν τις δυνατότητες για τον πίνακα D στην περίπτωση των δυαδικών περιπτώσεων και στη συνέχεια θα εξετάσουμε τη συνεχή περίπτωση.

1.3. Ομοιότητα των αντικειμένων με δυαδική δομή

Σαν μέτρο ομοιότητας μεταξύ παρατηρήσεων συγκρίνουμε πάντα ζεύγη παρατηρήσεων, της μορφής (x_i, x_j) όπου $x_i^T=(x_{i1},...,x_{ip})$ και $x_j^T=(x_{j1},...,x_{jp})$ με $x_{ik}, x_{jk}\in\{0,1\}$.

Προφανώς σε μία σύγκριση υπάρχουν τέσσερις περιπτώσεις:

$$x_{ik}=x_{jk}=1$$
$$x_{ik}=0, x_{jk}=1$$
$$x_{ik}=1, x_{jk}=0$$
$$x_{ik}=x_{jk}=0$$

Τα περισσότερα μέτρα ομοιότητας υπολογίζονται καταμετρώντας για κάθε ζεύγος παρατηρήσεων τον αριθμό των αποτελεσμάτων αυτών των συγκρίσεων για κάθε μία από τις μεταβλητές.

Στον Πίνακα 1.1 εμφανίζονται κάποια μέτρα ομοιότητας τα οποία κάνουν χρήση της σύγκρισης μία προς μία των συνιστωσών 2 διανυσμάτων.

Δηλαδή αν $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ ορίζονται ως:

$$\alpha_1 = \sum_{k=1}^{p} I(x_{ik} = x_{jk} = 1)$$

$$\alpha_2 = \sum_{k=1}^{p} I(x_{ik} = 0, x_{jk} = 1)$$

$$\alpha_3 = \sum_{k=1}^{p} I(x_{ik} = 1, x_{jk} = 0)$$

$$\alpha_4 = \sum_{k=1}^{p} I(x_{ik} = x_{jk} = 0)$$

όπου κάθε a_l με $l = 1,2,3,4$ εξαρτάται από το ζεύγος (x_i, x_j), ορίζεται ο παρακάτω Πίνακας 1.1 ο οποίος περιέχει τα ακόλουθα μέτρα ομοιότητας.

Name	δ	λ	Definition
Jaccard	0	1	$\frac{a_1}{a_1 + a_2 + a_3}$
Tanimoto	1	2	$\frac{a_1 + a_4}{a_1 + 2(a_2 + a_3) + a_4}$
Simple Matching (M)	1	1	$\frac{a_1 + a_4}{p}$
Russel and Rao (RR)	–	–	$\frac{a_1}{p}$
Dice	0	0.5	$\frac{2a_1}{2a_1 + (a_2 + a_3)}$
Kulczynski	–	–	$\frac{a_1}{a_2 + a_3}$

Πίνακας 1.1 Μέτρα ομοιότητας.

Στην πράξη χρησιμοποιείται το ακόλουθο μέτρο ομοιότητας:

$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)}$ όπου δ και λ είναι συντελεστές στάθμισης.

Τα μέτρα αυτά παρέχουν εναλλακτικούς τρόπους mismatchings στάθμισης και θετικά (παρουσία ενός κοινού χαρακτήρα) ή αρνητικά (απουσία ενός κοινού χαρακτήρα) ταιριάσματα. Κατ αρχήν, θα μπορούσαμε να εξετάσουμε και την Ευκλείδειο απόσταση. Ωστόσο, το μειονέκτημα αυτής της απόστασης είναι ότι αντιμετωπίζει τις

παρατηρήσεις 0 και 1 με τον ίδιο τρόπο. Αν $x_{ik}=1$ σημαίνει για παράδειγμα τη γνώση μιας συγκεκριμένης γλώσσας, τότε το αντίθετο, $x_{ik}=0$ (χωρίς να γνωρίζει τη γλώσσα), θα αντιμετωπιζόταν με διαφορετικό τρόπο.

ΠΑΡΑΔΕΙΓΜΑ 1.1.

Ο παρακάτω Πίνακας 1.2 δίνει τις τιμές για 8 $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$ μεταβλητές ($p=8$) που αφορούν σε 24 τύπους αυτοκινήτων ($n=24$). Οι μεταβλητές παίρνουν τιμές από 1 (πολύ καλό) έως 6 (πολύ κακό):

X_1 : A Οικονομία

X_2 : B Συντήρηση

X_3 : C Διατήρηση της χρηματικής αξίας

X_4 : D Τιμή (με 1 τα πολύ φθηνά αυτοκίνητα)

X_5 : E Σχεδίαση

X_6 : F Σπορ συμπεριφορά

X_7 : G Ασφάλεια

X_8 : H Εύκολος χειρισμός

Type	Model	Economy	Service	Value	Price	Design	Sport.	Safety	Easy h.
Audi	100	3.9	2.8	2.2	4.2	3.0	3.1	2.4	2.8
BMW	5 series	4.8	1.6	1.9	5.0	2.0	2.5	1.6	2.8
Citroen	AX	3.0	3.8	3.8	2.7	4.0	4.4	4.0	2.6
Ferrari		5.3	2.9	2.2	5.9	1.7	1.1	3.3	4.3
Fiat	Uno	2.1	3.9	4.0	2.6	4.5	4.4	4.4	2.2
Ford	Fiesta	2.3	3.1	3.4	2.6	3.2	3.3	3.6	2.8
Hyundai		2.5	3.4	3.2	2.2	3.3	3.3	3.3	2.4
Jaguar		4.6	2.4	1.6	5.5	1.3	1.6	2.8	3.6
Lada	Samara	3.2	3.9	4.3	2.0	4.3	4.5	4.7	2.9
Mazda	323	2.6	3.3	3.7	2.8	3.7	3.0	3.7	3.1
Mercedes	200	4.1	1.7	1.8	4.6	2.4	3.2	1.4	2.4
Mitsubishi	Galant	3.2	2.9	3.2	3.5	3.1	3.1	2.9	2.6
Nissan	Sunny	2.6	3.3	3.9	2.1	3.5	3.9	3.8	2.4
Opel	Corsa	2.2	2.4	3.0	2.6	3.2	4.0	2.9	2.4
Opel	Vectra	3.1	2.6	2.3	3.6	2.8	2.9	2.4	2.4
Peugeot	306	2.9	3.5	3.6	2.8	3.2	3.8	3.2	2.6
Renault	19	2.7	3.3	3.4	3.0	3.1	3.4	3.0	2.7
Rover		3.9	2.8	2.6	4.0	2.6	3.0	3.2	3.0
Toyota	Corolla	2.5	2.9	3.4	3.0	3.2	3.1	3.2	2.8
Volvo		3.8	2.3	1.9	4.2	3.1	3.6	1.6	2.4
Trabant	601	3.6	4.7	5.5	1.5	4.1	5.8	5.9	3.1
VW	Golf	2.4	2.1	2.0	2.6	3.2	3.1	3.1	1.6
VW	Passat	3.1	2.2	2.1	3.2	3.5	3.5	2.8	1.8
Wartburg	1.3	3.7	4.7	5.5	1.7	4.8	5.2	5.5	4.0

Πίνακας 1.2 Τιμές μεταβλητών τύπων αυτοκινήτων.

Ας εξετάσουμε δυαδικές μεταβλητές που υπολογίζονται από τα παραπάνω δεδομένα. Ορίζουμε τα νέα δυαδικά δεδομένα από:

$$y_{ik} = \begin{cases} 1 & \text{αν} \quad x_{ik} > \bar{x}_k \\ 0 & \text{αλλιώς} \end{cases} \qquad \text{για } i = 1,...,24 \text{ και } k = 1,...,8.$$

Αυτό σημαίνει ότι έχουμε μετατρέψει τις παρατηρήσεις της μεταβλητής k-τάξης σε 1 εάν είναι μεγαλύτερη από την μέση τιμή όλων των παρατηρήσεων της μεταβλητής

k-τάξης. Ας εξετάσουμε μόνο τα δεδομένα των παρατηρήσεων 17 έως 19 (Renault 19, Rover και Toyota Corolla) που οδηγούν σε (3 X 3) πίνακες απόσταση.

Το μέτρο Jaccard δίνει τον πίνακα ομοιότητας:

$$D = \begin{pmatrix} 1.000 & 0.000 & 0.333 \\ & 1.000 & 0.250 \\ & & 1.000 \end{pmatrix}$$

Το μέτρο Tanimoto δίνει τον πίνακα ομοιότητας:

$$D = \begin{pmatrix} 1.000 & 0.231 & 0.600 \\ & 1.000 & 0.455 \\ & & 1.000 \end{pmatrix}$$

Το Simple Matching (M) μέτρο δίνει τον πίνακα ομοιότητας:

$$D = \begin{pmatrix} 1.000 & 0.375 & 0.750 \\ & 1.000 & 0.625 \\ & & 1.000 \end{pmatrix}$$

Γίνεται ήδη φανερό ότι η επιλογή του μέτρου ομοιότητας ή απόστασης θα επηρεάσει τα επόμενα βήματα του αλγορίθμου μιας και οι προτάσεις θα είναι διαφορετικές.

1.4.Μέτρα απόστασης για συνεχείς μεταβλητές

Μια ευρεία ποικιλία των μέτρων απόσταση μπορεί να παραχθεί από τους

$L_r - norms$, $r \geq 1$,

$$d_{ij} = \left\| x_i - x_j \right\|_r = \left\{ \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^r \right\}^{\frac{1}{r}}.$$

Εδώ x_{ik} δηλώνει την τιμή της μεταβλητής k-τάξης της παρατήρησης i. Είναι σαφές ότι $d_{ii} = 0$ για $i = 1,...,n$. Η $L_1 - norm$, για παράδειγμα, δίνει μικρότερη βαρύτητα στις ακραίες τιμές από ότι η $L_2 - norm$ (Ευκλείδεια νόρμα). Συνηθέστερα χρησιμοποιούμε το τετράγωνο της $L_2 - norm.$

ΠΑΡΑΔΕΙΓΜΑ 1.2.

Έστω ότι έχουμε $x_1 = (0,0),\ x_2 = (1,0)$ και $x_3 = (5,5).$

Ο πίνακας αποστάσεων της $L_1 = norm$ είναι :

$$D_1 = \begin{pmatrix} 0 & 1 & 10 \\ 1 & 0 & 9 \\ 10 & 9 & 0 \end{pmatrix}$$

και το τετράγωνο της $L_2 - norm$ ή Ευκλείδειας απόστασης είναι:

$$D_2 = \begin{pmatrix} 0 & 1 & 50 \\ 1 & 0 & 41 \\ 50 & 41 & 0 \end{pmatrix}$$

Μπορούμε να παρατηρήσουμε ότι η τρίτη παρατήρηση x_3 έχει μεγαλύτερη βαρύτητα στην δεύτερη περίπτωση.

Μια βασική παραδοχή στην εφαρμογή αποστάσεων με βάση $L_r - norms$ είναι ότι οι μεταβλητές μετρούνται στην ίδια κλίμακα. Εάν αυτό δεν συμβαίνει, πρέπει πρώτα να γίνει τυποποίηση ως προς τη διασπορά, διαιρώντας την κάθε μεταβλητή με την τυπική της απόκλιση.

Αυτό αντιστοιχεί με μια γενικότερη $L_2 - norm$ ή Ευκλείδεια νόρμα με ένα πίνακα ιδιοτιμών A $(n \times p)$, όπου A> 0:

$$d_{ij}^2 = \left\| x_i - x_j \right\|_A = \left(x_i - x_j \right)^T \mathrm{A} \left(x_i - x_j \right).$$

Οι $L_2 - norms$ δίνονται από τον A = I_p όπου I_p ο p διάστασης μοναδιαίος πίνακας, δηλαδή ο τετραγωνικός $(p \times p)$ πίνακας όπου τα στοιχεία της κύριας διαγωνίου του είναι μονάδες και όλα τα υπόλοιπα μηδέν.

Αλλά αν είναι επιθυμητή η τυποποίηση κατάλληλος είναι ο πίνακας $A = diag(s_{x_1x_1}^{-1},...,s_{x_px_p}^{-1})$ όπου $s_{x_1x_1},...,s_{x_px_p}$ είναι τα στοιχεία της κύριας διαγωνίου του πίνακα συνδιακύμανσης

$$S = \begin{pmatrix} s_{x_1x_1} & \cdots & s_{x_1x_p} \\ \vdots & \ddots & \vdots \\ s_{x_px_1} & \cdots & s_{x_px_p} \end{pmatrix}$$

Και $diag\left(a_{11}, a_{22},..., a_{pp}\right)$ συμβολίζει ένα τετράγωνο $(p \times p)$ πίνακα με μόνα μη-μηδενικά στοιχεία εκείνα της κύριας διαγωνίου.

Αν θυμηθούμε ότι $s_{x_kx_k} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ik} - \bar{x}_k\right)^2$ είναι η διακύμανση της κ-τάξης μεταβλητής

(αν $n \le 20$ μπορούμε να αντικαταστήσουμε τον παράγοντα $\frac{1}{n}$ με $\frac{1}{n-1}$) τότε:

$$d_{ij}^2 = \sum_{k=1}^{p} \frac{\left(x_{ik} - x_{jk}\right)^2}{s_{x_kx_k}} \ .$$

Εδώ κάθε μεταβλητή έχει την ίδια στάθμιση στον υπολογισμό των αποστάσεων και οι αποστάσεις δεν εξαρτώνται από μια συγκεκριμένη επιλογή των μονάδων μέτρησης.

1.5. Αλγόριθμοι Cluster

Δύο μεγάλες κατηγορίες μεθόδων ομαδοποίησης είναι οι:

1 ιεραρχικοί (hierarchical) αλγόριθμοι, και
2. αλγόριθμοι διαχωρισμού (portioning).

Η ιεραρχικοί αλγόριθμοι μπορούν να χωριστούν σε ενωτικές (agglomerative) και διάσπασης (divisive) που είναι ισοδύναμοι αντίστροφοι, τρόποι εκτέλεσης του αλγορίθμου. Ο πρώτος τύπος ιεραρχικής ομαδοποίησης ξεκινά από την πιο λεπτή διαμέριση, δηλ. **κάθε παρατήρηση θα αποτελεί ένα cluster** και στην πορεία του αλγορίθμου τις ομαδοποιεί. Ο δεύτερος τύπος ξεκινά από το αντίστροφο σενάριο, ότι όλες οι παρατηρήσεις αποτελούν ένα cluster και προχωρά στη διάσπασή του σε μικρότερα.

1.5.1 Ιεραρχικοί Αλγόριθμοι, Agglomerative Τεχνικές

Οι Agglomerative αλγόριθμοι χρησιμοποιούνται αρκετά συχνά στην πράξη. Ο agglomerative αλγόριθμος αποτελείται από τα ακόλουθα βήματα:

1. Κατασκεύασε τον πίνακα αποστάσεων D.
2. Βρες τα δύο clusters με την πλησιέστερη απόσταση.
3. Ένωσε τα δύο αυτά cluster σε ένα cluster.
4. Υπολόγισε την απόσταση μεταξύ των νέων ομάδων και παρατήρησε την μειωμένη απόσταση από τον πίνακα D.

Θα ήταν χρήσιμο εδώ να δοθούν περισσότερες πληροφορίες για το $2^{ο}$ βήμα.

Ένωση Στοιχείων και Ομάδων (Clusters). (Ιουλία Παπαγεωργίου, σημειώσεις ΜΠΕ Εκπαιδευτικών)

Υποθέτουμε την Agglomerative μέθοδο, όπου όλα τα στοιχεία είναι μεμονωμένα και σε κάθε στάδιο ενώνονται τα πιο κοντινά μέχρις ότου αποτελούν μία ομάδα.

Στα διάφορα στάδια ακολουθούμε τα εξής:

1. Όταν η σύγκριση είναι στοιχείο με στοιχείο, επιλέγουμε να ενώσουμε εκείνο το ζευγάρι με την ελάχιστη απόσταση.
2. Όταν η σύγκριση είναι μεταξύ στοιχείου και μιας υπάρχουσας ομάδας ή γενικότερα μεταξύ 2 ομάδων, πρέπει να οριστεί η απόσταση μεταξύ ομάδων και να ενωθούν οι ομάδες με ελάχιστη και πάλι απόσταση.
3. Δεν υπάρχει μοναδικός τρόπος ορισμού απόστασης μεταξύ ομάδων ακόμα και για δοθέν μέτρο απόστασης. Αντίθετα, υπάρχουν πολλοί τρόποι (τεχνικές) και η κάθε μια από αυτές αποτελεί και μια τεχνική Clustering.

Συνολικά οι πιο γνωστοί από τους κανόνες/τεχνικές είναι οι εξής:

Single Linkage or Nearest Neighbour

Κανόνας: Σαν απόσταση 2 ομάδων θεωρείται η ελάχιστη απόσταση όλων των ανά δύο μελών των ομάδων (πλησιέστερος γείτονας)

Complete Linkage or Furthest Neighbour

Κανόνας: Σαν απόσταση θεωρείται η μέγιστη απόσταση όλων των ανά δύο μελών των ομάδων (μακρινότερος γείτονας).

Average Linkage

Κανόνας: Σαν απόσταση 2 ομάδων θεωρείται ο μέσος όρος όλων των ανά δύο αποστάσεων των μελών των ομάδων.

Centroid

Κανόνας: Σαν απόσταση 2 ομάδων θεωρείται η απόσταση των ΄΄κέντρων΄΄ τους. Διάφοροι ορισμοί για το κέντρο κάποιων σημείων μπορεί να δοθεί.

Ward

Κανόνας: Υπολογίζεται το άθροισμα των τετραγωνικών σφαλμάτων (ESS) για κάθε ομάδα που μπορεί να σχηματισθεί από ένωση 2 άλλων. Αυτό ορίζεται ως:

$$ESS(C)=\Sigma\left(x_i\text{-mean}(C)\right)^2.$$

Στη συνέχεια γίνεται η ένωση κατά την οποία θα έχουμε το ελάχιστο (ESS).

ΚΕΦΑΛΑΙΟ 2

ΠΕΡΙΓΡΑΦΗ

2.Διαχωριστική ανάλυση (Discriminant Analysis)

Σύμφωνα με τους Hardle and Simar (κεφ.12**,** 2003) η διαχωριστική ανάλυση χρησιμοποιείται σε περιπτώσεις όπου τα clusters είναι γνωστά εκ των προτέρων. Ο στόχος της διαχωριστική ανάλυσης είναι να ταξινομήσει μια παρατήρηση, ή διάφορες παρατηρήσεις, σε αυτές τις γνωστές ομάδες.

Για παράδειγμα, στο credit scoring, η τράπεζα γνωρίζει από την εμπειρία του παρελθόντος ότι υπάρχουν καλοί πελάτες (οι οποίοι εξοφλούν το δάνειό τους, χωρίς κανένα πρόβλημα) και κακοί πελάτες (οι οποίοι έδειξαν προβλήματα στην αποπληρωμή των δανείων τους). Όταν ένας νέος πελάτης ζητά ένα δάνειο, η τράπεζα πρέπει να αποφασίσει αν πρέπει ή όχι να δώσει το δάνειο.

Τα στοιχεία του παρελθόντος της Τράπεζας παρέχουν δύο σύνολα δεδομένων πολυμεταβλητών παρατηρήσεων x_i στις δύο κατηγορίες πελατών (όπως για παράδειγμα την ηλικία, το μισθό, την οικογενειακή κατάσταση, το ποσό του δανείου, κλπ.). Ο νέος πελάτης είναι μια νέα παρατήρηση x με τις ίδιες μεταβλητές. Ο κανόνας απαγόρευσης των διακρίσεων πρέπει να ταξινομήσει τον πελάτη σε μία από τις δύο υπάρχουσες ομάδες και η διαχωριστική ανάλυση πρέπει να αξιολογεί τον κίνδυνο μιας πιθανής «κακής απόφασης».

Στις περισσότερες εφαρμογές οι ομάδες συνδέονται με τις φυσικές ομαδοποιήσεις ή σε ομάδες που είναι γνωστές από την ιστορία (όπως στο παράδειγμα το credit scoring).

Αυτές οι ομάδες θα μπορούσαν να έχουν σχηματιστεί από μια ανάλυση των clusters η οποία εκτελείται σε δεδομένα του παρελθόντος.

Η παράγραφος 2.1 παρουσιάζει τους κανόνες κατάταξης, όταν είναι γνωστοί οι πληθυσμοί, δηλαδή, όταν γνωρίζουμε την κατανομή του κάθε πληθυσμού. Όπως περιγράφεται στην παράγραφο 2.2, στην πράξη, τα χαρακτηριστικά του πληθυσμού πρέπει να εκτιμώνται από την ιστορία. Οι μέθοδοι παρουσιάζονται σε διάφορα παραδείγματα.

2.1 Κανόνες κατανομής για γνωστές κατανομές

Η Διαχωριστική ανάλυση είναι ένα σύνολο μεθόδων και εργαλείων που χρησιμοποιούνται για τη διάκριση μεταξύ των ομάδων του πληθυσμού Π_j ώστε να καθορίσουν τον τρόπο με τον οποίο θα τοποθετηθούν οι νέες παρατηρήσεις σε ομάδες.

Θα μας ενδιέφερε επί παραδείγματι η διάκριση μεταξύ πλαστών και γνησίων χαρτονομισμάτων με βάση τις μετρήσεις αυτών των χαρτονομισμάτων. Σε αυτή την περίπτωση έχουμε δύο ομάδες (πλαστά και γνήσια χαρτονομίσματα) και θα θέλαμε να καθιερώσουμε έναν αλγόριθμο (κανόνα) ώστε να μπορούμε να τοποθετήσουμε μια νέα παρατήρηση (ένα νέο χαρτονόμισμα) σε μία από τις ομάδες.

Ένα άλλο παράδειγμα είναι η ανίχνευση του "γρήγορου" και "αργού" καταναλωτή σε ένα νέο προϊόν. Χρησιμοποιώντας τα χαρακτηριστικά του καταναλωτή, όπως η εκπαίδευση, το εισόδημα, το μέγεθος της οικογένειας, θέλουμε να ταξινομηθεί κάθε καταναλωτής στις δύο ομάδες μόλις εντοπιστεί.

Στην ποίηση και στις λογοτεχνικές μελέτες οι συχνότητες των προφορικών ή γραπτών λέξεων και τα μήκη των προτάσεων δείχνουν καλλιτέχνες και συγγραφείς με διαφορετικά προφίλ. Μπορεί να είναι ενδιαφέρον να αποδίδονται άγνωστα λογοτεχνικά ή καλλιτεχνικά έργα σε ορισμένους συγγραφείς με ένα συγκεκριμένο προφίλ. Όπως για παράδειγμα να γίνει αναγνώριση για συγγραφείς αρχαίων επιγραφών ή κειμένων.

Ανθρωπολογικές μετρήσεις σε αρχαία κρανία βοηθούν στην διάκριση μεταξύ ανδρικών και γυναικείων σκελετών.

Καλή και κακή βαθμολογία του πιστωτικού κινδύνου αποτελεί πρόβλημα των διακρίσεων που θα μπορούσαν να αντιμετωπιστούν με παρατηρήσεις σχετικά με το εισόδημα, την ηλικία, τον αριθμό των πιστωτικών καρτών, το μέγεθος της οικογένειας κλπ.

Σε γενικές γραμμές έχουμε πληθυσμούς $\Pi_j, j = 1, 2, ..., J$ και πρέπει να τοποθετήσουμε μια x παρατήρηση σε μία από αυτές τις ομάδες.

Ένας κανόνας διάκρισης (κανόνας διαχωρισμού) είναι ένας διαχωρισμός του δειγματικού χώρου (σε γενικές γραμμές $\mathbb{R}^p$) σε σύνολα $\mathbb{R}_j$ τέτοια που, αν $x \in \mathbb{R}_j$, να αναγνωρίζεται ως μέλος του Π_j πληθυσμού.

Το κύριο έργο της διαχωριστικής ανάλυσης είναι να βρεθεί *«καλή»* περιοχή $\mathbb{R}_j$ έτσι ώστε το σφάλμα της εσφαλμένης ταξινόμησης να είναι μικρό. Στα επόμενα περιγράφουμε αυτούς τους κανόνες όταν είναι γνωστές οι κατανομές του πληθυσμού.

2.2. Η γραμμική διαχωριστική ανάλυση (LDA) και η γραμμική συνάρτηση διαχωρισμού Fisher.

Για την γραμμική διαχωριστική ανάλυση και τη γραμμική συνάρτηση διαχωρισμού Fisher υπάρχουν εκτενείς αναφορές στους Fisher, 1936 McLachlan, 2004 Martinez & Kak, 2004 Abdi, 2007 Perriere, 2003 Venable & Ripley, 2002 Yu & Yang, 2001 Friedman, 1989 Ahdesmaki & Strimmer, 2010.

Σύμφωνα με τους προηγούμενους η γραμμική διαχωριστική ανάλυση (LDA) και η σχετιζόμενη μ' αυτήν γραμμική συνάρτηση διαχωρισμού του Fisher, είναι μέθοδοι που χρησιμοποιούνται στην στατιστική, στην αναγνώριση προτύπων και αλλού. Η χρησιμότητά του έγκειται στη δυνατότητα εύρεσης ενός γραμμικού συνδυασμού των χαρακτηριστικών τα οποία διαχωρίζουν με τον καλύτερο δυνατό τρόπο δύο ή περισσότερες κλάσεις αντικειμένων ή γεγονότων.

Το αριθμητικό αποτέλεσμα του συνδυασμού για νέα παρατήρηση μπορεί να χρησιμοποιηθεί ως ένας γραμμικός ταξινομητής.

2.2.1 Η γραμμική διαχωριστική ανάλυση (LDA).

Η γραμμική διαχωριστική ανάλυση (LDA) είναι στενά συνδεδεμένη με την ανάλυση διακύμανσης (ANOVA) και την ανάλυση παλινδρόμησης. Η ανάλυση παλινδρόμησης επιχειρεί να εκφράσει μια εξαρτημένη μεταβλητή ως γραμμικό συνδυασμό άλλων χαρακτηριστικών ή μετρήσεων (ανεξάρτητες μεταβλητές).

Ωστόσο στις μεθόδους ANOVA και ανάλυση παλινδρόμησης, η εξαρτημένη μεταβλητή είναι μία αριθμητική ποσότητα (ποσοτική μεταβλητή), ενώ στην LDA είναι κατηγορική μεταβλητή. Οι μέθοδοι ANOVA και ανάλυση παλινδρόμησης είναι προτιμότερες σε εφαρμογές που δεν είναι λογικό να υποθέσουμε ότι οι ανεξάρτητες μεταβλητές ακολουθούν κανονική κατανομή, η οποία αποτελεί θεμελιώδη παραδοχή της LDA.

Η λογιστική παλινδρόμηση είναι πλησιέστερη στην LDA καθώς και αυτή εξηγεί κατηγορική μεταβλητή.

Η LDA είναι επίσης στενά συνδεδεμένη με τη ανάλυση κυρίων συνιστωσών (Principal Component Analysis ή PCA) και την παραγοντική ανάλυση (factor analysis) καθώς και οι δύο αυτές αναζητούν γραμμικό συνδυασμό των μεταβλητών που εξηγεί καλύτερα είτε την συνολική μεταβλητότητα των δεδομένων είτε τις συσχετίσεις των μεταβλητών.

Η LDA επιχειρεί να μοντελοποιήσει τη διαφορά ανάμεσα στις κλάσεις των δεδομένων.

Η PCA από την άλλη δεν λαμβάνει υπόψη την τυχόν διαφορά στην κλάση, και η παραγοντική ανάλυση δημιουργεί τους συνδυασμούς των χαρακτηριστικών βασιζόμενη περισσότερο στις διαφορές παρά στις ομοιότητες.

Η διαχωριστική ανάλυση διαφέρει επίσης από την παραγοντική ανάλυση στο ότι δεν είναι μία τεχνική αλληλεξάρτησης. Στην διαχωριστική ανάλυση πρέπει να δημιουργηθεί μια διάκριση μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών.

Η LDA λειτουργεί όταν οι μετρήσεις που έγιναν σε ανεξάρτητες μεταβλητές για κάθε παρατήρηση είναι συνεχείς ποσότητες. Όταν ασχολούμαστε με κατηγορικές ανεξάρτητες μεταβλητές, κατάλληλη τεχνική είναι η διαχωριστική correspondence ανάλυση.

2.2.2 Η γραμμική διαχωριστική ανάλυση (LDA) για δύο κλάσεις.

Έστω ότι υπάρχουν δύο κλάσεις y ($y \in \{0,1\}$) και ένα σετ παρατηρήσεων $\boldsymbol{x}$. Αν οι συναρτήσεις πυκνότητας πιθανότητας $p(\boldsymbol{x} / y=0), p(\boldsymbol{x} / y=1)$ της $\boldsymbol{x}$. ακολουθούν κανονική κατανομή με μέσο και συνδιακύμανση $(\boldsymbol{\mu_0}, \Sigma_{y=0}), (\boldsymbol{\mu_1}, \Sigma_{y=1})$ ένα καλός δείκτης διαχωρισμού είναι ο λόγος του λογαρίθμου των πιθανοφανειών να είναι μικρότερος από μία σταθερά t δηλαδή:

$$(\boldsymbol{x}-\boldsymbol{\mu_0})^T \Sigma_{y=0}^{-1} (\boldsymbol{x}-\boldsymbol{\mu_0}) + \ln|\Sigma_{y=0}| - (\boldsymbol{x}-\boldsymbol{\mu_1})^T \Sigma_{y=1}^{-1} (\boldsymbol{x}-\boldsymbol{\mu_1}) + \ln|\Sigma_{y=1}| < t \text{ .(A).}$$

Αν υποθέσουμε ομοσκεδαστικότητα $(\Sigma_{y=0} = \Sigma_{y=1} = \Sigma)$ τότε η (A) οδηγείται στο:

$\boldsymbol{w} \cdot \boldsymbol{x} < c$ όπου c μία ελάχιστη σταθερά και $\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})$.

Αυτό σημαίνει ότι το κριτήριο να ανήκει μία καινούργια παρατήρηση $\boldsymbol{x}$ σε μία κλάση y είναι καθαρά γραμμική σχέση των γνωστών παρατηρήσεων $\boldsymbol{x}$.

Αν δεν υποθέσουμε ομοσκεδαστικότητα τότε οδηγούμαστε σε Quadratic Discriminant Analysis (QDA) περιγραφή της οποίας υπάρχει στην παράγραφο 2.2.4.

2.2.3 Η γραμμική συνάρτηση διαχωρισμού **Fisher**.

Μια άλλη προσέγγιση που πηγάζει από τον R. A. Fisher. Είναι η ιδέα του ήταν να στηριχτεί ο διαχωριστικός κανόνας σε μια προβολή $a^T x$ τέτοια ώστε να επιτευχθεί ένας καλός διαχωρισμός. Αυτή η μέθοδος προβολής LDA ονομάζεται γραμμική συνάρτηση διαχωρισμού του Fisher.

Αν

$$\boldsymbol{y = xa}$$

δηλώνει ένα γραμμικό συνδυασμό των παρατηρήσεων, τότε το συνολικό άθροισμα των τετραγώνων των y, $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$

είναι ίσο με $\boldsymbol{y}^T H\boldsymbol{y} = \alpha^T \boldsymbol{x}^T H\boldsymbol{x}\alpha = \alpha^T T\alpha$

με $H = I - n^{-1}1_n 1_n^T$ όπου $1_n = \underset{n-φορές}{[1,...,1]^T}$ και $T = X^T HX$.

Ο πίνακας H έχει διάσταση $n \text{ x } n$ και τα διαγώνια στοιχεία του είναι $h_{ii} = \frac{n-1}{n}, i = 1,...,n.$

Ας υποθέσουμε ότι έχουμε δείγματα $\boldsymbol{x}_j, j = 1,...,J$ που προέρχονται από J πληθυσμούς. Η εισήγηση του Fisher ήταν να βρεθεί ο γραμμικός συνδυασμός $a^T x$ που μεγιστοποιεί την αναλογία μεταξύ των αθροισμάτων τετραγώνων μεταξύ των ομάδων προς το άθροισμα των τετραγώνων στο εσωτερικό των ομάδων.

Το άθροισμα τετραγώνων στο εσωτερικό των ομάδων δίνεται από

$$\sum_{j=1}^{J} \boldsymbol{y}_j^T H_j \boldsymbol{y}_j = \sum_{j=1}^{J} \alpha^T \boldsymbol{x}_j^T H_j \boldsymbol{x}_j \alpha = \alpha^T W \alpha,$$

όπου $\boldsymbol{y}_j$ δηλώνει τον j υπο-πίνακα του $\boldsymbol{y}$ που αντιστοιχεί στις παρατηρήσεις της ομάδας j και H_j δηλώνει τον $(n_j \text{ x } n_j)$ κεντραρισμένο πίνακα.

Το άθροισμα τετραγώνων στο εσωτερικό των ομάδων μετράει το άθροισμα των διακυμάνσεων σε κάθε ομάδα.

Το άθροισμα των τετραγώνων μεταξύ των ομάδων δίνεται από:

$$\sum_{j=1}^{J} n_j \left(\bar{y}_j - \bar{y}\right)^2 = \sum_{j=1}^{J} n_j \left\{\alpha^T \left(\bar{x}_j - \bar{x}\right)^2\right\} = \alpha^T Ba$$

όπου $\bar{y}_j$ και $\bar{x}_j$ χαρακτηρίζει το μέσο των y_j και x_j και $\bar{y}$ και $\bar{x}$ τους δειγματικούς μέσους των $\boldsymbol{y}$ και $\boldsymbol{x}$ αντίστοιχα. Το άθροισμα των τετραγώνων μεταξύ των ομάδων μετρά την διακύμανση των μέσων μεταξύ των ομάδων. Το συνολικό άθροισμα των τετραγώνων είναι το άθροισμα των τετραγώνων στο εσωτερικό των ομάδων και των τετραγώνων μεταξύ των ομάδων, δηλαδή, $\alpha^T T\alpha = \alpha^T W\alpha + \alpha^T B\alpha$.

Η ιδέα του Fisher ήταν να επιλέξει ένα διάνυσμα προβολής που μεγιστοποιεί το λόγο

$$\frac{\alpha^T B\alpha}{\alpha^T W\alpha}.$$

Η λύση βρίσκεται εφαρμόζοντας το παρακάτω θεώρημα (2.1) των Hardle-Simar, 2003 σελ 68:

ΘΕΩΡΗΜΑ 2.1.

Αν Α και B είναι συμμετρικοί με $B>0$, τότε το μέγιστο των $x^T Ax$ κάτω από τον περιορισμό $x^T Bx=1$ δίνεται από τη μέγιστη ιδιοτιμή του $B^{-1}A$ που αντιστοιχεί στη μεγαλύτερη (μικρότερη) ιδιοτιμή (Παράρτημα Α) του $B^{-1}A$. Γενικά:

$$\max_{\{x:x^T Bx=1\}} \boldsymbol{x}^T A\boldsymbol{x} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = \min_{\{x:x^T Bx=1\}} \boldsymbol{x}^T A\boldsymbol{x}$$

όπου $\lambda_1,...,\lambda_p$ δηλώνουν τις ιδιοτιμές του $B^{-1}A$ (Hardle and Simar κεφ.12**,** 2003).

Το διάνυσμα που μεγιστοποιεί (ελαχιστοποιεί) τον $\boldsymbol{x}^T A\boldsymbol{x}$ κάτω από τον περιορισμό $\boldsymbol{x}^T B\boldsymbol{x}=1$ είναι το ιδιοδιάνυσμα του $B^{-1}A$ που αντιστοιχεί στη μεγαλύτερη (μικρότερη) ιδιοτιμή του $B^{-1}A$.

Τώρα ένας κανόνας διάκρισης είναι εύκολο να επιτευχθεί:

Ταξινομώντας τα $\boldsymbol{x}$ σε J ομάδες όπου $a^T\bar{\boldsymbol{x}}_j$ είναι το κοντινότερο στο $a^T\boldsymbol{x}$ π.χ

$\boldsymbol{x}\to\Pi_j$ όπου $j=\arg\min_i \left|a^T\left(\boldsymbol{x}-\bar{\boldsymbol{x}}\right)\right|$.

Όταν $J=2$ ομάδες, ο διαχωριστικός κανόνας είναι εύκολο να υπολογιστεί. Ας υποθέσουμε ότι η ομάδα 1 έχει n_1 στοιχεία και η ομάδα 2 έχει n_2 στοιχεία.

Σε αυτή την περίπτωση $B=\left(\frac{n_1 n_2}{n}\right)dd^T$ όπου $d=\left(\bar{x}_1-\bar{x}_2\right)$.

Ο πίνακας $W^{-1}B$ έχει μία μόνο ιδιοτιμή που ισούται με: $tr\left(W^{-1}B\right)=\left(\frac{n_1 n_2}{n}\right)d^T W^{-1}d$

και αντίστοιχο ιδιοδιάνυσμα (Παράρτημα Α) που ισούται με: $\alpha=W^{-1}d$.

Ο αντίστοιχος διαχωριστικός κανόνας είναι:

$$x \to \Pi_1 \quad \alpha\nu \quad \alpha^T \left\{ x - \frac{1}{2}\left(\bar{x}_1 + \bar{x}_2\right) \right\} > 0$$

$$x \to \Pi_2 \quad \alpha\nu \quad \alpha^T \left\{ x - \frac{1}{2}\left(\bar{x}_1 + \bar{x}_2\right) \right\} \le 0 .$$

Η Fisher LDA είναι στενά συνδεδεμένη με την επιδίωξη προβολής δεδομένου ότι η στατιστική τεχνική βασίζεται σε ένα τρισδιάστατο δείκτη $a^T x$.

2.2.4 Η τετραγωνική διαχωριστική ανάλυση (Quadratic Discriminant Analysis)

Η QDA (Sathyanarayana, Shashi 2010) είναι στενά συσχετισμένη με την LDA, έχουμε υποθέσει ότι υπάρχουν μόνο δύο κλάσεις y ($y \in \{0,1\}$) και ότι οι τιμές των παρατηρήσεων της $\boldsymbol{x}$ ακολουθούν κανονική κατανομή.

Αντίθετα από την LDA όμως, στην QDA δεν υπάρχει η υπόθεση των ίσων συνδιακυμάνσεων σε κάθε κλάση.

Όταν η υπόθεση $\Sigma_{y=0} \neq \Sigma_{y=1}$ ισχύει, ο καλύτερος έλεγχος για την υπόθεση ότι μία τυχαία μέτρηση προέρχεται από μία από τις δεδομένες κλάσεις είναι ο έλεγχος του λόγου των πιθανοφανειών (likelihood ratio test).

Αν $\mu_{y=0}, \mu_{y=1}$ οι γνωστοί μέσοι για κάθε κλάση και $\Sigma_{y=0}, \Sigma_{y=1}$ οι πίνακες συνδιακυμάνσεων τότε ο λόγος των πιθανοφανειών δίνεται από:

$$\text{Likelihood ratio} = \frac{\sqrt{2\pi\left|\Sigma_{y=1}\right|}^{-1} e^{\left(-\frac{1}{2}\left(x-\mu_{y=1}\right)^T \Sigma_{y=1}^{-1}\left(x-\mu_{y=1}\right)\right)}}{\sqrt{2\pi\left|\Sigma_{y=0}\right|}^{-1} e^{\left(-\frac{1}{2}\left(x-\mu_{y=0}\right)^T \Sigma_{y=0}^{-1}\left(x-\mu_{y=0}\right)\right)}} < t$$ όπου t μια ελάχιστη σταθερή τιμή όπου μετά από πράξεις μπορεί να δειχθεί ότι ό τρόπος διαχωρισμού μεταξύ των κλάσεων είναι τετραγωνικός.

2.3. Κανόνας μέγιστης πιθανοφάνειας (ML).

Χαρακτηρίζει την πυκνότητα του κάθε πληθυσμού Π_j από την κατανομή της, έστω $f_j(x)$. Ο κανόνας μέγιστης πιθανοφάνειας (ML rule) δίνεται από την τοποθέτηση του $\boldsymbol{x}$ στον πληθυσμό Π_j ώστε να μεγιστοποιείται η πιθανοφάνεια $L_j(x) = f_j(x) = \max_i f_i(x)$.

Ορισμός Πιθανοφάνειας.

Με τον όρο πιθανοφάνεια όπως αναφέρεται στο βιβλίο του Spiegel,1975 σελ 196,206,207 ορίζουμε τη συνάρτηση:

$$L(\theta) = \prod_{\iota=1}^{n} f_j(x_\iota;\theta)$$

όπου $f_j(x_i;\theta)$ η συνάρτηση πυκνότητας των x_j ανεξάρτητων n παρατηρήσεων (i.i.d) και θ η προς εκτίμηση παράμετρος. Η συνάρτηση $L(\theta)$ ονομάζεται συνάρτηση πιθανοφάνειας.

Επίσης μερικές φορές η κατανομή f_j είναι ίδια για όλες τις κλάσεις, π.χ κανονική, και αυτό που αλλάζει είναι οι παράμετροι από κλάση σε κλάση. Τότε θα η συνάρτηση θα ήταν $L_j(\theta) = \prod_{\iota=1}^{n} f_j(x_\iota;\theta_j)$.

Ως μία εκτίμηση της θ μπορεί να ληφθεί η τιμή που δίνει η μέγιστη πιθανοφάνεια $L(\theta)$ και ο εκτιμητής ονομάζεται εκτιμητής μέγιστης πιθανοφάνειας (MLE). Ένας εκτιμητής που χρησιμοποιείται πολύ συχνά στην πράξη λόγω των πολύ καλών ιδιοτήτων του. Αντί να μηδενίσουμε την παράγωγο της $L(\theta)$ μηδενίζουμε την παράγωγο του λογαρίθμου της $L(\theta)$, $\log L(\theta)$ που συμβολίζεται και $l(\theta)$.

Δηλαδή αν $\log L(\theta) = \log \prod_{i=1}^{n} f_j(x_i;\theta) = \sum_{i=1}^{n} \log f_j(x_i;\theta)$ τότε

$$\frac{\partial l}{\partial \theta} = \frac{1}{f_1(x_1,\theta)} \frac{\partial f_1(x_1,\theta)}{\partial \theta} + \ldots + \frac{1}{f_n(x_n,\theta)} \frac{\partial f_n(x_n,\theta)}{\partial \theta} \equiv 0$$

από όπου υπολογίζεται το θ συναρτήσει των x_j.

Η μέθοδος μέγιστης πιθανοφάνειας μπορεί να γενικευθεί. Έτσι στην περίπτωση που έχουμε να προσδιορίσουμε πολλές παραμέτρους θ_j συγχρόνως, παίρνουμε τις μερικές παραγώγους ως προς τις παραμέτρους, τις μηδενίζουμε και λύνουμε το σύστημα. Μια εξήγηση της μεθόδου , ως εύρεση μεγίστου, είναι ότι η L σε μια διακριτή περίπτωση κατανομής f_j ταυτίζεται ως συνάρτηση των x με την πιθανότητα να έχουμε παρατηρήσει τα συγκεκριμένα δεδομένα. Συνεπώς, μεγιστοποιώντας την L ή ισοδύναμα την l ως προς θ, βρίσκουμε ποια τιμή του θ

(ή των $\underset{\sim}{\theta}$) ήταν η πιο πιθανή, με μέγιστη πιθανότητα, να έδωσε σαν δεδομένα αυτά που παρατηρήσαμε.

2.3.1. Παραδείγματα εκτιμήσεων μέγιστης πιθανοφάνειας.

Παράδειγμα 1.

Έστω ότι n μετρήσεις σ΄ έναν κανονικό πληθυσμό με άγνωστη μέση τιμή και γνωστή τυπική απόκλιση έδωσαν τις τιμές $x_1,...,x_n$. Ζητείται η εκτίμηση της μέσης τιμής του πληθυσμού με τη μέθοδο της μέγιστης πιθανοφάνειας. ($\theta = \mu$).

Επειδή $f(x_k,\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}$

έχουμε $L = f(x_1,\mu)\cdots f(x_k\mu) = \frac{1}{\sqrt[n]{2\pi\sigma^2}} e^{-\sum_{k=1}^{n}\frac{(x_k-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\sum_{k=1}^{n}\frac{(x_k-\mu)^2}{2\sigma^2}}$.

Λογαριθμόντας την προηγούμενη σχέση έχουμε:

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2 . \qquad \text{(A)}$$

Παίρνοντας την μερική παράγωγο ως προς μ έχουμε:

$$\frac{1}{L}\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k-\mu).$$

Θέτοντας $\frac{\partial L}{\partial \mu} \equiv 0$ έχουμε $\sum_{k=1}^{n}(x_k-\mu) = 0 \Rightarrow \sum_{k=1}^{n}x_k - n\mu = 0 \Rightarrow \mu = \frac{\sum_{k=1}^{n}x_k}{n} = \bar{x}.$

Για να είναι max και όχι min πρέπει να εξετασθεί και η $\frac{\partial^2 L}{\partial \mu^2}$ και να είναι αρνητική στο $\mu = \bar{x}$ με μ γνωστό.

Άρα η μέθοδος της μέγιστης πιθανοφάνειας δίνει ως εκτίμηση της μέσης τιμής του πληθυσμού σε κανονική κατανομή την μέση τιμή του δείγματος.

Παράδειγμα 2.

Αν στο Παράδειγμα 1 ζητάμε την εκτίμηση της διασποράς του πληθυσμού ($\theta = \sigma^2$) τότε παίρνοντας την μερική παράγωγο της σχέσης (Α) ως προς σ^2 έχουμε:

$$\frac{1}{L}\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{k=1}^{n}(x_k - \mu)^2.$$

Θέτοντας $\frac{\partial L}{\partial \sigma^2} = 0$ έχουμε:

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{k=1}^{n}(x_k - \mu)^2 = 0 \Rightarrow \frac{n}{2\sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{k=1}^{n}(x_k - \mu)^2 \Rightarrow$$

$$\Rightarrow \sigma^2 = \frac{\sum_{k=1}^{n}(x_k - \mu)^2}{n}.$$

Εάν περισσότεροι f_i δίνουν το ίδιο μέγιστο τότε οποιαδήποτε από αυτές μπορεί να επιλεγεί. Από μαθηματική άποψη, τα σύνολα R_j που έδωσε ο ML κανόνας διαχωριστικής ανάλυσης ορίζονται ως:

$$R_j = \left\{x : L_j(x) > L_i(x) \quad \gamma\iota\alpha \quad i = 1, 2, ..., J \quad , \quad i \neq j\right\}$$

Με την ταξινόμηση των παρατηρήσεων σε συγκεκριμένη ομάδα μπορεί να υποπέσουμε σε σφάλμα εσφαλμένης κατάταξης.

Για J = 2 ομάδες η πιθανότητα να τεθεί η x στην ομάδα 2 αν και είναι από τον πληθυσμό 1 μπορεί να υπολογιστεί ως

$$p_{21} = P(X \in R_2 / \Pi_1) = \int_{R_2} f_1(x)dx.$$

Ομοίως, η υπό όρους πιθανότητα χαρακτηρισμού ενός αντικειμένου που ανήκει στην πρώτη πληθυσμού Π1 αν και πρόκειται στην πραγματικότητα για αντικείμενο του Π2 είναι

$$p_{12} = P(X \in R_1 / \Pi_2) = \int_{R_1} f_2(x)dx$$

Οι εσφαλμένα ταξινομημένες παρατηρήσεις δημιουργούν ένα κόστος C (i / j) όταν μια παρατήρηση Π_j έχει ανατεθεί σε R_i Στο παράδειγμα του πιστωτικού κινδύνου, αυτό θα μπορούσε να είναι το κόστος μιας «κακής» αξιολόγησης της πιστοληπτικής ικανότητας. Η διάρθρωση του κόστους μπορεί να καταχωρηθεί σε έναν πίνακα κόστους (Πίνακας 2.1):

		classified	population
		Π_1	Π_2
True polulation	Π_1	0	$C(2/1)$
	Π_2	$C(1/2)$	0

Πίνακας 2.1 Κόστος "κακής" αξιολόγησης πιστοληπτικής ικανότητας

Έστω π_j η prior πιθανότητα του πληθυσμού Π_j όπου το "prior", είναι η εκ των προτέρων πιθανότητα ότι ένα άτομο που επιλέγεται τυχαία ανήκει στο Π_j (δηλαδή, πριν εξετάσει την τιμή του x). Για τις prior πιθανότητες θα πρέπει να εξετάζεται εκ των προτέρων αν είναι σαφές ότι μια παρατήρηση είναι πιο πιθανό να προέρχεται από ένα συγκεκριμένο Π_j πληθυσμού: Ένα παράδειγμα είναι η ταξινόμηση των μουσικών έργων. Αν είναι γνωστό ότι για μια ορισμένη χρονική περίοδο η πλειοψηφία των μουσικών έργων γράφτηκαν από ένα συγκεκριμένο συνθέτη, τότε υπάρχει μεγαλύτερη πιθανότητα ότι ένα μουσικό έργο να είναι σύνθεση αυτού του συνθέτη. Ως εκ τούτου, θα πρέπει να θεωρούμε μεγαλύτερη την prior πιθανότητα, όταν μια μελωδία καταγράφεται σε συγκεκριμένη ομάδα.

Το αναμενόμενο κόστος της εσφαλμένης τοποθέτησης (ECM) δίνεται

$$ECM = C(2/1)\pi_1 p_{21} + C(1/2)\pi_2 p_{12}$$

Θα πρέπει οι κανόνες κατάταξης να κρατούν το ECM μικρό ή να την ελαχιστοποιούν μέσω κανόνων. Ο κανόνας διαχωρισμού που ελαχιστοποιεί το ECM για δύο πληθυσμούς δίνεται στο παρακάτω θεώρημα.

Θεώρημα 2.2.

Για δυο δοθέντες πληθυσμούς ο κανόνας ελαχιστοποίησης του ECM δίνεται από

$$R_1 = \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{C(1/2)\pi_2}{C(2/1)\pi_1} \right\}$$

$$R_2 = \left\{ x : \frac{f_2(x)}{f_1(x)} \geq \frac{C(2/1)\pi_1}{C(1/2)\pi_2} \right\}$$

ή ισοδύναμα μπορούμε να γράψουμε:

$$R_1 = \left\{ x : \frac{f_2(x)}{f_1(x)} < \frac{C(2/1)\pi_1}{C(1/2)\pi_2} \right\} \quad R_2 = \left\{ x : \frac{f_1(x)}{f_2(x)} < \frac{C(1/2)\pi_2}{C(2/1)\pi_1} \right\}.$$

Ο κανόνας ML της Διαχωριστικής ανάλυσης είναι μια ειδική περίπτωση του κανόνα ECM για την ισοκατανομή του κόστους εσφαλμένης τοποθέτησης και την ισότητα των prior πιθανοτήτων. Για την απλότητα στα επόμενα θα θεωρήσουμε ότι
C(1 / 2) = C (2 / 1) = 1, και ίσες οι prior πιθανότητες, $\pi_2 = \pi_1$.

Το θεώρημα 2.1 θα αποδειχθεί μέσω ενός παραδείγματος πάνω στο credit scoring.

ΠΑΡΑΔΕΙΓΜΑ 3

Ας υποθέσουμε ότι Π_1 αντιπροσωπεύει τον πληθυσμό των επισφαλών πελατών που δημιουργεί το κόστος C (2 / 1) εάν ταξινομηθούν ως καλοί πελάτες. Κατ 'αναλογία καθορίζεται C (1 / 2), το κόστος της απώλειας ενός καλού πελάτη χαρακτηρίζοντάς τον ως επισφαλή. Αν με γ συμβολίσουμε τα κέρδη της τράπεζας για την ορθή κατάταξη ενός καλού πελάτη, τότε το συνολικό κέρδος της τράπεζας είναι:

$$G(R_2) = -C(2/1)\pi_1 \int_{R_2} f_1(x)dx - C(1/2)\pi_2 \int_{R_1} f_2(x)dx + \gamma\pi_2 \int_{R_2} f_2(x)dx \quad \text{(A)}$$

Επειδή $\int_{R_1} f_2(x)dx = 1 - \int_{R_2} f_2(x)dx$ (B_1) και $\left(\int_a^x f(x)dx\right)' = f(x)$ (B_2)

η σχέση (A) μέσω της (B_1) γράφεται:

$$G(R_2) = -C(2/1)\pi_1 \int_{R_2} f_1(x)dx - C(1/2)\pi_2 \int_{R_1} f_2(x)dx + \gamma\pi_2 \int_{R_2} f_2(x)dx =$$
$$= -C(2/1)\pi_1 \int_{R_2} f_1(x)dx - C(1/2)\pi_2 \left(1 - \int_{R_2} f_2(x)dx\right) + \gamma\pi_2 \int_{R_2} f_2(x)dx =$$
$$= -C(2/1)\pi_1 \int_{R_2} f_1(x)dx - C(1/2)\pi_2 + C(1/2)\pi_2 \int_{R_2} f_2(x)dx + \gamma\pi_2 \int_{R_2} f_2(x)dx =$$
$$= -C(1/2)\pi_2 + \int_{R_2} -C(2/1)\pi_1 f_1(x)dx + \int_{R_2} C(1/2)\pi_2 f_2(x)dx + \int_{R_2} \gamma\pi_2 f_2(x)dx =$$
$$= -C(1/2)\pi_2 + \int_{R_2} -C(2/1)\pi_1 f_1(x)dx + \int_{R_2} \left[C(1/2) + \gamma\right]\pi_2 f_2(x)dx =$$
$$= -C(1/2)\pi_2 + \int_{R_2} \left\{-C(2/1)\pi_1 f_1(x) + \left[C(1/2) + \gamma\right]\pi_2 f_2(x)\right\}dx.$$

Δεδομένου ότι ο πρώτος όρος της εξίσωσης αυτής καθώς και ο $-C(1/2)\pi_2$ είναι σταθεροί και χρησιμοποιώντας την (B_2) (ορισμός της αρχικής συνάρτησης $F(x)$ μιας συνεχούς συνάρτησης $f(x)$) και παραγωγίζοντας έχουμε:

$$(G(R_2))' = (-C(1/2)\pi_2 + \int_{R_2} \left\{-C(2/1)\pi_1 f_1(x) + \left[C(1/2) + \gamma\right]\pi_2 f_2(x)\right\}dx)' \Rightarrow$$
$$\Rightarrow 0 = -C(2/1)\pi_1 f_1(x) + \left[C(1/2) + \gamma\right]\pi_2 f_2(x)$$

Το μέγιστο λαμβάνεται αν:

$$-C(2/1)\pi_1 f_1(x)+\left[C(1/2)+\gamma\right]\pi_2 f_2(x)\geq 0\Rightarrow$$
$$\Rightarrow\left[C(1/2)+\gamma\right]\pi_2 f_2(x)\geq C(2/1)\pi_1 f_1(x)\Rightarrow$$
$$\Rightarrow\frac{f_2(x)}{f_1(x)}\geq\frac{C(2/1)\pi_1}{\left[C(1/2)+\gamma\right]\pi_2}$$

το οποίο αντιστοιχεί στο σύνολο R_2 του Θεωρήματος 2.1 για ένα κέρδος γ = 0.

ΠΑΡΑΔΕΙΓΜΑ 4

Υποθέτουμε ότι x $\in$ {0,1} και

$$\Pi_1 : P(X=0)=P(x=1)=\frac{1}{2}$$

$$\Pi_2 : P(X=0)=\frac{1}{4}=1-P(X=1)$$

Ο δειγματικός χώρος είναι το σύνολο{0,1}. Ο ML rule είναι να τοποθετηθεί το x = 0 στο Π_1 και το x = 1 στο Π_2, ορίζοντας τα σύνολα

R_1 = {0}, R_2 = {1} και $R_1 \cup R_2$ = {0,1}.

ΠΑΡΑΔΕΙΓΜΑ 5

Έστω δυο κανονικοί πληθυσμοί

$$\Pi_1 : N\left(\mu_1,\sigma_1^2\right)$$

$$\Pi_2 : N\left(\mu_2,\sigma_2^2\right)$$

είναι:

$$L_i(x)=\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}=(2\pi\sigma_i^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right\}.$$

L_1 η πιθανοφάνεια για $\underset{\sim}{\theta}=\left(\mu_1,\sigma_1^2\right)$ και L_2 η πιθανοφάνεια για $\underset{\sim}{\theta}=\left(\mu_2,\sigma_2^2\right)$.

Ως εκ τούτου το x κατανέμεται στον Π_1 $(x\in R)$ αν $L_1(x)\geq L_2(x)$. Ας σημειωθεί ότι η σχέση $L_1(x)\geq L_2(x)$ είναι ισοδύναμη με τον επόμενο τύπο(2.2):

$$\frac{\sigma_2}{\sigma_1}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]\right\}\geq 1$$

ή

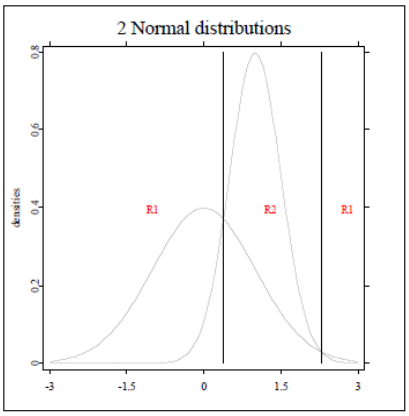
$$x^2\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)-2x\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)+\left(\frac{\mu_1^2}{\sigma_1^2}-\frac{\mu_2^2}{\sigma_2^2}\right)\le 2\log\frac{\sigma_2}{\sigma_1}.$$

Αν υποθέσουμε ότι $\mu_1 = 0$, $\sigma_1 = 1$ και $\mu_2 = 1$, $\sigma_2 = 1/2$ ο τύπος 2.2 οδηγεί σε:

$$R_1=\left\{x: x\le\frac{1}{3}(4-\sqrt{4+6\log 2})\,\acute{\eta}\quad x\ge\frac{1}{3}(4+\sqrt{4+6\log 2})\right\}$$

$$R_2=\mathbb{R}-R_1.$$

Η κατάσταση αυτή φαίνεται στο Γράφημα 2.1



Γράφημα 2.1 Περίπτωση 2 κανονικών πληθυσμών και ML μέθοδος εύρεσης διαχωριστικής ανάλυσης

Με τις κάθετες ευθείες στο εσωτερικό του γραφήματος σημειώνονται τα σημεία

$$x_1=\frac{1}{3}\left(4-\sqrt{4+6\log 2}\right)\ \text{και}\ x_2=\frac{1}{3}\left(4+\sqrt{4+6\log 2}\right).$$

Η κατάσταση απλοποιείται σε περίπτωση ίσων διακυμάνσεων $\sigma_1 = \sigma_2$ και $\mu_1<\mu_2$.

$$x\to\Pi_1\quad\alpha\nu\quad x\in R_1=\left\{x: x\le\frac{1}{2}(\mu_1+\mu_2)\right\}$$

$$x\to\Pi_2\quad\alpha\nu\quad x\in R_2=\left\{x: x>\frac{1}{2}(\mu_1+\mu_2)\right\}.$$

Το Θεώρημα 2.3 δείχνει ότι ο ML κανόνας διαχωρισμού για πολυμεταβλητές κανονικές παρατηρήσεις είναι στενά συνδεδεμένες με την απόσταση Mahalanobis.

Ορισμός της απόστασης Mahalanobis.

Ο ορισμός που δόθηκε από τον Mahalanobis ,1936 ορίζει ως απόσταση Mahalanobis σε πολυμεταβλητές $\boldsymbol{x}=(x_1,x_2,...,x_N)^T$ από ένα σύνολο τιμών, με μέσο $\boldsymbol{\mu}=(\mu_1,\mu_2,...,\mu_N)^T$ και πίνακα συνδιακύμανσης S την:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \quad \text{(M1)}$$

Η απόσταση Mahalanobis μπορεί επίσης να ορισθεί ως ένα μέτρο ανομοιότητας μεταξύ δύο τυχαίων διανυσμάτων $\boldsymbol{x}$ και $\boldsymbol{y}$ με ίδια κατανομή και πίνακα συνδιακύμανσης S :

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T S^{-1}(\boldsymbol{x}-\boldsymbol{y})} \quad \text{(M2).}$$

Αν ο πίνακας συνδιακύμανσης είναι ο μοναδιαίος τότε η απόσταση Mahalanobis ταυτίζεται με την Ευκλείδεια απόσταση.

Όπου ως Ευκλείδεια απόσταση σε χώρο $\mathbb{R}^N$ δύο διανυσμάτων $\boldsymbol{x} = (x_1, x_2, ..., x_N)$ και $\boldsymbol{y} = (y_1, y_2, ..., y_N)$ ορίζουμε την:

$$d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\| = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}\ .$$

Αν ο πίνακας συνδιακύμανσης είναι διαγώνιος τότε η (M2) γίνεται

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{N}\frac{(x_i - y_i)^2}{s_{ii}^2}}$$

όπου s_{ii} είναι η τυπική απόκλιση των x_i και y_i, και ονομάζεται κανονικοποιημένη Ευκλείδεια απόσταση.

Ο κανόνας διαχωρισμού βασίζεται σε γραμμικούς συνδυασμούς και ανήκει στην οικογένεια των γραμμικών μεθόδων της διαχωριστικής ανάλυσης (LDA).

Θεώρημα 2.3. Υποθέτουμε κανονικούς πληθυσμούς με ίσους πίνακες Σ, δηλ.

$$\Pi_i = N_p(\mu_i, \Sigma)$$

Τότε:

α) Ο κανόνας ML τοποθετεί το x στο Π_j όπου j$\in$ {1,…,j} είναι η τιμή που ελαχιστοποιεί το τετράγωνο της Mahalanobis απόστασης μεταξύ της παρατήρησης x από τη μέση τιμή του πληθυσμού (λαμβάνοντας υπόψη και τη διασπορά αυτού):

$$\delta^2(x,\mu_i) = (x-\mu_i)^T\Sigma^{-1}(x-\mu_i), i = 1,...,J$$

β)Στην περίπτωση όπου J = 2,

$$x \in R_1 \Leftrightarrow \alpha^T(x-\mu) \geq 0$$

όπου $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ και $\mu = \frac{1}{2}(\mu_1 + \mu_2)$.

Στην περίπτωση αυτή ταυτίζεται με τον κανόνα του Fisher.

Απόδειξη.

Το μέρος α του Θεωρήματος προκύπτει άμεσα από τη σύγκριση των πιθανοφανειών. Για J = 2 το μέρος α λέει ότι το x τοποθετείται στο Π_1 αν

$$(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) \le (x-\mu_2)^T \Sigma^{-1}(x-\mu_2).$$

Με αναδιάταξη των όρων έχουμε

$$-2\mu_1^T \Sigma^{-1} x + 2\mu_2^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1}\mu_1 - \mu_2^T \Sigma^{-1}\mu_2 \le 0$$

το οποίο ισοδυναμεί με:

$$2(\mu_2 - \mu_1)^T \Sigma^{-1} x + (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2) \le 0 \Rightarrow$$
$$(\mu_1 - \mu_2)^T \Sigma^{-1}\left\{ x - \frac{1}{2}(\mu_1 + \mu_2)\right\} \ge 0 \Rightarrow$$
$$\alpha^T (x - \mu) \ge 0.$$

2.4. Ο Κανόνας διαχωρισμού Bayes.

Συμβολίζουμε ως π_j την εκ των προτέρων γνώση της πιθανότητας ταξινόμησης στον Π_j(prior) με $\sum_{j=1}^{j} \pi_j = 1.$

Ο διαχωριστικός κανόνας Bayes για τον προσδιορισμό του x στον Π_j που δίνει τη μεγαλύτερη τιμή για $\pi_i f_i(x),\ \pi_j f_j(x) = \max_i \pi_i f_i(x)$

Ως εκ τούτου ο κανόνας διαχωρισμού καθορίζεται από $R_j = \left\{ x : \pi_j f_j(x) \ge \pi_i f_i(x), i = 1,..., j \right\}$. Προφανώς ο κανόνας Bayes είναι ταυτόσημος με τον ML διαχωριστικό κανόνα για $\pi_j = \frac{1}{j}$.

Μια άλλη τροποποίηση είναι να προσδιοριστεί το x στον Π_j με μια ορισμένη πιθανότητα $\varphi_j(x)$, τέτοια ώστε για όλα τα x να ισχύει $\sum_{j=1}^{J} \phi_j(x) = 1$.

Αυτό ονομάζεται ένας τυχαιοποιημένος κανόνας διαχωρισμού. Ένας τέτοιος κανόνας είναι μια γενίκευση των διαχωριστικών κανόνων αφού

$$\varphi_j(x)=\begin{cases}1 & \text{αν} \quad \pi_j f_j(x)=\max_i \pi_i f_i(x)\\ 0 & \text{αλλιώς}\end{cases}.$$

Ποιοι διαχωριστικοί κανόνες είναι καλοί; Χρειαζόμαστε ένα μέτρο σύγκρισης.
Συμβολίζουμε με:

$$p_{ij}=\int \varphi_i(x) f_j(x)dx$$

ως την πιθανότητα να προσδιοριστεί το x στο Π_i αν στην πραγματικότητα ανήκει στον Π_j. Ένας κανόνας διαχωρισμού με p_{ij} είναι τόσο καλός όσο κάθε άλλος κανόνας με πιθανότητα p'_{ij} αν

$$p_{ii}\geq p'_{ii} \quad \text{για} \quad \text{κάθε} \quad i=1,...,j. \qquad (2.1)$$

Καλούμε τον πρώτο κανόνα καλύτερο αν η αυστηρή ανισότητα (2.1) ισχύει για τουλάχιστον ένα i.
Ένας κανόνας διαχωρισμού ονομάζεται παραδεκτός, εάν δεν υπάρχει καλύτερος.

ΘΕΩΡΗΜΑ 2.4. Όλοι οι κανόνες διαχωρισμού Bayes (συμπεριλαμβανομένου του κανόνα ML) είναι παραδεκτοί.

2.4.1 Πιθανότητα λανθασμένης ταξινόμησης με ίσους πίνακες συνδιακύμανσης για τον κανόνα ML (J=2).

Ας υποθέσουμε ότι $\Pi_i=N_p(\mu_i,\Sigma)$. Ίσοι πίνακες διασποράς για κάθε υποπληθυσμό.
Στην περίπτωση των δύο ομάδων, δεν είναι δύσκολο να αντλήσουμε τις πιθανότητες εσφαλμένης ταξινόμησης για τον διαχωριστικό κανόνα ML.
Αν θεωρήσουμε $p_{12}=P\left(x\in R_1/\Pi_2\right)$ ως την πιθανότητα ταξινόμησης του x στον R_1 δεδομένου ότι προέρχεται από τον Π_2.

Με το μέρος (β) στο θεώρημα 2.3 έχουμε

$$p_{12}=P\left\{\alpha^T(x-\mu)>0/\Pi_2\right\}.$$

Αν

$$X\in R_2, \alpha^T(X-\mu)\sim N(-\frac{1}{2}\delta^2,\delta^2)$$

όπου

$$\delta^2=(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)$$

είναι το τετράγωνο της Mahalanobis απόστασης μεταξύ των μέσων τιμών δυο πληθυσμών τότε προκύπτει

$$p_{12} = \Phi(-\frac{1}{2}\delta).$$

Ομοίως η πιθανότητα να ταξινομηθεί στον πληθυσμό R_2 αν και προέρχεται από τον πληθυσμό Π_1 ισούται με

$$p_{21} = \Phi(-\frac{1}{2}\delta).$$

2.4.2. Ταξινόμηση με διαφορετικούς πίνακες συνδιακύμανσης

Το ελάχιστο ECM εξαρτάται από το λόγο των πυκνοτήτων $\frac{f_1(x)}{f_2(x)}$

ή ισοδύναμα από τη διαφορά $\ln\{f_1(x)\} - \ln\{f_2(x)\}$.

Όταν η συνδιακύμανση και για τις δυο συναρτήσεις πυκνότητας διαφέρει, ο κανόνας προσδιορισμού γίνεται πιο περίπλοκος

$$R_1 = \left\{ x : \frac{1}{2} x^T \left(\Sigma_1^{-1} - \Sigma_2^{-1}\right) x + \left(\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1}\right) x - k \geq \ln\left[\left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{\pi_2}{\pi_1}\right)\right]\right\}$$

$$R_2 = \left\{ x : \frac{1}{2} x^T \left(\Sigma_1^{-1} - \Sigma_2^{-1}\right) x + \left(\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1}\right) x - k < \ln\left[\left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{\pi_2}{\pi_1}\right)\right]\right\}$$

όπου $k = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_2^T \Sigma_2^{-1} \mu_2)$.

Οι περιοχές ταξινόμησης καθορίζονται από τετραγωνικές συναρτήσεις. Ως εκ τούτου, ανήκουν στην οικογένεια των τετραγωνικών μεθόδων διαχωριστικής ανάλυσης (QDA). Αυτός ο τετραγωνικός κανόνας ταξινόμησης συμπίπτει με τους κανόνες που χρησιμοποιούνται κατά την $\Sigma_1 = \Sigma_2$, δεδομένου ότι ο όρος $\frac{1}{2} x^T (\Sigma_1^{-1} - \Sigma_2^{-1}) x$

εξαφανίζεται.

2.5 Οι διαχωριστικοί κανόνες στην πράξη.

Ο κανόνας ML χρησιμοποιείται εάν η κατανομή των στοιχείων είναι γνωστών παραμέτρων. Ας υποθέσουμε για παράδειγμα ότι τα δεδομένα που προέρχονται από την πολυμεταβλητή κανονική κατανομή $N_p(\mu_j, \Sigma)$. Αν έχουμε J ομάδες με n_j

παρατηρήσεις σε κάθε ομάδα, χρησιμοποιούμε το $\overline{x}_j$ για την εκτίμηση των μ_j, και Σ_j για την εκτίμηση της συνδιακύμανσης Σ.

Η κοινή συνδιακύμανση μπορεί να εκτιμηθεί από

$$\Sigma_u = \sum_{j=1}^{J} n_j \left(\frac{\Sigma_j}{n-J} \right) \text{ με } n = \sum_{j=1}^{J} n_j .$$

Έτσι, η εμπειρική εκδοχή του ML κανόνα του Θεωρήματος 2.2 είναι για να ταξινομήσει μια νέα παρατήρηση x στον Π_j τέτοια ώστε να ελαχιστοποιεί το

$\left(x-\overline{x}_i\right)^T \Sigma_u^{-1} \left(x-\overline{x}_i\right)$ για $i = 1,...,J$.

2.5.1 Εκτίμηση πιθανοτήτων εσφαλμένης ταξινόμησης

Οι πιθανότητες εσφαλμένης ταξινόμησης δίνονται από

$$p_{ij} = \int \varphi_i(x) f_j(x) dx$$

και εκτιμώνται αντικαθιστώντας την άγνωστη παράμετρο από τον αντίστοιχο εκτιμητή της.

Για τον κανόνα ML για δυο κανονικούς πληθυσμούς προκύπτει $\hat{p}_{12} = \hat{p}_{21} = \Phi\left(-\frac{1}{2}\hat{\delta}\right)$

όπου $\hat{\delta}^2 = \left(\overline{x}_1 - \overline{x}_2\right)^T \Sigma_u^{-1} \left(\overline{x}_1 - \overline{x}_2\right)$ είναι ο εκτιμητής του δ^2.

Οι πιθανότητες της εσφαλμένης ταξινόμησης μπορεί επίσης να εκτιμώνται με τη μέθοδο της εκ νέου αντικατάστασης.

Ανακατατάσσεται κάθε αρχική παρατήρηση x_i, $i = 1,...,n$ σε $\Pi_1,...,\Pi_j$ σύμφωνα με το επιλεγμένο κανόνα. Στη συνέχεια, δηλώνεται ο αριθμός των ατόμων που προέρχονται από το Π_j τα οποία έχουν ταξινομηθεί στο Π_i από n_{ij} έχουμε $\hat{p}_{ij} = \frac{n_{ij}}{n_j}$

έναν εκτιμητή του p_{ij}.

Προφανώς, η μέθοδος αυτή οδηγεί σε υπερβολικά αισιόδοξους εκτιμητές του p_{ij}, αλλά παρέχει μια πρόχειρη μέτρηση της ποιότητας του διαχωριστικού κανόνα.

ΠΑΡΑΔΕΙΓΜΑ 2.4

Πρόβλημα ταξινόμησης των χαρτονομισμάτων της Ελβετικής τράπεζας.

Χρησιμοποιήθηκαν 100 γνήσια και 100 πλαστά Ελβετικά χαρτονομίσματα των 1000 φράγκων .

Ελήφθησαν μετρήσεις για 6 μεταβλητές:

X_1 : Μήκος χαρτονομίσματος

X_2 : Ύψος από τα αριστερά

X_3 : Ύψος από τα δεξιά

X_4 : Απόσταση εσωτερικού πλαισίου στο κάτω άκρο

X_5 : Απόσταση εσωτερικού πλαισίου στο πάνω άκρο

X_6 : Μήκος διαγωνίου

(δεδομένα από Flury και Riedwyl (1988))

και προέκυψε ο ακόλουθος πίνακας:

		true membership	
		genuine(Π_1)	countrtfeit(Π_2)
predicted	Π_1	100	1
	Π_2	0	99

Πίνακας 2.2 Ταξινόμηση ελβετικών χαρτονομισμάτων.

Το προφανές ποσοστό σφάλματος ($APER$) ορίζεται ως το κλάσμα των παρατηρήσεων που έχουν ταξινομηθεί εσφαλμένα.

Το $APER$, εκφρασμένο ως ποσοστό, είναι:

$$APER = \left(\frac{1}{200} \right) 100\% = 0.5\%.$$

Για τον υπολογισμό του $APER$ χρησιμοποιούμε τις παρατηρήσεις δύο φορές: την πρώτη φορά για να κατασκευάσουμε τον κανόνα ταξινόμησης και τη δεύτερη φορά για την αξιολόγηση αυτού του κανόνα. Ένα $APER$ του 0.5% θα μπορούσε συνεπώς να είναι υπερβολικά αισιόδοξο. Μια προσέγγιση που διορθώνει αυτή την απόκλιση βασίζεται στη διαδικασία holdout των Lachenbruch και Mickey (1968). Για δύο πληθυσμούς αυτή η διαδικασία έχει ως εξής:

1. Ξεκινάμε με το πρώτο Π_1 πληθυσμό. Παραλείπουμε μια παρατήρηση και αναπτύσσουμε τον κανόνα κατάταξης με βάση τις εναπομείνασες παρατηρήσεις $n_1 - 1, n_2$.

2. Ταξινομούμε τη "holdout" παρατήρηση χρησιμοποιώντας τον κανόνα των διακρίσεων στο Βήμα 1.

3. Επαναλαμβάνουμε τα βήματα 1 και 2 έως ότου όλες οι Π_1 παρατηρήσεις είναι ταξινομημένες. Μετρήστε τον αριθμό των n'_{21} εσφαλμένα ταξινομημένων παρατηρήσεων.

4. Επαναλάβετε τα βήματα 1 έως 3 για τον πληθυσμό Π_2. Μετρήστε τον αριθμό των n'_{12} εσφαλμένα ταξινομημένων παρατηρήσεων.

Οι εκτιμήσεις των πιθανοτήτων εσφαλμένης ταξινόμησης δίνονται από

$$\hat{p}'_{12} = \frac{n'_{12}}{n_2}$$

$$\hat{p}'_{21} = \frac{n'_{21}}{n_1}$$

Ένας πιο ρεαλιστικός εκτιμητής ενός πραγματικού ποσοστού (AER) σφαλμάτων δίνεται από

$$\frac{n'_{12} + n'_{21}}{n_2 + n_1}$$

Οι ερευνητές στη στατιστική (James W. Sayre, Neil J. Durans) κλίνουν υπέρ της AER (για την αμεροληψία της) παρά στο $APER$. Σε μεγάλα δείγματα, ωστόσο, το υπολογιστικό κόστος θα μπορούσε να αντισταθμίσει το στατιστικό πλεονέκτημα. Αυτό δεν είναι ένα πραγματικό πρόβλημα δεδομένου ότι τα δύο μέτρα ταξινόμησης είναι ασυμπτωτικά ισοδύναμα.

ΚΕΦΑΛΑΙΟ 3

ΠΕΡΙΓΡΑΦΗ

3 Correspondence Analysis

Η Correspondence Ανάλυση, (Hardle & Simar 2003, κεφ. 13) παρέχει εργαλεία για την ανάλυση των σχέσεων μεταξύ γραμμών και στηλών των πινάκων συνάφειας. Πίνακας συνάφειας είναι ένας πίνακας συχνοτήτων δύο εισόδων όπου αναφέρονται οι κοινές συχνότητες των δύο ποιοτικών μεταβλητών. Για παράδειγμα, ένας (2 × 2) πίνακας θα μπορούσε να διαμορφωθεί με την παρατήρηση από ένα δείγμα n ατόμων δύο ποιοτικών μεταβλητών π.χ:

X_1 : του φύλου του ατόμου, και

X_2 : εάν το άτομο καπνίζει.

Ο πίνακας αναφέρει τις παρατηρούμενες κοινές συχνότητες. Σε γενικές γραμμές μπορεί να θεωρηθούν πίνακες ($n \times p$).

Η κύρια ιδέα της correspondence ανάλυσης είναι η ανάπτυξη νέων μεταβλητών που θα δείχνουν τις σχέσεις μεταξύ των γραμμών και των κατηγοριών ανά στήλη. Οι δείκτες αυτοί θα μας πουν ταυτόχρονα τις κατηγορίες στηλών που έχουν περισσότερο βάρος σε μια κατηγορία γραμμής και αντίστροφα.

Η correspondence ανάλυση σχετίζεται επίσης με το ζήτημα του περιορισμού της διάστασης του πίνακα, θέμα παρόμοιο μ΄αυτό της ανάλυση κυριών συνιστωσών (PCA).

Η ιδέα είναι η δημιουργία των νέων αυτών μεταβλητών κατά φθίνουσα σειρά σπουδαιότητας έτσι ώστε η κύρια πληροφορία του πίνακα να μπορεί να συνοψιστεί σε χώρους με μικρότερες διαστάσεις. Για παράδειγμα, εάν χρησιμοποιούνται μόνο δύο παράγοντες (δείκτες), τα αποτελέσματα μπορεί να δειχθούν σε γραφήματα δύο διαστάσεων, που δείχνουν τη σχέση μεταξύ των σειρών και των στηλών του πίνακα. Η παράγραφος 3.1 ορίζει το βασικό συμβολισμό και η παράγραφος 3.2 δίνει την βασική θεωρία.

Οι νέες μεταβλητές θα πρέπει να χρησιμοποιούνται για να περιγράψουν την X^2 στατιστική μέτρηση των ενώσεων στον πίνακα.

Αρκετά παραδείγματα δείχνουν πώς να παρέχουν και να ερμηνεύουν, στην πράξη, γραφήματα δύο διαστάσεων εμφανίζοντας τη σχέση μεταξύ των σειρών και των στηλών ενός πίνακα συνάφειας.

3.1 Κίνητρα

Ο στόχος της correspondence ανάλυσης είναι η ανάπτυξη νέων μεταβλητών που δείχνουν τις σχέσεις μεταξύ των γραμμών και των στηλών ενός πινάκα συνάφειας.

Οι πίνακες συνάφειας είναι πολύ χρήσιμοι για να περιγράψουν τη σχέση μεταξύ των δύο μεταβλητών σε πολύ γενικές καταστάσεις. Οι δύο μεταβλητές μπορεί να είναι ποιοτικές (ονομαστικές), στην οποία περίπτωση αναφέρονται ως κατηγορικές μεταβλητές.

Κάθε γραμμή και κάθε στήλη του πίνακα αντιστοιχεί σε μια κατηγορία της αντίστοιχης μεταβλητής.

Η εγγραφή x_{ij} στον πίνακα X (με διάσταση ($n \times p$)) είναι ο αριθμός των παρατηρήσεων σε ένα δείγμα η οποία εμπίπτει ταυτόχρονα στην i γραμμή και στην j στήλη,

για $i = 1,...,n$ και $j = 1,...,p$.

Μερικές φορές μια «κατηγορία», με ονομαστική μεταβλητή ονομάζεται επίσης "τυπικότητα" της μεταβλητής.

Οι μεταβλητές των τόκων μπορεί επίσης να είναι διακριτές ποσοτικές μεταβλητές, όπως ο αριθμός των μελών της οικογένειας ή του αριθμού των ατυχημάτων μιας ασφαλιστικής εταιρείας που πρέπει να καλυφθούν κατά τη διάρκεια ενός έτους, κ.λπ.

Εδώ, κάθε δυνατή τιμή που η μεταβλητή μπορεί να πάρει, καθορίζει μια κατηγορία σε μια σειρά ή στήλη. Μπορούν να ληφθούν υπόψη είτε κατηγορικές είτε συνεχείς που έχουν κατηγοριοποηθεί τα αποτελέσματά τους.

Έτσι πίνακες συνάφειας μπορούν να χρησιμοποιηθούν σε πολλές καταστάσεις, πράγμα που σημαίνει ότι η correspondence ανάλυση είναι ένα πολύ χρήσιμο εργαλείο σε πολλές εφαρμογές.

Η γραφική αναπαράσταση για τη σχέση μεταξύ των σειρών και των στηλών του πίνακα X που προκύπτουν από την correspondence ανάλυση βασίζεται στην ιδέα του να περιέχουν οι νέες μεταβλητές όλες τις κατηγορίες γραμμών και στηλών σε μια μικρότερη διάσταση και στην ερμηνεία των σχετικών θέσεων των σημείων από την άποψη των συντελεστών στάθμισης που αντιστοιχεί στη στήλη και τη σειρά.

Αυτό επιτυγχάνεται με το νέο σύνολο δεδομένων που θα προκύψει υπολογίζοντας τις συντεταγμένες της κάθε γραμμής και κάθε στήλης πάνω στις μεταβλητές. Οι συντεταγμένες αυτές λέγονται και scores των γραμμών και στηλών και η γραφική τους παράσταση δίνει τη σχετική τους θέση σε 2 ή 3 διαστάσεις.

Στη συνέχεια είναι σαφές να δούμε ποιες κατηγορίες στηλών είναι πιο σημαντικές για τις κατηγορίες σειρών του πίνακα (και το αντίστροφο).

Όπως είχε ήδη αναφερθεί η κατασκευή των δεικτών βασίζεται σε μια ιδέα παρόμοια με αυτή της PCA.

Χρησιμοποιώντας PCA η συνολική διακύμανση χωρίστηκε σε ανεξάρτητες συνεισφορές που απορρέουν από τις κύριες συνιστώσες.

Στην correspondence ανάλυση, από την άλλη πλευρά, η κατασκευή των νέων μεταβλητών γίνεται με κριτήριο την ερμηνεία (διάσπαση) της X^2 ποσότητας που αντιστοιχεί στον εν λόγω πίνακα συνάφειας.

ΠΑΡΑΔΕΙΓΜΑ 3.1 Ας θεωρήσουμε n κατηγορίες εταιρειών και τις p τοποθεσίες αυτών των εταιρειών. Υπάρχει ένας συγκεκριμένος τύπος εταιρείας που προτιμά μια συγκεκριμένη τοποθεσία; Ή υπάρχει ένας δείκτης θέσης που αντιστοιχεί σε ένα συγκεκριμένο είδος της εταιρείας εκεί;

Ας υποθέσουμε ότι n = 3, p = 3, και ότι οι συχνότητες που έχουν είναι ως εξής:

$$X = \begin{pmatrix} 4 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{matrix} \leftarrow \text{ Finance} \\ \leftarrow \text{ Energy} \\ \leftarrow \text{ HiTech} \end{matrix}$$

$$\begin{matrix} \uparrow \text{ Frankfurt} \\ \uparrow \text{ Berlin} \\ \uparrow \text{ Munich} \end{matrix}$$

Ο παραπάνω πίνακας σημαίνει ότι π.χ: τέσσερις εταιρείες τύπου 3 (Hi Tech) είναι στη θέση 3 (Μόναχο), και ούτω καθεξής. Ας υποθέσουμε ότι υπάρχει ένα διάνυσμα βάρους (weight vector) για τις εταιρείες $\boldsymbol{r} = (r_1, ..., r_n)^T$, έτσι ώστε ένας δείκτης θέσης s_j θα μπορούσε να οριστεί ως

$$s_j = c \sum_{i=1}^{n} r_j \frac{x_{ij}}{x_{\cdot j}} \qquad (3.1.1)$$

όπου

$$x_{\cdot j} = \sum_{i=1}^{n} x_{ij}$$

είναι ο αριθμός των εταιρειών στην περιοχή j και c είναι μια σταθερά.

Το s_1, για παράδειγμα θα δίνει τη μέση σταθμισμένη συχνότητα (από r) των εταιρειών στη τοποθεσία 1 (Frankfurt).

Λαμβάνοντας υπόψη ένα διάνυσμα βάρους(weight vector) για τις τοποθεσίες

$$\boldsymbol{s}^* = (s_1^*, ..., s_p^*)^T$$

μπορούμε να ορίσουμε ένα δείκτη για την εταιρεία με τον ίδιο τρόπο όπως και πριν:

$$\boldsymbol{r}_i^* = c^* \sum_{j=1}^{p} s_j^* \frac{x_{ij}}{x_{i\bullet}} \qquad (3.1.2)$$

όπου c^* είναι μια σταθερά και $x_{i\bullet} = \sum_{j=1}^{p} x_{ij}$ είναι το άθροισμα της i-οστής σειράς του X, δηλαδή, ο αριθμός των τύπου i εταιρειών. Έτσι r_2^*, για παράδειγμα, θα δώσει τη μέση σταθμισμένη συχνότητα (από s) των εταιρειών ενέργειας.

Αν οι 3.1.1 και 3.1.2 μπορούν να λυθούν ταυτόχρονα για ένα " row weight " διάνυσμα $\boldsymbol{r} = (r_1, ..., r_n)^T$ και ένα " column weight " διάνυσμα $\boldsymbol{s} = (s_1, ..., s_p)^T$, μπορούμε να αντιπροσωπεύσουμε κάθε σειρά από το r_i, $i = 1, ..., n$ και κάθε στήλη από το s_j, $j = 1, ..., p$ σε ένα μονοδιάστατο γράφημα.

Αν σε αυτό το γράφημα τα r_i και s_j είναι σε κοντινή απόσταση (μακριά από την προέλευση), αυτό θα έδειχνε ότι η i-οστή γραμμή έχει μια σημαντική συχνότητα $\frac{x_{ij}}{x_{\bullet j}}$ στην (3.1.1) και ότι η j-οστή στήλη έχει σημαντική συχνότητα $\frac{x_{ij}}{x_{i\bullet}}$ στην (3.1.2).

3.2 Αποσύνθεση του X^2 (X^2- Decomposition)

Ένας εναλλακτικός τρόπος μέτρησης της σύνδεσης μεταξύ των γραμμών και των στηλών είναι μια αποσύνθεση της τιμής του στατιστικού αποτελέσματος του X^2-test. Το γνωστό X^2-test για την ανεξαρτησία σε ένα δισδιάστατο πίνακα συνάφειας αποτελείται από δύο βήματα. Πρώτον η αναμενόμενη τιμή του κάθε στοιχείου του πίνακα υπολογίζεται με την υπόθεση της ανεξαρτησίας. Δεύτερον, οι αντίστοιχες τιμές που παρατηρούνται συγκρίνονται με τις αναμενόμενες τιμές χρησιμοποιώντας το στατιστικό

$$t = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{\left(x_{ij} - E_{ij}\right)^2}{E_{ij}}, \qquad (3.2.1)$$

όπου x_{ij} είναι η παρατηρούμενη συχνότητα του στοιχείου (i, j) και E_{ij} είναι η προβλεπόμενη αντίστοιχη αναμενόμενη τιμή με βάση την παραδοχή της ανεξαρτησίας, δηλαδή,

$$E_{ij} = \frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}.$$

Εδώ $x_{..} = \sum_{i=1}^{n} x_{i.}$ Υπό την υπόθεση ανεξαρτησίας το t ακολουθεί X^2 κατανομή με $(n-1)(p-1)$ βαθμούς ελευθερίας.

Στο παράδειγμα της βιομηχανικής τοποθεσίας που έγινε εισαγωγή πάνω από την τιμή του $t = 6.26$ έχει επίπεδο σημαντικότητας σχεδόν 5%. Είναι επομένως χρήσιμη η διερεύνηση των ειδικών λόγων για την παρέκκλιση από την ανεξαρτησία.

Η μέθοδος της X^2 decomposition αποτελείται από την εύρεση του SVD (singular value decomposition) του πίνακα $C(n\text{x}p)$ με στοιχεία

$$c_{ij} = \frac{(x_{ij} - E_{ij})}{E_{ij}^{\frac{1}{2}}} \qquad (3.2.2).$$

Τα στοιχεία c_{ij} μπορούν να θεωρηθούν ως μέτρηση των (σταθμισμένων) παρεκκλίσεων μεταξύ των παρατηρούμενων x_{ij} και των θεωρητικών τιμών E_{ij} υπό την προϋπόθεση της ανεξαρτησίας.

Για απλούστευση καθορίζονται οι πίνακες $A(n\text{x}n)$ και $B(p\text{x}p)$ ως:

$$A = diag(x_{i.}) \text{ και } B = diag(x_{.j}) .(\text{Παράρτημα Α}).$$

Οι πίνακες παρέχουν τις οριακές συχνότητες της σειράς $a(n\text{x}1)$ και τις οριακές συχνότητες της στήλης $b(p\text{x}1)$ με $a = A1_n$ και $b = B1_p$.

Είναι εύκολο να επαληθεύσουμε ότι $C\sqrt{b} = 0$ και $C^T\sqrt{a} = 0$.

Όπου η τετραγωνική ρίζα του διανύσματος λαμβάνεται στοιχείο προς στοιχείο και $R = rank(C) \le \min\{(n-1),(p-1)\}$.(Παράρτημα Α).

Θεώρημα 3.1

Έστω ο πίνακας X. Οι πίνακες X^TX και XX^T έχουν τις ίδιες μη μηδενικές ιδιοτιμές $\lambda_1,...,\lambda_r$ με $r = rank(X)$.(Hardle & Simar, 2003 σελ 227)

Τα ιδιονύσματα u_k και v_k των X^TX και XX^T αντίστοιχα ($k \le r$) συνδέονται με:

$$u_k = \frac{1}{\sqrt{\lambda_k}} X^T v_k \text{ και } v_k = \frac{1}{\sqrt{\lambda_k}} X u_k .$$

Οι συντεταγμένες που αντιστοιχούν στις μεταβλητές (στήλες) του X σ΄ένα p-διαστάσεων χώρο μπορούν να υπολογισθούν από:

$$w_k = X^T v_k = \frac{1}{\sqrt{\lambda_k}} X^T X u_k = \sqrt{\lambda_k} u_k . \qquad (3.2.3)$$

Τα u_k και v_k παρέχουν το SVD του X . Δηλαδή αν $U = [u_1 u_2 ... u_r]$, $V = [v_1 v_2 ... v_r]$ και

$\Lambda = diag(\lambda_1, ..., \lambda_r)$ τότε $X = V \Lambda^{\frac{1}{2}} U^T$ έτσι ώστε $x_{ij} \sum_{k=1}^{r} \lambda_k^{\frac{1}{2}} v_{ik} u_{jk}$.

Το SVD (Singular Value Decomposition) του C είναι

$$C = \Gamma \Lambda \Delta^T \qquad (3.2.4)$$

όπου Γ περιέχει τα ιδιοδιανύσματα του CC^T, Δ περιέχει τα ιδιοδιανύσματα του $C^T C$ και $\Lambda = \text{diag}(\lambda_1^{\frac{1}{2}}, ..., \lambda_R^{\frac{1}{2}})$ με $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_R$ οι ιδιοτιμές του CC^T όπου R η maximum διάσταση.

Ο (3.2.2) γίνεται $c_{ij} = \sum_{l=1}^{R} \lambda_k^{\frac{1}{2}} \gamma_{ik} \delta_{jk}$. (3.2.5)

Το (3 2.1) μπορεί να γραφτεί

$$tr(CC^T) = \sum_{k=1}^{R} \lambda_k = \sum_{i=1}^{n} \sum_{j=1}^{p} c_{ij}^2 = t .$$

Αυτή η σχέση δείχνει ότι η SVD του C αποσυνθέτει τη συνολική τιμή X^2 και όχι, τη συνολική διακύμανση.

Η δυαδικότητα των σχέσεων μεταξύ της γραμμής και του χώρου της στήλης δίνεται τώρα για $k = 1, ..., R$ από:

$$\delta_k = \frac{1}{\sqrt{\lambda_k}} C^T \gamma_k ,$$
$$\gamma_k = \frac{1}{\sqrt{\lambda_k}} C \delta_k .$$

Οι προβολές των γραμμών και των στηλών του C δίνονται από

$$C\delta_k = \sqrt{\lambda_\kappa} \gamma_k ,$$
$$C^T \gamma_k = \sqrt{\lambda_k} \delta_k . \qquad (3.2.6)$$

Σημειωτέον ότι τα ιδιοδιανύσματα ικανοποιούν τις

$$\delta_k^T \sqrt{b} = 0, \gamma_k^T \sqrt{a} = 0.$$

Από τον (3.2.5) βλέπουμε ότι τα ιδιοδιανύσματα δ_k και γ_k είναι τα αντικείμενα που παρουσιάζουν ενδιαφέρον κατά την correspondence ανάλυση μεταξύ των σειρών και των στηλών. Ας υποθέσουμε ότι η πρώτη ιδιοτιμή στο 3.2.5 κατέχει δεσπόζουσα θέση, έτσι ώστε $c_{ij} \approx \lambda_1^{\frac{1}{2}} \gamma_{i1} \delta_{j1}$.

Σε αυτή την περίπτωση, όταν οι συντεταγμένες γ_{i1} και δ_{j1} είναι και οι δυο μεγάλες (με το ίδιο πρόσημο) σε σχέση με τις άλλες συντεταγμένες, τότε το c_{ij} θα είναι μεγάλο, καθώς, δείχνοντας μια θετική σχέση μεταξύ της σειράς i και της j στήλης στον πίνακα συνάφειας. Αν γ_{i1} και δ_{j1} ήταν και οι δύο μεγάλες με αντίθετα πρόσημα, τότε θα υπήρχε μια αρνητική συσχέτιση μεταξύ της γραμμής i και της στήλης j.

Σε πολλές εφαρμογές, οι δύο πρώτες ιδιοτιμές, λ_1 και λ_2, κυριαρχούν και το ποσοστό του συνολικού X^2 που εξηγείται από τα ιδιοδιανύσματα γ_1 και γ_2 και δ_1 και δ_2 είναι μεγάλο. Στην περίπτωση (3.2.6) τα (γ_1, γ_2) μπορούν να χρησιμοποιηθούν για την απόκτηση μιας γραφικής απεικόνισης των n γραμμών του πίνακα (τα (δ_1, δ_2) διαδραματίζουν παρόμοιο ρόλο για τις p στήλες του πίνακα).

Αυτό ισχύει, δηλ 1 ή 2 ιδιοτιμές είναι αρκετές, γιατί το R είναι μικρό $(\min\{i-1,\ j-1\})$.

Η ερμηνεία της μικρής απόστασης μεταξύ των γραμμών και των σημείων των στηλών θα πρέπει να ερμηνευθεί ως ανωτέρω όσον αφορά το (3.2.5).

Στην correspondence ανάλυση, χρησιμοποιούμε τις προβολές των σταθμισμένων σειρών του C και τις προβολές των σταθμισμένων στηλών του C για γραφικές απεικονίσεις. Έστω $r_k(n \text{ x} 1)$ είναι οι προβολές των $A^{-\frac{1}{2}}C$ στο δ_k και $s_k(p \text{ x} 1)$ είναι οι προβολές του $B^{-\frac{1}{2}}C^T$ στις γ_k με $k = 1, ..., R$:

$$r_k = A^{-\frac{1}{2}} C \delta_k = \sqrt{\lambda_k} A^{-\frac{1}{2}} \gamma_k, \qquad s_k = B^{-\frac{1}{2}} C^T \gamma_k = \sqrt{\lambda_k} B^{-\frac{1}{2}} \delta_k. \tag{3.2.7}$$

Αυτά τα διανύσματα έχουν την ιδιότητα ότι

$$r_k^T a = 0,$$
$$s_k^T b = 0.$$

Οι προβολές που λαμβάνονται σε κάθε άξονα $k = 1,...,R$ με επίκεντρο στο μηδέν με το φυσικό βάρος δίνεται από το a (οι οριακές συχνότητες των γραμμών του X) για τις συντεταγμένες των γραμμών r_k και από το b (η οριακή συχνότητα των στηλών X) για για τις συντεταγμένες των στηλών s_k. Ως αποτέλεσμα, η προέλευση είναι το κέντρο βάρους για όλες τις παραστάσεις.

Ξέρουμε επίσης ότι:

$$r_k^T A r_k = \lambda_k,$$
$$s_k^T B s_k = \lambda_k.$$

Από τη δυαδικότητα των σχέσεων μεταξύ δ_k και γ_k προκύπτει:

$$r_k = \frac{1}{\sqrt{\lambda_k}} A^{-\frac{1}{2}} C B^{\frac{1}{2}} s_k,$$
$$s_k = \frac{1}{\sqrt{\lambda_k}} B^{-\frac{1}{2}} C A^{\frac{1}{2}} r_k,$$

το οποίο μπορεί να απλοποιηθεί σε:

$$r_k = \sqrt{\frac{x_{..}}{\lambda_k}} A^{-1} X s_k,$$
$$s_k = \sqrt{\frac{x_{..}}{\lambda_k}} B^{-1} X^T r_k.$$

Αυτά τα διανύσματα ικανοποιούν τις σχέσεις (3.1.1) και (3.1.2) για κάθε $k = 1,...,R$ ταυτόχρονα. Τα διανύσματα r_k και s_k αναφέρονται ως παράγοντες (παράγοντας σειράς και παράγοντας στήλης, αντίστοιχα). Έχουν τα ακόλουθα μέσα και διακυμάνσεις:

$$\overline{r_k} = \frac{1}{x_{..}} r_k^T a = 0,$$
$$\overline{s_k} = \frac{1}{x_{..}} s_k^T b = 0,$$

και

$$Var(r_k) = \frac{1}{x_{..}} \sum_{i=1}^{n} x_{i.} r_{ki}^2 = \frac{r_k^T A r_k}{x_{..}} = \frac{\lambda_k}{x_{..}},$$
$$Var(s_k) = \frac{1}{x_{..}} \sum_{j=1}^{p} x_{.j} s_{kj}^2 = \frac{s_k^T B s_k}{x_{..}} = \frac{\lambda_k}{x_{..}}.$$

Ως εκ τούτου το, $\frac{\lambda_k}{\sum_{j=1}^{R} \lambda_j}$, που είναι το μέρος του k-ιοστού παράγοντα στη αποσύνθεση του στατιστικού t της X^2, μπορεί επίσης να ερμηνευθεί ως το ποσοστό της διακύμανσης που εξηγείται από το συντελεστή k.

Οι αναλογίες $C_a(i, r_k) = \frac{x_{i.} r_{ki}^2}{\lambda_k}$ για $i = 1, ..., n, k = 1, ..., R$ καλούνται η απόλυτη συμβολή της σειράς i για τη διακύμανση του παράγοντα r_k. Δείχνουν τις κατηγορίες σειρών που είναι οι πιο σημαντικές στη διασπορά του k-ιοστου παράγοντα της σειράς.

Ομοίως, οι αναλογίες $C_a(j, s_k) = \frac{x_{.j} s_{ki}^2}{\lambda_k}$ για $j = 1, ..., p, k = 1, ..., R$ καλούνται απόλυτη συμβολή της j στήλης για τη διακύμανση του παράγοντας s_k. Αυτές οι απόλυτες εισφορές μπορούν να βοηθήσουν να γίνει αντιληπτή η γραφική παράσταση που λαμβάνεται με την correspondence ανάλυση.

3.3 Correspondence Analysis in Practice

Οι γραφικές παραστάσεις στους άξονες $k = 1, ..., R$ των σειρών n και των στηλών p του X παρέχονται από τα στοιχεία των r_k και s_k. Συνήθως, δύο διαστάσεων πίνακες είναι συχνά ικανοποιητικοί αν το συνολικό ποσοστό της διακύμανσης που εξηγείται από τους δύο πρώτους παράγοντες, $\psi_2 = \frac{\lambda_1 + \lambda_2}{\sum_{\kappa=1}^{R} \lambda_\kappa}$, είναι αρκετά μεγάλο.

Η ερμηνεία των γραφημάτων συνοψίζεται ως εξής:

-Η εγγύτητα των δύο σειρών (δύο στηλών) δείχνει παρόμοιο προφίλ σε αυτές τις δύο γραμμές (δύο στήλες), όπου "προφίλ" αναφέρεται στην υπό συνθήκη κατανομή συχνοτήτων μιας γραμμής (στήλης) Οι δύο αυτές σειρές (στήλες) είναι σχεδόν

ανάλογες. Αντίθετη ερμηνεία ισχύει όταν οι δύο γραμμές (δύο στήλες) είναι πολύ μακριά.

-Η εγγύτητα μιας συγκεκριμένης γραμμής σε μια συγκεκριμένη στήλη υποδεικνύει ότι αυτή η γραμμή (στήλη) έχει έναν ιδιαίτερα σημαντικό βάρος σε αυτή τη στήλη (γραμμή). Σε αντίθεση με αυτό, μια σειρά που είναι αρκετά μακριά από μια συγκεκριμένη στήλη δείχνει ότι σχεδόν δεν υπάρχουν παρατηρήσεις σε αυτή τη στήλη για αυτή τη σειρά (και αντίστροφα).

Φυσικά, όπως αναφέρθηκε παραπάνω, τα συμπεράσματα αυτά είναι ιδιαίτερα αληθή όταν τα σημεία βρίσκονται μακριά από το 0.

-Το σημείο αναφοράς είναι ο μέσος όρος των συντελεστών r_k και s_k . Ως εκ τούτου, ένα συγκεκριμένο σημείο (γραμμή ή στήλη) που προβάλλεται κοντά στο σημείο αναφοράς δείχνει ένα μέσο προφίλ.

-Η απόλυτη συμβολή χρησιμοποιείται για την αξιολόγηση του βάρους της κάθε γραμμής (στήλης) στις διακυμάνσεις των παραγόντων.

-Όλες οι ερμηνείες που περιγράφονται ανωτέρω πρέπει να πραγματοποιούνται εν όψει της ποιότητας της γραφικής παράστασης η οποία αξιολογείται, όπως και στην PCA, χρησιμοποιώντας το σωρευτικό ποσοστό της διακύμανσης.

ΠΑΡΑΤΗΡΗΣΗ 3.1 Ας σημειωθεί ότι η correspondence ανάλυση μπορεί επίσης να εφαρμοστεί σε πιο γενικούς (n x p) πίνακες X οι οποίοι σε μια αυστηρή έννοια του όρου δεν είναι πίνακες συνάφειας.

Όσο η στατιστική (ή φυσική) έννοια μπορεί να δοθεί σε αθροίσματα γραμμών και στηλών, η παρατήρηση 3.1 ισχύει.

Αυτό συνεπάγεται, μεταξύ άλλων, ότι όλες οι μεταβλητές μετρώνται στις ίδιες μονάδες. Στην περίπτωση αυτή, αποτελεί τη συνολική συχνότητα του φαινομένου που παρατηρήθηκε, και είναι κοινή μεταξύ των ατόμων (n γραμμές) και μεταξύ των μεταβλητών (p στήλες). Απεικονίσεις των γραμμών και των στηλών του X , r_k και s_k, έχουν τη βασική ιδιότητα (3.2.6) και δείχνουν ποιες μεταβλητές έχουν σημαντικό βάρος για κάθε άτομο και το αντίστροφο. Αυτού του είδους η ανάλυση χρησιμοποιείται ως εναλλακτική λύση στην PCA. Η PCA αφορά κυρίως συνδιασπορές και συσχετίσεις, ενώ η correspondence ανάλυση αναλύει ένα γενικότερο είδος συσχέτισης.

Παράδειγμα 3.1

Μια έρευνα σε Βέλγους πολίτες που διαβάζουν τακτικά μια εφημερίδα διεξήχθη στη δεκαετία του 1980. Ερωτήθηκαν που ζούσαν. Οι πιθανές απαντήσεις ήταν 10 περιφέρειες: 7 επαρχίες (Αμβέρσα, Δυτική Φλάνδρα, Ανατολική Φλάνδρα, Hainant, Λιέγη, Λίμπουργκ, Λουξεμβούργο) και 3 περιοχές γύρω από τις Βρυξέλλες (Φλαμανδική-Brabant, Βαλλονική-Brabant και την πόλη των Βρυξελλών). Τους ζητήθηκε επίσης το είδος των εφημερίδων που διαβάζουν σε τακτική βάση. Υπήρχαν 15 πιθανές απαντήσεις που χωρίζονταν σε 3 κατηγορίες: Φλαμανδική εφημερίδες (ετικέτα αρχίζει με το γράμμα v), γαλλικές εφημερίδες (ετικέτα ξεκινά με f) και τις δύο γλώσσες μαζί (ετικέτα ξεκινά με b).

Το σύνολο των στοιχείων δίδεται στον παρακάτω Πίνακα 3.1.

X_1:	WaBr	Walloon Brabant
X_2:	Brar	Brussels area
X_3:	Antw	Antwerp
X_4:	FlBr	Flemish Brabant
X_5:	OcFl	Occidental Flanders
X_6:	OrFl	Oriental Flanders
X_7:	Hain	Hainaut
X_8:	Lièg	Liège
X_9:	Limb	Limburg
X_{10}:	Luxe	Luxembourg

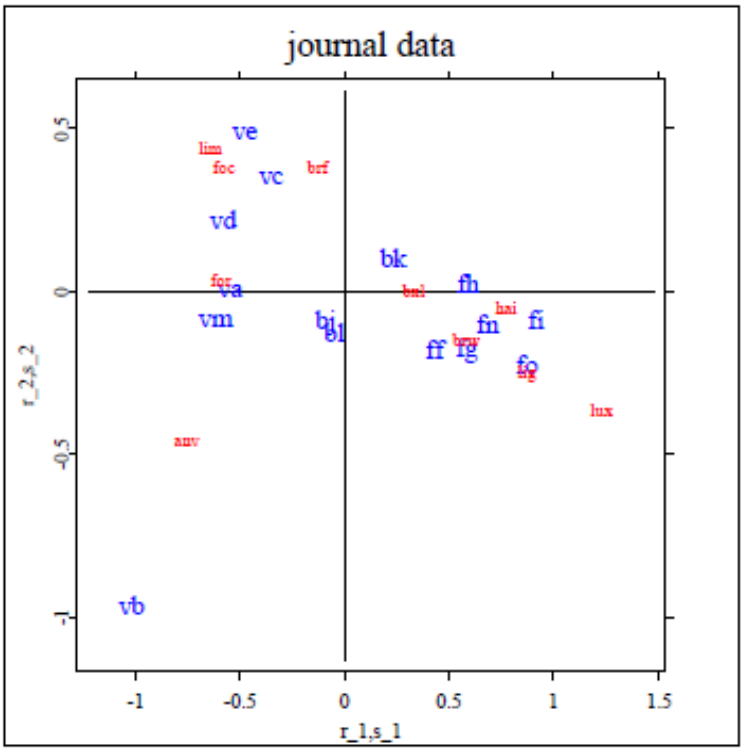
	WaBr	Brar	Antw	FlBr	OcFl	OrFl	Hain	Lièg	Limb	Luxe
v_a	1.8	7.8	9.1	3.0	4.3	3.9	0.1	0.3	3.3	0.0
v_b	0.1	3.4	17.8	1.0	0.7	4.1	0.0	0.0	0.2	0.0
v_c	0.1	9.4	4.6	7.7	4.4	5.8	1.6	0.1	1.4	0.0
v_d	0.5	15.6	6.1	12.0	10.5	10.2	0.7	0.3	5.4	0.0
v_e	0.1	5.2	3.3	4.8	1.6	1.4	0.1	0.0	3.5	0.0
f_f	5.6	13.7	3.1	2.4	0.5	1.7	1.9	2.3	0.2	0.2
f_g	4.1	16.5	1.9	1.0	1.0	0.9	2.4	3.2	0.1	0.3
f_h	8.3	29.5	1.8	7.3	0.8	0.4	5.1	3.2	0.2	0.3
f_i	0.9	7.8	0.2	2.6	0.1	0.1	5.6	3.8	0.1	0.8
b_j	6.1	18.2	10.8	4.1	4.5	5.3	2.0	2.6	3.4	0.2
b_k	8.3	35.4	6.2	11.0	5.0	6.1	5.5	3.3	1.5	0.3
b_l	4.4	9.9	6.7	3.4	1.1	3.9	2.1	1.5	2.1	0.0
v_m	0.3	11.6	14.2	4.7	5.1	7.9	0.3	0.5	3.0	0.0
f_n	5.1	21.0	1.3	3.4	0.2	0.2	2.3	4.4	0.0	0.4
f_0	2.2	9.8	0.1	0.3	0.0	0.7	2.3	3.0	0.3	1.0

Πίνακας 3.1 Δεδομένα για την ανάγνωση εφημερίδων.

Οι ιδιοτιμές της παραγοντικής correspondence ανάλυσης δίνονται στον Πίνακα 3.2. Δισδιάστατες απεικονίσεις θα είναι αρκετά ικανοποιητικές, καθώς οι δύο πρώτες ιδιοτιμές υπολογίζονται στο 81% της inertia. Το Γράφημα 3.1 δείχνει τις προβολές των γραμμών (οι 15 εφημερίδες) και των στηλών (10 περιφέρειες).

λ_j	percentage of variance	cumulated percentage
183.40	0.653	0.653
43.75	0.156	0.809
25.21	0.090	0.898
11.74	0.042	0.940
8.04	0.029	0.969
4.68	0.017	0.985
2.13	0.008	0.993
1.20	0.004	0.997
0.82	0.003	1.000
0.00	0.000	1.000

Πίνακας 3.2 Ιδιοτιμές και αποτελέσματα Correspondence Analysis των δεδομένων.



Γράφημα 3.1 Γραφική παράσταση των κατηγοριών των εφημερίδων στις δυο πρώτες διαστάσεις.

	$C_a(i,r_1)$	$C_a(i,r_2)$	$C_a(i,r_3)$
v_a	0.0563	0.0008	0.0036
v_b	0.1555	0.5567	0.0067
v_c	0.0244	0.1179	0.0266
v_d	0.1352	0.0952	0.0164
v_e	0.0253	0.1193	0.0013
f_f	0.0314	0.0183	0.0597
f_g	0.0585	0.0162	0.0122
f_h	0.1086	0.0024	0.0656
f_i	0.1001	0.0024	0.6376
b_j	0.0029	0.0055	0.0187
b_k	0.0236	0.0278	0.0237
b_l	0.0006	0.0090	0.0064
v_m	0.1000	0.0038	0.0047
f_n	0.0966	0.0059	0.0269
f_0	0.0810	0.0188	0.0899
Total	1.0000	1.0000	1.0000

Πίνακας 3.3 Απόλυτη συνεισφορά των παραγόντων της γραμμής r_k.

Όπως ήταν αναμενόμενο, υπάρχει ισχυρή συσχέτιση μεταξύ των περιφερειών και του είδους των εφημερίδων που διαβάζονται. Ειδικότερα, η u_b (Gazet Van Αμβέρσα) διαβάζεται σχεδόν αποκλειστικά στην επαρχία της Αμβέρσας (αυτό είναι ένα ακραίο σημείο στο γράφημα). Τα σημεία στα αριστερά ανήκουν όλα στην Φλάνδρα, ενώ εκείνες στα δεξιά ανήκουν όλα στην Βαλονία. Ας σημειωθεί ότι η Βαλλονική-Brabant και η Φλαμανδική-Brabant δεν είναι μακριά από τις Βρυξέλλες. Οι Βρυξέλλες είναι κοντά στο κέντρο (μέσο όρο) και επίσης κοντά στις δίγλωσσες εφημερίδες. Μετατοπίζονται λίγο προς τα δεξιά του σημείου αναφοράς λόγω της πλειοψηφίας των Γαλλόφωνων στην περιοχή.

Η απόλυτες συνεισφορές των 3 πρώτων παραγόντων της σειράς r_k αναφέρονται στον Πίνακα 3.3. Το Γράφημα 3.1 παριστάνει γραφικά την προβολή σε δύο διαστάσεις των γραμμών (15 εφημερίδες) και στηλών (10 περιοχές).

Δείχνουν, για παράδειγμα, το σημαντικό ρόλο της Αμβέρσας και της εφημερίδας u_b για τον προσδιορισμό της διασποράς και των δύο παραγόντων. Σαφώς, ο πρώτος άξονας εκφράζει γλωσσικές διαφορές μεταξύ των 3 τμημάτων του Βελγίου. Ο δεύτερος άξονας δείχνει μια μεγαλύτερη διασπορά μεταξύ της περιφέρειας της Φλάνδρας από τις γαλλόφωνες περιφέρειες. Σημειώστε επίσης ότι ο 3-ος άξονας παρουσιάζει σημαντικό ρόλο της κατηγορίας "fi" (άλλες γαλλικές εφημερίδες) με την

Wallon-Brabant «brw» και το Hainant «hai», που δείχνει τις πιο σημαντικές συνεισφορές.

Οι συντεταγμένες της «fi» για τον άξονα αυτό είναι αρνητικές (δεν φαίνεται εδώ), και το ίδιο ισχύει για τις συντεταγμένες των "brw" και "hai". . Προφανώς, αυτές τις δύο περιοχές, επίσης, φαίνεται να διαθέτουν μεγαλύτερο ποσοστό από τους αναγνώστες των περισσότερων τοπικών εφημερίδων.

3.4 Πρακτικοί υπολογισμοί-Ορισμός inertia

Στην πράξη η εκτέλεση των τεχνικών που παρουσιάστηκαν ξεκινά με τον υπολογισμό των ιδιοτιμών $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ και των αντίστοιχων ιδιοδιανυσμάτων $u_1, ..., u_p$ του $X^T X$. Επειδή συνήθως το p είναι μικρότερο από το n, η διαδικασία αυτή απαιτεί λιγότερους αριθμητικούς υπολογισμούς από το να υπολογισθούν πρώτα τα u_k για $k = 1, ..., p$.

Η απεικόνιση των n παρατηρήσεων λαμβάνεται σε ένα διάγραμμα (z_1, z_2) όπου $z_1 = Xu_1$ και $z_2 = Xu_2$. Αν είναι απαραίτητο χρησιμοποιείται και τρίτη διάσταση με $z_3 = Xu_3$.

Χρησιμοποιώντας τη δυαδική σχέση

$$w_k = X^T v_k = \frac{1}{\sqrt{\lambda_k}} X^T X u_k = \sqrt{\lambda_k} u_k \qquad (3.4.1)$$

μπορούν εύκολα να ληφθούν οι απεικονίσεις των p μεταβλητών. Αυτές οι απεικονίσεις μπορούν να οπτικοποιηθούν σε ένα scatter plot (w_1, w_2) με $w_1 = \sqrt{\lambda_1} u_1$ και $w_2 = \sqrt{\lambda_2} u_2$ (πιθανόν και σε τρίτη διάσταση με $w_3 = \sqrt{\lambda_3} u_3$).

Μπορεί βεβαίως να γίνει παραγοντική ανάλυση σε περισσότερες διαστάσεις (υπολογίζοντας τα z_k και w_k για $k > 3$) αλλά δεν μπορούν φυσικά να απεικονισθούν.

Ένας τυποποιημένος τρόπος υπολογισμού της ποσότητας που απεικονίζεται σε έναν υποχώρο q ($q < p$) δίνεται από το λόγο

$$\tau_q = \frac{\lambda_1 + \lambda_2 + ... + \lambda_q}{\lambda_1 + \lambda_2 + ... + \lambda_p} \qquad (3.4.2) \text{ όπου } 0 \leq \tau_q \leq 1.$$

Γενικά το εσωτερικό γινόμενο $y^T y$ καλείται **inertia** του $y \in \mathbb{R}^n$.

Γι' αυτό το λόγο το μέγεθος τ_p συνήθως ερμηνεύεται ως το επί τοις εκατό ποσοστό της συνολικής inertia που ερμηνεύεται από τους πρώτους q παράγοντες.

Δεδομένου ότι $\lambda_j = \left(Xu_j\right)^T \left(Xu_j\right) = z_j^T z_j$ το λ_j είναι η inertia της j-ιοστής μεταβλητής,

Ο παρανομαστής της σχέσης (3.4.1) δίνει τη συνολική inertia των p μεταβλητών.

Πράγματι: $\sum_{j=1}^{p} \lambda_j = tr\left(X^T X\right) = \sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2 = \sum_{j=1}^{p} x_j^T x_j$.

Είναι σαφές ότι το $\sum_{l=1}^{q} \lambda_j$ είναι η συνολική inertia των q πρώτων παραγοντικών μεταβλητών $z_1, z_2, ..., z_q$.

Στην πράξη, ένας άλλος τρόπος για να κοιτάξουμε την CA είναι να θεωρήσουμε μια μέθοδο αποσύνθεσης του συνολικού X^2 στατιστικού αφού η inertia και το X^2 συνδέονται με τη σχέση $\text{inertia} = \frac{\text{chi-square}}{\text{total} \quad N}$ προσδιορίζοντας έτσι ένα μικρότερο αριθμό διαστάσεων στις οποίες να εκπροσωπούνται οι αποκλίσεις από τις αναμενόμενες τιμές και να αιτιολογούν την απόρριψη ή όχι της στατιστικής υπόθεσης ότι υπάρχει συσχέτιση μεταξύ των μεταβολών γραμμών και στηλών.

3.5 Biplots

Το biplot είναι (Hardle & Simar, 2003 σελ 356-357) μια μικρών διαστάσεων απεικόνιση ενός πίνακα δεδομένων X, όπου οι γραμμές και οι στήλες εκπροσωπούνται από σημεία. Ας υποθέσουμε ότι έχουμε ένα (10 × 5) πίνακα δεδομένων με στοιχεία x_{ij}. Η ιδέα της biplot είναι να βρει 10 σημεία γραμμής $q_i \in R^k \ \left(k < p, i = 1,...,10\right)$ και 5 σημεία στήλης $t_j \in R^k \ \left(j = 1,...,5\right)$ έτσι ώστε τα 50 κλιμακωτά εξαγόμενα μεταξύ των γραμμών και των στηλών να προσεγγίζουν τα 50 αντίστοιχα στοιχεία του πίνακα δεδομένων X. Συνήθως επιλέγουμε $k = 2$. Για παράδειγμα, το κλιμακωτά εξαγόμενα μεταξύ q_7 και t_4 θα πρέπει να προσεγγίσει την τιμή των δεδομένων x_{74} στην έβδομη σειρά και την τέταρτη στήλη. Σε γενικές γραμμές, τα biplot μοντελοποιούν τα δεδομένα x_{ij} όπως το άθροισμα ενός κλιμακωτού εξαγόμενου σε ορισμένους μικρών διαστάσεων υποχώρους και έναν όρο κατάλοιπο: $x_{ij} = q_i^T t_j + e_{ij} = \sum_k q_{ik} t_{jk} + e_{ij}$.

Για να κατανοήσουμε τη σχέση μεταξύ της correspondence ανάλυσης και του biplot, πρέπει να εισαχθεί μια φόρμουλα η οποία να εκφράζει τα x_{ij} από τα αρχικά δεδομένα του πίνακα σε σχέση με τις συχνότητες των γραμμών και των στηλών. Ένας τέτοιος τύπος, που είναι γνωστός ως το "τύπος ανασύστασης", είναι :

$$x_{ij} = E_{ij}\left(1 + \frac{\sum_{k=1}^{R} \lambda_k^{\frac{1}{2}} \gamma_{ik} \delta_{jk}}{\sqrt{\frac{x_{i\bullet} x_{j\bullet}}{x_{\bullet\bullet}}}}\right)$$

$$\left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{i\bullet}}{x_{\bullet\bullet}}\right) = \sum_{k=1}^{R} \lambda_k^{\frac{1}{2}} \gamma_{ik} \left(\sqrt{\frac{x_{\bullet j}}{x_{i\bullet} x_{\bullet\bullet}}}\right) \delta_{jk}$$

$$\left(\frac{x_{ij}}{x_{\bullet j}} - \frac{x_{\bullet j}}{x_{\bullet\bullet}}\right) = \sum_{k=1}^{R} \lambda_k^{\frac{1}{2}} \gamma_{ik} \left(\sqrt{\frac{x_{i\bullet}}{x_{\bullet j} x_{\bullet\bullet}}}\right) \delta_{jk}.$$

Τώρα, αν $\lambda_1 >> \lambda_2 >> \lambda_3 \ldots$, μπορούμε να προσεγγίσουμε αυτά τα ποσά από έναν πεπερασμένο αριθμό K (συνήθως $K = 2$) και προκύπτει:

$$\left(\frac{x_{ij}}{x_{\bullet j}} - \frac{x_{i\bullet}}{x_{\bullet\bullet}}\right) = \sum_{k=1}^{K} \left(\frac{x_{i\bullet}}{\sqrt{\lambda_k x_{\bullet\bullet}}} r_{ki}\right) s_{kj} + e_{ij},$$

$$\left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{\bullet j}}{x_{\bullet\bullet}}\right) = \sum_{k=1}^{K} \left(\frac{x_{\bullet j}}{\sqrt{\lambda_k x_{\bullet\bullet}}} s_{kj}\right) r_{ki} + e'_{ij},$$

όπου e_{ij} και e'_{ij} οι όροι των σφαλμάτων. Που δείχνουν ότι αν αναλογιστούμε την εμφάνιση των διαφορών μεταξύ των προφίλ των γραμμών και το μέσο προφίλ, η προβολή του προφίλ σειράς r_k και η αναπροσαρμοσμένη σε κλίμακα έκδοση των προβολών του προφίλ στήλη s_k αποτελεί biplot αυτών των διαφορών. Συνεπάγεται το ίδιο και για τις διαφορές μεταξύ των προφίλ στήλης και μέσο προφίλ των στηλών.

Παράδειγμα εφαρμογής της Correspondense Analysis.

Παρουσιάζοντας τα στοιχεία για τις συνήθειες καπνίσματος διαφορετικών υπαλλήλων σε μία εταιρεία (τα στοιχεία παρουσιάσθηκαν στο Greenacre 1984, p.55) έχουμε τον παρακάτω Πίνακα 3.4.:

Correspondence Table

	SMOKING CATEGORY				
STAFF GROUP	(1) None	(2) Light	(3) Medium	(4) Heavy	Active Margin
(1) Senior Managers	4	2	3	2	11
(2) Junior Managers	4	3	7	4	18
(3) Senior Employees	25	10	12	4	51
(4) Junior Employees	18	24	33	13	88
(5) Secretaries	10	6	7	2	25
Active Margin	61	45	62	25	193

Πίνακας 3.4 Πίνακας συνάφειας συνηθειών καπνίσματος.

Αν θεωρήσουμε τις τιμές των 4 στηλών σε κάθε σειρά του πίνακα σαν συντεταγμένες σε ένα τεσσάρων διαστάσεων χώρο ($\mathbb{R}^4$) θα υπολογίζαμε τις Ευκλείδειες αποστάσεις μεταξύ των σημείων των 5 γραμμών στον $\mathbb{R}^4$. Οι αποστάσεις μεταξύ των σημείων στον $\mathbb{R}^4$ συνοψίζουν όλες τις πληροφορίες σχετικά με τις ομοιότητες μεταξύ των γραμμών στον παραπάνω πίνακα.

Σκοπός είναι να βρούμε ένα χώρο μικρότερων διαστάσεων, στον οποίο να διατηρούνται όλες ή σχεδόν όλες οι διαφορές μεταξύ των γραμμών.

Θα μπορούσαμε τότε να παρουσιάσουμε όλες τις πληροφορίες σχετικά με τις ομοιότητες μεταξύ των γραμμών (κατηγορία υπαλλήλων στην προκειμένη περίπτωση) σε ένα απλό 1,2 ή 3 διαστάσεων χώρο.

Και ενώ ο τρόπος αυτός μπορεί να μην φαίνεται χρήσιμος για ένα μικρό πίνακα όπως ο παραπάνω, μπορούμε να καταλάβουμε πως η παρουσίαση και η ερμηνεία πολύ μεγάλων πινάκων θα είχαν μεγάλα οφέλη από την απλοποίηση η οποία μπορεί να επιτευχθεί μέσω της correspondence analysis.

Ο επόμενος Πίνακας 3.5 εμφανίζει τα ποσοστά (παρατηρούμενες συχνότητες) ανά γραμμή.

Row Profiles

STAFF GROUP	SMOKING CATEGORY (1) None	(2) Light	(3) Medium	(4) Heavy	Active Margin
(1) Senior Managers	,364	,182	,273	,182	1,000
(2) Junior Managers	,222	,167	,389	,222	1,000
(3) Senior Employees	,490	,196	,235	,078	1,000
(4) Junior Employees	,205	,273	,375	,148	1,000
(5) Secretaries	,400	,240	,280	,080	1,000
Mass	,316	,233	,321	,130	

Πίνακας 3.5 Παρατηρούμενες συχνότητες καπνίσματος ανά κατηγορία υπαλλήλων.

Δηλαδή για την κατηγορία (2) Junior Managers έχουμε:

$0.222 = \frac{4}{18}, 0.167 = \frac{3}{18}, 0.389 = \frac{7}{18}, 0.222 = \frac{4}{18}$ και φυσικά όλες οι γραμμές αθροίζουν στο 1.

Δηλαδή για παράδειγμα το 16.7% των Junior Managers καπνίζουν λίγο (Light).

Η τελευταία γραμμή (mass) εμφανίζει το συνολικό ποσοστό της κάθε στήλης ως προς συνολικό μέγεθος. ($0.316 = \frac{61}{193}, 0.233 = \frac{45}{193}, 0.321 = \frac{62}{193}, 0{,}130 = \frac{25}{193}$).

Δηλαδή για παράδειγμα 32.1% στο σύνολο των 193 είναι καπνιστές με μέτρια κατανάλωση καπνού.

Αντίστοιχα στον επόμενο Πίνακα 3.6 έχουμε τα ποσοστά ανά στήλη και τα συνολικά (mass) ποσοστά ανά γραμμή.

Column Profiles

STAFF GROUP	SMOKING CATEGORY (1) None	(2) Light	(3) Medium	(4) Heavy	Mass
(1) Senior Managers	,066	,044	,048	,080	,057
(2) Junior Managers	,066	,067	,113	,160	,093
(3) Senior Employees	,410	,222	,194	,160	,264
(4) Junior Employees	,295	,533	,532	,520	,456
(5) Secretaries	,164	,133	,113	,080	,130
Active Margin	1,000	1,000	1,000	1,000	

Πίνακας 3.6 Παρατηρούμενες συχνότητες ανά κατηγορία καπνίσματος και συνολικές ανά γραμμή.

Ο επόμενος Πίνακας 3.7 εμφανίζει τα αποτελέσματα της. Correspondese analysis σε διάστημα εμπιστοσύνης $\alpha = 0.05$.

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia: Accounted for	Proportion of Inertia: Cumulative	Confidence Singular Value: Standard Deviation	Confidence Singular Value: Correlation 2
1	,273	,075			,878	,878	,070	,020
2	,100	,010			,118	,995	,076	
3	,020	,000			,005	1,000		
Total		,085	16,442	,172[a]	1,000	1,000		

a. 12 degrees of freedom

Πίνακας 3.7 Αποτελέσματα Correspondence Analysis.

Η στήλη Dimension αναφέρεται στο R, που εδώ είναι 3 μιας και $R = \min\{i-1, j-1\}$ όπου με $i = 5$ και $j = 4$ είναι $R = \min\{4,3\} = 3$.

Η στήλη chi-square εμφανίζει το συνολικό X^2 κάτω από τον έλεγχο H_0 της ανεξαρτησίας δηλ. $X^2 = 16.442$.

Το σημείο σύγκρισης, εάν $\alpha = 0.05$ θα ήταν το $X^2_{0.05,12}$ όπου 12 οι βαθμοί ελευθερίας αφού γενικά είναι $(i-1)\cdot(j-1)$ δηλ είναι $(5-1)\cdot(4-1) = 12$. Από πίνακες βλέπουμε ότι $X^2_{0.05,12} = 21.03$.

Είναι $X^2 = 16.442 < 21.03$ δηλ δεν ανήκει στην κρίσιμη περιοχή άρα δεν απορρίπτεται η μηδενική υπόθεση .

Αυτό επιβεβαιώνεται και από p-value=0.173>0.05.

Στη στήλη proportion of inertia παρατηρούμε ότι το μεγαλύρερο μέρος της inertia (87.8%) γίνεται από τον πρώτο άξονα $(R = 1)$.Με τον δεύτερο άξονα $(R = 2)$ θα αποκτηθεί το 99.5%

Επειδή οι δύο πρώτες διαστάσεις εξηγούν το 99.5% της συνολικής inertia, η δισδιάστατη απεικόνιση θα είναι αρκετή για την ανάλυση των σχέσεων μεταξύ του προσωπικού της εταιρείας και συνήθειες καπνίσματός τους.

Οι επόμενοι δύο Πίνακες 3.7.1 και 3.7.2 εμφανίζουν τα προηγούμενα (Πίνακας 3.7) συνολικά αποτελέσματα αναλυτικά ανά γραμμή και στήλη.

Overview Row Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
STAFF GROUP	Mass	1	2	Inertia	1	2	1	2	Total
(1) Senior Managers	,057	-,126	,612	,003	,003	,214	,092	,800	,893
(2) Junior Managers	,093	,495	,769	,012	,084	,551	,526	,465	,991
(3) Senior Employees	,264	-,728	,034	,038	,512	,003	,999	,001	1,000
(4) Junior Employees	,456	,446	-,183	,026	,331	,152	,942	,058	1,000
(5) Secretaries	,130	-,385	-,249	,006	,070	,081	,865	,133	,999
Active Total	1,000			,085	1,000	1,000			

a. Symmetrical normalization

Πίνακας 3.7.1 Αποτελέσματα ανά κατηγορία υπαλλήλων.

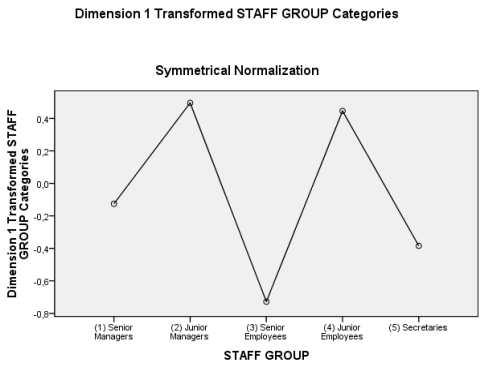
Overview Column Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
SMOKING CATEGORY	Mass	1	2	Inertia	1	2	1	2	Total
(1) None	,316	-,752	,096	,049	,654	,029	,994	,006	1,000
(2) Light	,233	,190	-,446	,007	,031	,463	,327	,657	,984
(3) Medium	,321	,375	-,023	,013	,166	,002	,982	,001	,983
(4) Heavy	,130	,562	,625	,016	,150	,506	,684	,310	,995
Active Total	1,000			,085	1,000	1,000			

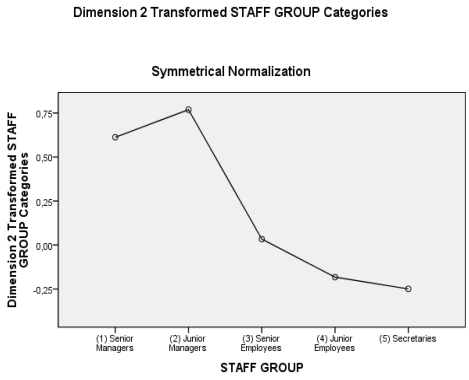
a. Symmetrical normalization

Πίνακας 3.7.2 Αποτελέσματα ανά κατηγορία καπνίσματος.

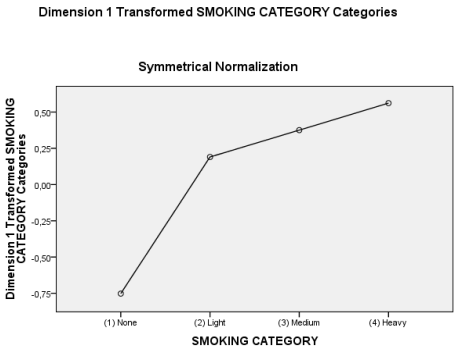
Στα επόμενα 6 (3.1, 3.2, 3.3, 3.4, 3.5, 3.6) διαγράμματα εμφανίζονται οι προηγούμενες πληροφορίες χρησιμοποιώντας τις δύο διαστάσεις όπως εξηγήθηκε παραπάνω.



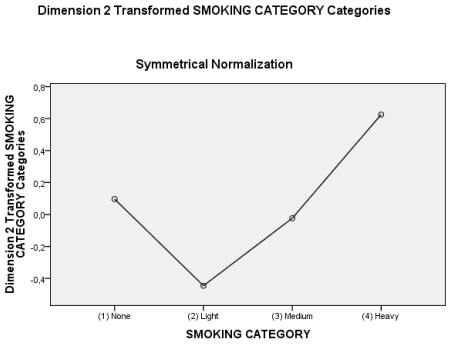
Διάγραμμα 3.1 Διαγραμματική απεικόνιση Πίνακα 3.7.1 σε D1



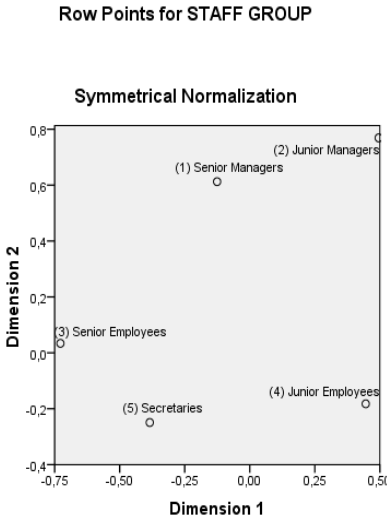
Διάγραμμα 3.2 Διαγραμματική απεικόνιση Πίνακα 3.7.1 σε D2



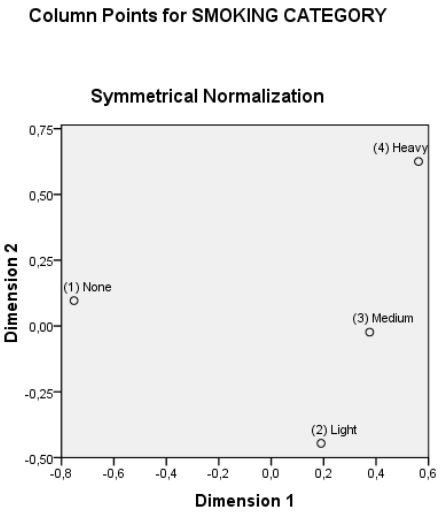
Διάγραμμα 3.3 Διαγραμματική απεικόνιση Πίνακα 3.7.2 σε D1



Διάγραμμα 3.4 Διαγραμματική απεικόνιση Πίνακα 3.7.2 σε D2



Διάγραμμα 3.5 Διαγραμματική απεικόνιση Κατηγορίας Υπαλλήλων σε D1 D2



Διάγραμμα 3.6 Διαγραμματική απεικόνιση Κατηγορίας καπνίσματος σε D1 D2

ΚΕΦΑΛΑΙΟ 4

Εφαρμογή της διαχωριστικής ανάλυσης και της correspondence analysis σε δεδομένα που προκύπτουν από την εκπαιδευση

Όπως παρουσιάστηκε συνοπτικά στη εισαγωγή σκοπός της παρούσας είναι να μελετήσει τα εκπαιδευτικά χαρακτηριστικά των μαθητών σε σχεσή με την καταγωγή τους. Τα δεδομένα προέρχονται από σχολική κοινότητα της οποίας οι μαθητές χωρίζονται σε τρεις κατηγορίες,

- Οι ελληνόπαιδες, δηλαδή παιδιά που έχουν γεννηθεί στην Ελλάδα από έλληνες γονείς
- Οι παλινοστούντες δηλαδή παιδιά που έχουν γεννηθεί στην Ελλάδα από γονείς έλληνικής καταγωγής επαναπατρισθέντες κυρίως από χώρες της πρών Σοβιετικής Ένωσης, και
- Οι αλλοδαπόι παιδιά από μη έλληνες γονείς

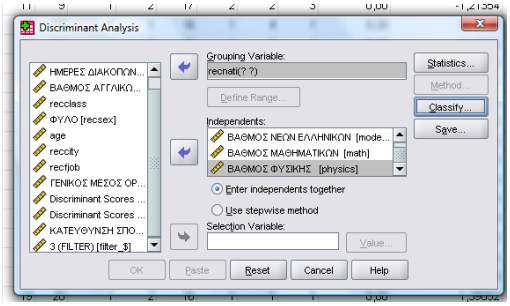
Προκειμένου να διαμορφωθεί μια άποψη σχετικά με την διαφορετικότητα που πιθανόν υφίσταται μεταξύ των κατηγοριών αυτών στο χώρο της εκπάιδευσης μελετώνται δυο προβλήματα

- Πόσο διαφέρουν οι επιδόσεις των παιδιών ανά μάθημα και κατηγορία;
- Τi επιλογές κάνουν σε σχέση με την κατεύθυνση των σπουδών τους και κατ επέκταση το επάγγελμά τους;

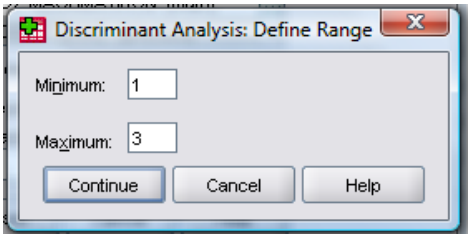
Για το πρώτο ερώτημα μελετώνται οι βαθμολογίες των μαθητών στα τέσσερα βασικά μαθήματα δηλαδή Αρχαία Ελληνικά, Νέα Ελληνικά , Μαθηματικά και Φυσική. Δεν θεωρήθηκε σκόπιμο να μελετηθούν όλες οι βαθμολογίες σε όλα τα μαθήματα καθώς τα τέσσερα παραπάνω ορίζουν την κλασική εκπαίδευση. Οι τρεις κατηγορίες είναι οι τρεις ομάδες στις οποίες θέλουμε να δούμε αν μπορεί να γίνει ο διαχωρισμός σε σχέση με τις βαθμολογίες των μαθημάτων. Στη συγκεκριμένη περίπτωση έχοντας μια ποιοτική μεταβλητή, την καταγωγή, και τέσσερις ποσοτικές μεταβλητές, τις βαθμολογίες στα τέσσερα βασικά μαθήματα το κατάλληλο εργαλείο είναι η διαχωριστική ανάλυση. Περαιτέρω σύγκριση έγινε ανά κατηγορία στους βαθμούς των Μαθηματικών και των discriminant scores.

Το στατιστικό πακέτο που χρησιμοποιήθηκε είναι το SPSS 17.0 με την ακόλουθη διαδικασία εντολών:

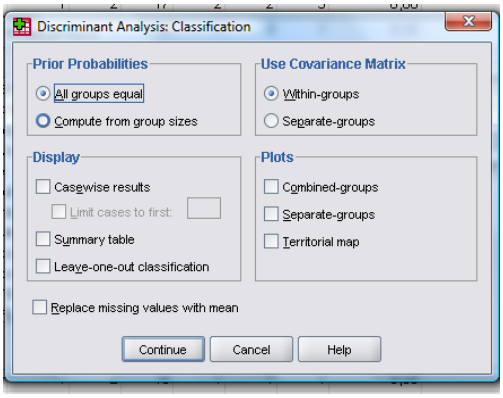
ANALYSE → CLASSIFY → DISCRIMINANT → ……



Εικόνα 4.1 Πρώτο βήμα



Εικόνα 4.2 Δεύτερο βήμα



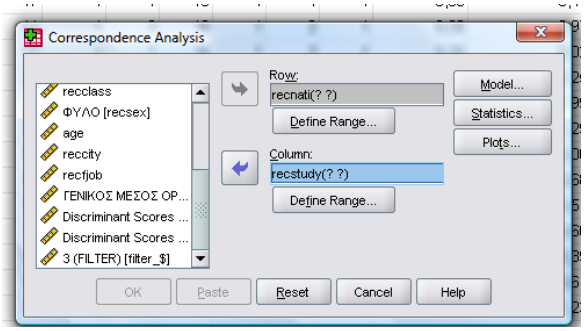
Εικόνα 4.3 Τρίτο βήμα

Οι εντολές statistics classify και save χρησιμοποιήθηκαν ανάλογα με την περίπτωση.

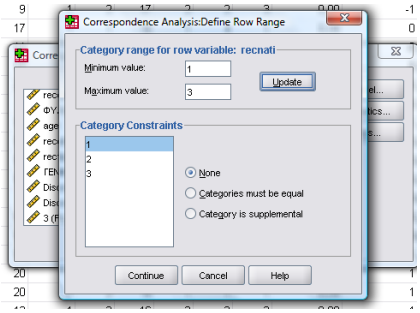
Για το δεύτερο ερώτημα μελετώνται οι επιλογές των κατευθύνσεων ανά κατηγορία. Επειδή και οι δυο είναι ποιοτικές μεταβλητές, για την δημιουργία των πινάκων συνάφειας έγινε χρήση της correspondence analysis της οποίας η ελεγχοσυνάρτηση ακολουθεί την X^2 κατανομή με 4 βαθμούς ελευθερίας $\left[(r-1)(k-1)\right]=(3-1)(3-1)=4$.

Το στατιστικό πακέτο που χρησιμοποιήθηκε είναι το SPSS 17.0 με την ακόλουθη διαδικασία εντολών:

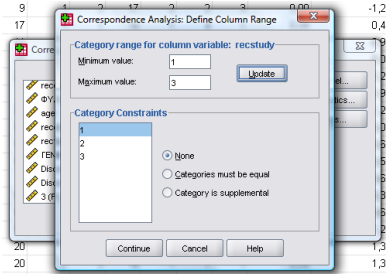
ANALYSE → DIMENSION-REDUCTION → CORRESPONDENCE ANALYSIS → ……



Εικόνα 4.4 Πρώτο βήμα



Εικόνα 4.5 Δεύτερο βήμα



Εικόνα 4.6 Τρίτο βήμα.

Χρησιμοποιήθηκαν και οι εντολές statistics και plots.

Εκτελώντας τις παραπάνω εντολές στα δεδομένα του παραδείγματος προέκυψαν τα output του προγράμματος spss 17.0 για την διαχωριστική ανάλυση και την correspondence anlysis αντίστοιχα και παρατείθεται η ερμηνεία τους.

Πίνακες, διαγράμματα και συμπεράσματα για την διαχωριστική ανάλυση.

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		496	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		496	100,0

Πίνακας 4.1 Πλήθος δείγματος.

Group Statistics

ΚΑΤΗΓΟΡΙΑ		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
ΕΛΛΗΝΟΠΑΙΔΕΣ	ΒΑΘΜΟΣ ΑΡΧΑΙΩΝ ΕΛΛΗΝΙΚΩΝ	15,27	3,388	301	301,000
	ΒΑΘΜΟΣ ΝΕΩΝ ΕΛΛΗΝΙΚΩΝ	14,97	3,050	301	301,000
	ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	14,39	3,405	301	301,000
	ΒΑΘΜΟΣ ΦΥΣΙΚΗΣ	13,75	3,547	301	301,000
ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	ΒΑΘΜΟΣ ΑΡΧΑΙΩΝ ΕΛΛΗΝΙΚΩΝ	14,79	3,203	151	151,000
	ΒΑΘΜΟΣ ΝΕΩΝ ΕΛΛΗΝΙΚΩΝ	13,90	2,775	151	151,000
	ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	12,86	2,941	151	151,000
	ΒΑΘΜΟΣ ΦΥΣΙΚΗΣ	12,40	3,086	151	151,000
ΑΛΛΟΔΑΠΟΙ	ΒΑΘΜΟΣ ΑΡΧΑΙΩΝ ΕΛΛΗΝΙΚΩΝ	14,36	3,564	44	44,000
	ΒΑΘΜΟΣ ΝΕΩΝ ΕΛΛΗΝΙΚΩΝ	13,95	2,965	44	44,000
	ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	12,67	3,146	44	44,000
	ΒΑΘΜΟΣ ΦΥΣΙΚΗΣ	12,61	3,322	44	44,000
Total	ΒΑΘΜΟΣ ΑΡΧΑΙΩΝ ΕΛΛΗΝΙΚΩΝ	15,04	3,356	496	496,000
	ΒΑΘΜΟΣ ΝΕΩΝ ΕΛΛΗΝΙΚΩΝ	14,55	3,001	496	496,000
	ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	13,77	3,331	496	496,000
	ΒΑΘΜΟΣ ΦΥΣΙΚΗΣ	13,24	3,447	496	496,000

Πίνακας 4.2 Μέσοι ανά κατηγορία και μάθημα.

Οι παραπάνω Πίνακες 4.1 και 4.2 παρουσιάζουν τους μέσους όρους βαθμολογίας σε τέσσερα μαθήματα Γενικής Παιδείας ανά κατηγορία και συνολικά, ανεξαρτήτως τάξης.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
ΒΑΘΜΟΣ ΑΡΧΑΙΩΝ ΕΛΛΗΝΙΚΩΝ	,992	2,029	2	493	,133
ΒΑΘΜΟΣ ΝΕΩΝ ΕΛΛΗΝΙΚΩΝ	,970	7,546	2	493	,001
ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	,947	13,867	2	493	,000
ΒΑΘΜΟΣ ΦΥΣΙΚΗΣ	,965	8,816	2	493	,000

Πίνακας 4.3 Πίνακας ελέγχου ισότητας μέσων.

Από τον πίνακα ελέγχου για τη ισότητα (Πίνακας 4.3) των μέσων ανά κατηγορία προκύπτει ότι απορρίπτεται σε διάστημα εμπιστοσύνης α=5% η ισότητα των μέσων για τα μαθήματα των Νέων Ελληνικών των Μαθηματικών και της Φυσικής $(p-value<0.05)$ ενώ δεν απορρίπτεται η ισότητα για τα Αρχαία Ελληνικά $(p-value=0.133>0.05)$.

Box's Test of Equality of Covariance Matrices

Log Determinants

ΚΑΤΗΓΟΡΙΑ	Rank	Log Determinant
ΕΛΛΗΝΟΠΑΙΔΕΣ	4	7,148
ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	4	6,463
ΑΛΛΟΔΑΠΟΙ	4	6,130
Pooled within-groups	4	6,894

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Πίνακας 4.4 Πίνακας βαθμών και λογάριθμοι των οριζουσών των πινάκων διακύμανσης.

Test Results

Box's M		21,751
F	Approx.	1,063
	df1	20
	df2	56765,053
	Sig.	,382

Tests null hypothesis of equal population covariance matrices.

Πίνακας 4.5 Πίνακας ελέγχου ισότητας διακυμάνσεων.

Επειδή p-value=0.382 > 0,05 δεν απορρίπτεται η ισότητα των διακυμάνσεων

Αυτό σημαίνει ότι μπορεί να προχωρήσει η ανάλυση των δεδομένων, με υπόθεση ίσες διακυμάνσεις.

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,069[a]	92,0	92,0	,253
2	,006[a]	8,0	100,0	,077

a. First 2 canonical discriminant functions were used in the analysis.

Πίνακας 4.6 Πίνακας ιδιοτιμών των δύο πρώτων διαχωριστικών συναρτήσεων.

Οι ιδιοτιμές των διαχωριστικών συναρτήσεων δείχνουν το ποσοστό της διακύμανσης (το πηλίκο του αθροίσματος των τετραγώνων μεταξύ των ομάδων προς το άθροισμα των τετραγώνων μέσα στις ομάδες) που εξηγείται. Χρειάστηκαν δύο συναρτήσεις ώστε να εξηγηθεί πλήρως ο διαχωρισμός.

Η στήλη canonical correlation δείχνει τη συσχέτιση μεταξύ των discriminant scores για κάθε συνάρτηση και των επιπέδων της εξαρτημένης μεταβλητής. Μία υψηλή συσχέτιση υποδηλώνει μία συνάρτηση που κάνει καλό διαχωρισμό (discriminant). Εδώ οι παρούσες συσχετίσεις 0,253 και 0,077 δεν είναι υψηλές. Άρα δεν αναμένουμε έναν ισχυρό διαχωρισμό.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,930	35,554	8	,000
2	,994	2,939	3	,401

Πίνακας 4.7 Πίνακας Wilks' Lambda.

Ορισμός του Wilks lambda

Ο Wilks lambda ορίζεται ως η ποσοστό της inertia το οποίο **δεν εξηγείται** από την μεταβλητή η οποία προσδιορίζει τις κλάσεις κατά τη διαδικασία αποδόμησης της διακύμανσης.

Άρα $\text{Wilks lambda} = \frac{\text{intra} - \text{class} \quad \text{inertia}}{\text{total} \quad \text{inertia}}$ (A)

(Η διαφορά του Wilks lambda από το ANOVA's F στατιστικό είναι ότι το F είναι ο λόγος $F = \frac{\text{Explaind} \quad \text{Sum} \quad \text{of} \quad \text{Squares}}{\text{Residual(unexplaind)} \quad \text{Sum} \quad \text{of} \quad \text{Squares}}$).

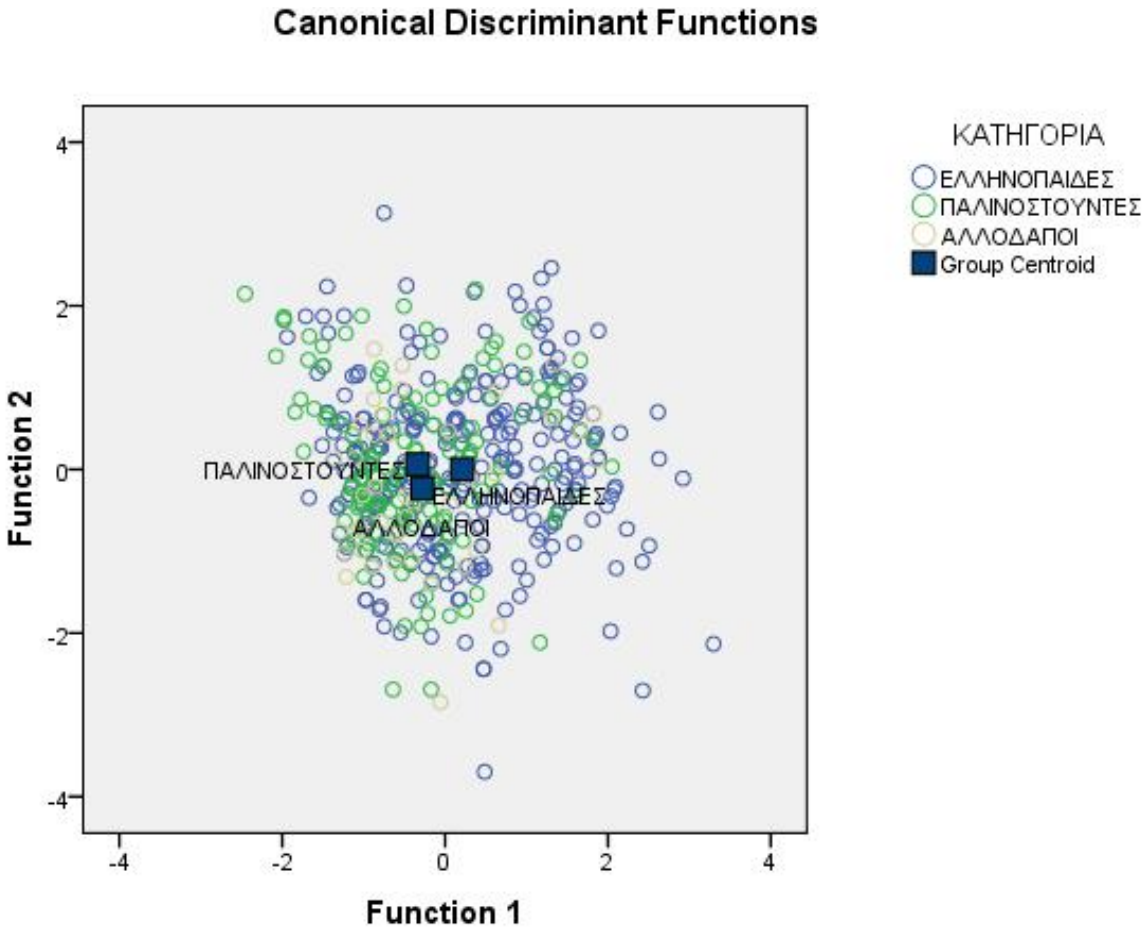
Λόγω της σχέσης (Α) ο Wilks lambda παίρνει τιμές μεταξύ του 0 και 1.

Ως εκ τούτου αν ένα μικρό μόνο μέρος από την συνολική inertia δεν εξηγείται τότε οι ομάδες διαχωρίζονται καλώς και οι μέσοι τους διαφέρουν σημαντικά.

Δηλαδή αν ο Wilks lambda είναι κοντά στο 0 τότε οι ομάδες διαχωρίζονται καλώς ενώ αν είναι κοντά στο 1 δεν υπάρχει σαφής διαχωρισμός.

Στη συγκεκριμένη εφαρμογή η τιμή του Wilks lambda είναι πολύ υψηλή και για τις δύο συναρτήσεις (0.930 και 0,994) άρα δεν υπάρχει σαφής διαχωρισμός των μέσων των κατηγοριών.

Τα παραπάνω φαίνονται στο Διάγραμμα 4.1 που ακολουθεί (CDF)



Διάγραμμα 4.1 Canonical Discriminant Functions

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
ΒΑΘΜΟΣ ΑΡΧΑΙΩΝ ΕΛΛΗΝΙΚΩΝ	-,696	1,209
ΒΑΘΜΟΣ ΝΕΩΝ ΕΛΛΗΝΙΚΩΝ	,665	-,955
ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	,794	1,059
ΒΑΘΜΟΣ ΦΥΣΙΚΗΣ	,085	-,962

Πίνακας 4.8 Συντελεστές συναρτήσεων

Structure Matrix

	Function	
	1	2
ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	,898*	,380
ΒΑΘΜΟΣ ΦΥΣΙΚΗΣ	,722*	-,072
ΒΑΘΜΟΣ ΝΕΩΝ ΕΛΛΗΝΙΚΩΝ	,668*	,072
ΒΑΘΜΟΣ ΑΡΧΑΙΩΝ ΕΛΛΗΝΙΚΩΝ	,314	,494*

Πίνακας 4.9 Συντελεστές συναρτήσεων

Οι παραπάνω Πίνακες 4.8 και 4.9 δίνουν τους συντελεστές των συναρτήσεων

Functions at Group Centroids

	Function	
ΚΑΤΗΓΟΡΙΑ	1	2
ΕΛΛΗΝΟΠΑΙΔΕΣ	,210	,004
ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	-,337	,061
ΑΛΛΟΔΑΠΟΙ	-,278	-,233

Unstandardized canonical discriminant functions evaluated at group means

Πίνακας 4.10 Μέσος των scores

Ο παραπάνω Πίνακας 4.10 δείχνει το μέσο των scores της κάθε συνάρτησης ανά κατηγορία.

Ειδικότερα είναι τα scores για κάθε ομάδα όταν εισάγονται στις συναρτήσεις οι μέσοι των μεταβλητών παρά οι ανεξάρτητες τιμές .

Ο επόμενος Πίνακας 4.11 (CR) περιλαμβάνει τα τελικά αποτελέσματα της ομαδοποίησης.

Classification Results[a]

			Predicted Group Membership			
		ΚΑΤΗΓΟΡΙΑ	ΕΛΛΗΝΟΠΑΙΔΕΣ	ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	ΑΛΛΟΔΑΠΟΙ	Total
Original	Count	ΕΛΛΗΝΟΠΑΙΔΕΣ	146	70	85	301
		ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	40	53	58	151
		ΑΛΛΟΔΑΠΟΙ	10	11	23	44
	%	ΕΛΛΗΝΟΠΑΙΔΕΣ	48,5	23,3	28,2	100,0
		ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	26,5	35,1	38,4	100,0
		ΑΛΛΟΔΑΠΟΙ	22,7	25,0	52,3	100,0

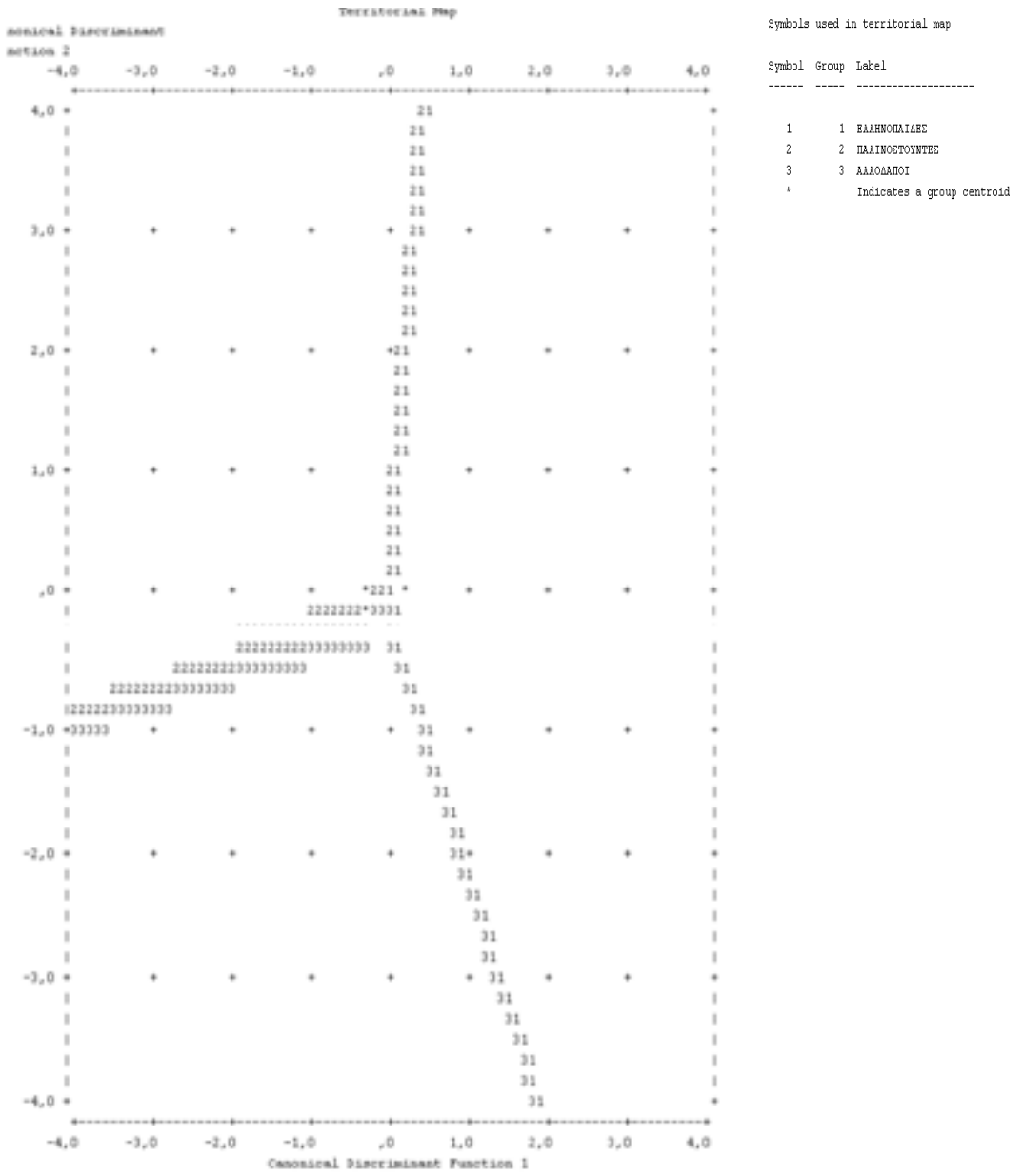
a. 44,8% of original grouped cases correctly classified.

Πίνακας 4.11 Πίνακας αποτελεσμάτων ομαδοποίησης.

Ο τελικός πίνακας συνοψίζει τα προηγούμενες αναλύσεις δηλ πλήθος και ποσοστό παρατηρήσεων που ταξινομήθηκαν σωστά και όχι.

Δηλαδή το 44,8% των παρατηρήσεων μόνο ομαδοποιείται σωστά. Το ποσοστό αυτό είναι χαμηλό όπως παρατηρήθηκε και με τις συσχετίσεις των discriminate scores στη

σελ. 61, ως εκ τούτου δεν μπορούμε να ισχυρισθούμε ότι έχουμε έναν ισχυρό διαχωρισμό.



Διάγραμμα 4.2 Διαγραμματική απεικόνιση των scores DF_1 DF_2.

Στο προηγούμενο Διάγραμμα 4.2 (territorial map) έχουμε την διαγραμματική απεικόνιση των scores των DF_1 και DF_2 όπου στο δεξί τμήμα εμφανίζονται οι υψηλές τιμές των συναρτήσεων δηλαδή το ποσοστό των σωστά ομαδοποιημένων παρατηρήσεων ενώ στα άλλα δύο τμήματα τα χαμηλά scores των συναρτήσεων.

Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	ΕΛΛΗΝΟΠΑΙΔΕΣ	301	14,39	3,405	,196	14,00	14,77	9	20
	ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	151	12,86	2,941	,239	12,39	13,33	9	20
	ΑΛΛΟΔΑΠΟΙ	44	12,67	3,146	,474	11,71	13,63	9	20
	Total	496	13,77	3,331	,150	13,48	14,06	9	20
Discriminant Scores from Function 1 for Analysis 1	ΕΛΛΗΝΟΠΑΙΔΕΣ	301	,2098200	1,06344681	,06129605	,0891953	,3304447	-1,93971	3,29691
	ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	151	-,3372160	,89887614	,07314950	-,4817525	-,1926795	-2,45661	2,05045
	ΑΛΛΟΔΑΠΟΙ	44	-,2780957	,86974360	,13111878	-,5425219	-,0136695	-1,35945	1,86773
	Total	496	,0000000	1,03164584	,04632227	-,0910125	,0910125	-2,45661	3,29691
Discriminant Scores from Function 2 for Analysis 1	ΕΛΛΗΝΟΠΑΙΔΕΣ	301	,0036871	1,00849534	,05812871	-,1107045	,1180788	-3,69683	3,13615
	ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	151	,0606579	1,00583617	,08185379	-,1010774	,2223932	-2,69119	2,20428
	ΑΛΛΟΔΑΠΟΙ	44	-,2333902	,91658090	,13817977	-,5120563	,0452759	-2,84690	1,47058
	Total	496	,0000000	1,00096560	,04494468	-,0883059	,0883059	-3,69683	3,13615

Πίνακας 4.12 Περιγραφικά στοιχεία Βαθμού Μαθηματικών και Discriminant scores σε DF_1 και DF_2

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	7,347	2	493	,001
Discriminant Scores from Function 1 for Analysis 1	7,404	2	493	,001
Discriminant Scores from Function 2 for Analysis 1	,168	2	493	,845

Πίνακας 4.13 Test ομοιογένειας διακυμάνσεων

ANOVA

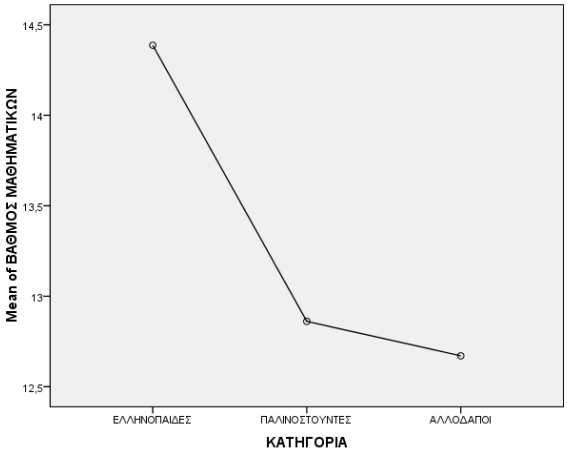
		Sum of Squares	df	Mean Square	F	Sig.
ΒΑΘΜΟΣ ΜΑΘΗΜΑΤΙΚΩΝ	Between Groups	292,588	2	146,294	13,867	,000
	Within Groups	5201,211	493	10,550		
	Total	5493,798	495			
Discriminant Scores from Function 1 for Analysis 1	Between Groups	33,825	2	16,913	16,913	,000
	Within Groups	493,000	493	1,000		
	Total	526,825	495			
Discriminant Scores from Function 2 for Analysis 1	Between Groups	2,956	2	1,478	1,478	,229
	Within Groups	493,000	493	1,000		
	Total	495,956	495			

Πίνακας 4.14 Πίνακας ANOVA για Βαθμό Μαθηματικών

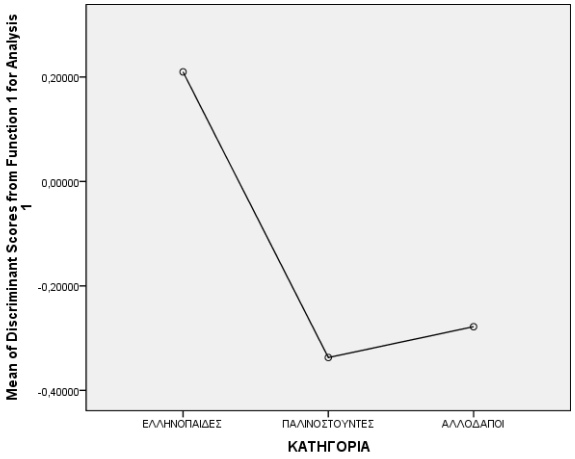
Η ανάλυση ANOVA με Factor την κατηγορία και dependent list τον βαθμό στα Μαθηματικά και τα discriminant scores των DF_1 και DF_2 όπου DF_1 και DF_2 οι δύο discriminant συναρτήσεις που δημιουργήθηκαν προηγούμενα μας έδωσε τα αποτελέσματα των τριών προηγούμενων Πινάκων (4.12, 4.13, 4.14) όπου :

1. Απορρίπτεται η ομοιογένεια των διακυμάνσεων στα Μαθηματικά και στα DF_1 ενώ δεν απορρίπτεται η ομοιογένεια στη διακύμανση των DF_2.
2. Απορρίπτεται η ισότητα των μέσων μεταξύ των ομάδων στα Μαθηματικά και στα DF_1 ενώ δεν απορρίπτεται η ισότητα των μέσων για τα DF_2.

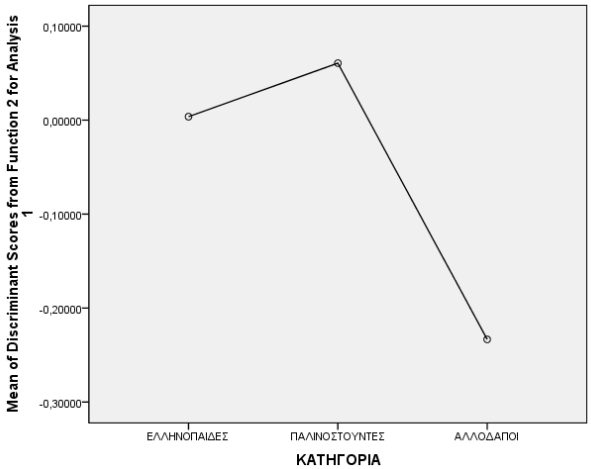
Τα αποτελέσματα που προέκυψαν από την ANOVA εμφανίζονται διαγραμματικά παρακάτω.



Διάγραμμα 4.3 Μέσοι ανά κατηγορία



Διάγραμμα 4.4 Μέσοι DF_1 ανά κατηγορία



Διάγραμμα 4.5 Μέσοι DF_2 ανά κατηγορία

Συμπεράσματα της διαχωριστικής ανάλυσης.

Σύμφωνα με την προηγούμενη ανάλυση φαίνεται ότι δεν μπορεί να γίνει σαφής διαχωρισμός μεταξύ των ομάδων εφόσον μόνο το 44.8% των παρατηρήσεων ομαδοποιείτε σωστά.

Αυτό μπορεί να μας οδηγήσει στο συμπέρασμα ότι και οι μη Έλληνες μαθητές ακολουθώντας τη Ελληνική δημόσια εκπαίδευση ενσωματώνονται σε μεγάλο βαθμό στην ελληνική κοινωνία και κουλτούρα.

Αυτό όμως εξαρτάται από το ποσοστό των μη ελλήνων σε σχέση με τους έλληνες μαθητές.

Φάνηκε να υπάρχει στατιστικά σημαντική διαφορά για την ισότητα των μέσων στα Νέα Ελληνικά, κάτι το οποίο αναμενόταν λόγω μητρικής γλώσσας των ελληνοπαίδων.

Στα μαθηματικά και στη Φυσική, παρότι οι έννοιες είναι διεθνείς και θα αναμενόταν ισότητα των μέσων εντούτοις δεν παρουσιάζεται κάτι τέτοιο. Ίσως να οφείλεται στον διαφορετικό επαγγελματικό προσανατολισμό των μαθητών.

Εντύπωση προκαλεί η ισότητα των μέσων στα Αρχαία Ελληνικά παρότι θα αναμενόταν να είναι διαφορετικά. Αυτό πιθανόν να οφείλεται στο ότι οι περισσότεροι μαθητές αντιμετωπίζουν το μάθημα αυτό ως ξένη γλώσσα.

Άλλος λόγος θα μπορούσε να είναι η διδακτική προσέγγιση μέσω των σχολικών βιβλίων ή διδακτικών μεθόδων για τα Αρχαία Ελληνικά.

Από την ανάλυση των δεδομένων που προηγήθηκε προκύπτει ότι η ομαδοποίηση στις κατηγορίες δεν είναι εφικτή.

Για το λόγο αυτό τα συμπεράσματα της έρευνας δεν μας δίνουν αρκετή πληροφορία ώστε να τα ανάγουμε στον γενικό πληθυσμό.

Πίνακες, διαγράμματα και συμπεράσματα για την correspondence analysis

Τα δεδομένα που επελέγησαν για την μελέτη κατεύθυνσης σπουδών ανά κατηγορία αφορούν μαθητές της Γ΄ Λυκείου διότι η επιλογή τους αυτή είναι καθοριστική για την διαδικασία των Πανελλαδικών Εξετάσεων και κατ΄ επέκταση για τον επαγγελματικό τους προσανατολισμό..

Επειδή $r = 3$και $k = 3$ έχουμε παραγοντική correspondence analysis.

Correspondence Table

	ΚΑΤΕΥΘΥΝΣΗ ΣΠΟΥΔΩΝ			
ΚΑΤΗΓΟΡΙΑ	ΘΕΩΡΗΤΙΚΗ	ΘΕΤΙΚΗ	ΤΕΧΝΟΛΟΓΙΚΗ 2	Active Margin
ΕΛΛΗΝΟΠΑΙΔΕΣ	45	9	28	82
ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	27	12	13	52
ΑΛΛΟΔΑΠΟΙ	8	1	2	11
Active Margin	80	22	43	145

Πίνακας 4.15 Πίνακας συνάφειας ανά κατηγορία και κατεύθυνση σπουδών.

Ο correspondence table είναι ο πίνακας συχνοτήτων. Π.χ 2 μαθητές που είναι αλλοδαποί και παρακολουθούν την κατεύθυνση ΤΕΧΝΟΛΟΓΙΚΗ 2.

Row Profiles

	ΚΑΤΕΥΘΥΝΣΗ ΣΠΟΥΔΩΝ			
ΚΑΤΗΓΟΡΙΑ	ΘΕΩΡΗΤΙΚΗ	ΘΕΤΙΚΗ	ΤΕΧΝΟΛΟΓΙΚΗ 2	Active Margin
ΕΛΛΗΝΟΠΑΙΔΕΣ	,549	,110	,341	1,000
ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	,519	,231	,250	1,000
ΑΛΛΟΔΑΠΟΙ	,727	,091	,182	1,000
Mass	,552	,152	,297	

Πίνακας 4.16 Πίνακας ποσοστών ανά γραμμή

Ο Row Profiles Πίνακας 4.16 είναι ο πίνακας ποσοστών των τιμών κάθε στήλης ανά γραμμή. Π.χ σε σύνολο 82 ελληνοπαίδων οι 9 ακλουθούν θετική κατεύθυνση δηλ το 11% κ.λ.π.

Column Profiles

	ΚΑΤΕΥΘΥΝΣΗ ΣΠΟΥΔΩΝ			
ΚΑΤΗΓΟΡΙΑ	ΘΕΩΡΗΤΙΚΗ	ΘΕΤΙΚΗ	ΤΕΧΝΟΛΟΓΙΚΗ 2	Mass
ΕΛΛΗΝΟΠΑΙΔΕΣ	,563	,409	,651	,566
ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	,338	,545	,302	,359
ΑΛΛΟΔΑΠΟΙ	,100	,045	,047	,076
Active Margin	1,000	1,000	1,000	

Πίνακας 4.17 Πίνακας ποσοστών ανά στήλη

Ο Column Profiles Πίνακας 4.17 είναι ο πίνακας ποσοστών των τιμών κάθε γραμμής ανά στήλη. Π.χ σε σύνολο 22 μαθητών της θετικής κατεύθυνσης οι 9 είναι ελληνόπαιδες δηλ το 40.9% κ.λ.π.

Summary

					Proportion of Inertia		Confidence Singular Value	
								Correlation
Dimension	Singular Value	Inertia	Chi Square	Sig.	Accounted for	Cumulative	Standard Deviation	2
1	,168	,028			,744	,744	,086	,031
2	,099	,010			,256	1,000	,076	
Total		,038	5,503	,239[a]	1,000	1,000		

a. 4 degrees of freedom

Πίνακας 4.18 Συνοπτικός Πίνακας Correspondence Analysis

Ο συνοπτικός Πίνακας 4.18 (Summary table) μας δίνει ποικιλία από χρήσιμες πληροφορίες.

Οι παράγοντες (Dimension-1 και Dimension-2) έχουν εξαχθεί με τέτοιο τρόπο ώστε να μεγιστοποιούν τις αποστάσεις μεταξύ των τιμών των γραμμών ή των στηλών.

Διαδοχικοί παράγοντες ,οι οποίοι είναι ανεξάρτητοι ή ορθογώνιοι μεταξύ τους, θα εξηγούν όλο και λιγότερο από την συνολική τιμή του Chi-square και κατά συνέπεια και της συνολικής τιμής της inertia.

Οι παράγοντες αυτοί προσομοιάζουν με τους αντίστοιχους παράγοντες της PCA.

Στο συγκεκριμένο παράδειγμα έχουμε:

$$\text{Total inertia} = \frac{\text{chi-square}}{\text{n}} = \frac{5.503}{145} = 0.038.$$

Ο πρώτος παράγοντας εξηγεί το 74.4% της συνολικής inertia.

Κάτω από την υπόθεση H_0: Πλήρης ανεξαρτησία (Total inertia =0) σε διάστημα εμπιστοσύνης 5% είναι p-value=0,239 > 0,05 άρα δεν απορρίπτεται η H_0 δηλαδή φαίνεται ότι δεν συσχετίζεται η κατηγορία (Ελληνόπαιδες, Παλιννοστούντες, Αλλοδαποί) με την επιλογή της κατεύθυνσης σπουδών.

Το ίδιο αποτέλεσμα προκύπτει και από την τιμή της X^2 με $(r-1)(k-1)=(3-1)(3-1)=4$ βαθμούς ελευθερίας.

Είναι $X^2_{4.0,95} = 9,49$ δηλαδή 5,503 < 9,49 άρα δεν απορρίπτεται η H_0.

Σημειωτέον ότι η X^2 εδώ δεν ελέγχει μοντέλα διαφορετικών διακυμάνσεων όπως συνήθως συμβαίνει αλλά ελέγχει μόνο την τιμή της inertia γύρω από το 0.

Overview Row Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
ΚΑΤΗΓΟΡΙΑ	Mass	1	2	Inertia	1	2	1	2	Total
ΕΛΛΗΝΟΠΑΙΔΕΣ	,566	-,315	-,132	,010	,335	,100	,907	,093	1,000
ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	,359	,547	-,021	,018	,640	,002	,999	,001	1,000
ΑΛΛΟΔΑΠΟΙ	,076	-,237	1,081	,009	,025	,899	,076	,924	1,000
Active Total	1,000			,038	1,000	1,000			

a. Symmetrical normalization

Πίνακας 4.19 Πίνακας αξιολόγησης ανά γραμμή

Overview Column Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
ΚΑΤΕΥΘΥΝΣΗ ΣΠΟΥΔΩΝ	Mass	1	2	Inertia	1	2	1	2	Total
ΘΕΩΡΗΤΙΚΗ	,552	-,097	,273	,005	,031	,417	,178	,822	1,000
ΘΕΤΙΚΗ	,152	,945	-,164	,023	,807	,041	,983	,017	1,000
ΤΕΧΝΟΛΟΓΙΚΗ 2	,297	-,303	-,424	,010	,162	,542	,465	,535	1,000
Active Total	1,000			,038	1,000	1,000			

a. Symmetrical normalization

Πίνακας 4.20 Πίνακας αξιολόγησης ανά στήλη

Οι δύο παραπάνω Πίνακες (4.19, 4.20) δίνουν τις τιμές που επιτρέπουν να αξιολογηθεί πως κάθε γραμμή ή στήλη συνεισφέρει στου παράγοντες και πως κάθε παράγοντας συνεισφέρει στις γραμμές ή στις στήλες.

Confidence Row Points

	Standard Deviation in Dimension		Correlation
ΚΑΤΗΓΟΡΙΑ	1	2	1-2
ΕΛΛΗΝΟΠΑΙΔΕΣ	,154	,189	-,798
ΠΑΛΙΝΟΣΤΟΥΝΤΕΣ	,148	,314	,141
ΑΛΛΟΔΑΠΟΙ	1,061	,443	,312

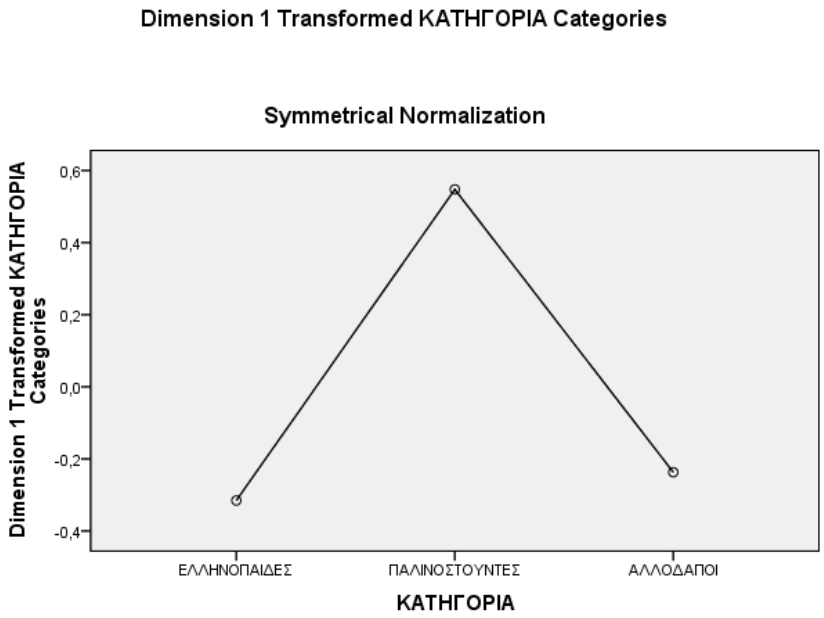
Πίνακας 4.21 Πίνακας τυπικής απόκλισης ανά γραμμή

Confidence Column Points

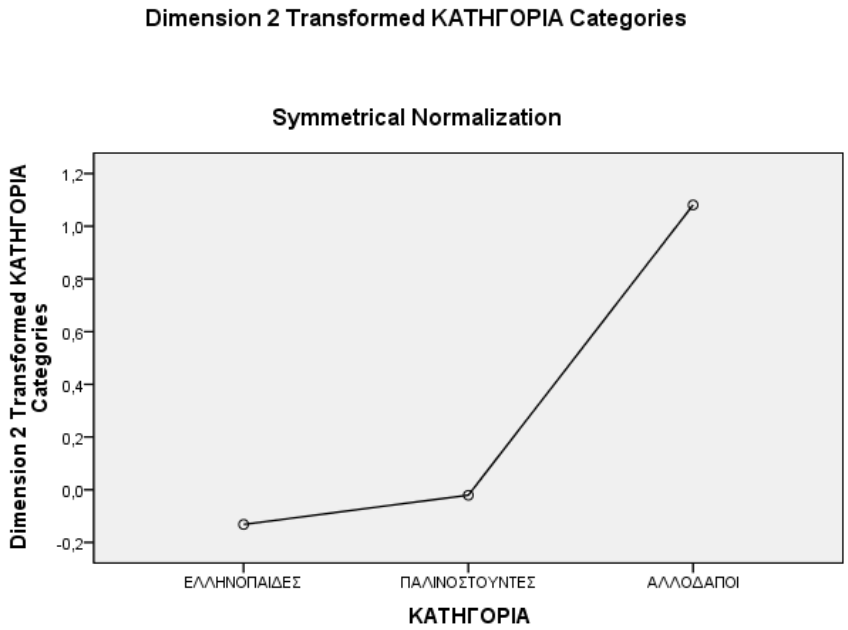
	Standard Deviation in Dimension		Correlation
ΚΑΤΕΥΘΥΝΣΗ ΣΠΟΥΔΩΝ	1	2	1-2
ΘΕΩΡΗΤΙΚΗ	,301	,125	,495
ΘΕΤΙΚΗ	,305	,600	,577
ΤΕΧΝΟΛΟΓΙΚΗ 2	,461	,252	-,742

Πίνακας 4.22 Πίνακας τυπικής απόκλισης ανά στήλη

Οι παραπάνω πίνακες (4.21, 4.22) δίνουν την τυπική απόκλιση και τη συσχέτιση μεταξύ των τιμών των παραγόντων

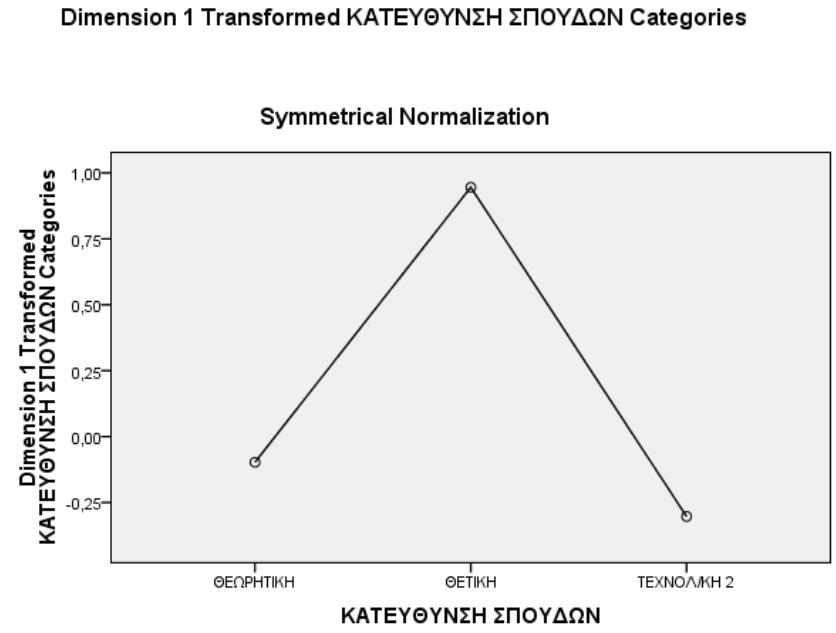


Διάγραμμα 4.6 Διαγραμματική απεικόνιση D1 ανά κατηγορία

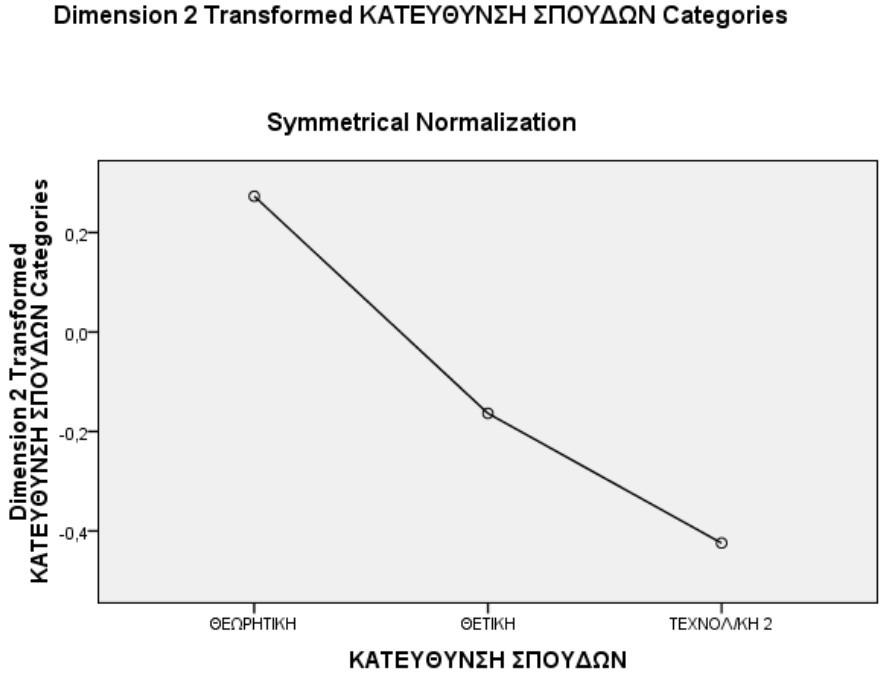


Διάγραμμα 4.7 Διαγραμματική απεικόνιση D2 ανά κατηγορία

Τα παραπάνω δύο διαγράμματα (4.6, 4.7) δείχνουν τα scores για κάθε κατηγορία για τους παράγοντες 1 και 2.

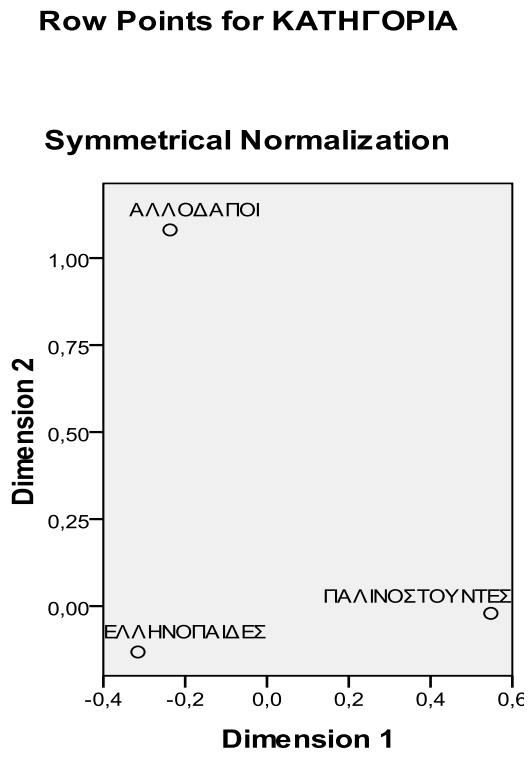


Διάγραμμα 4.8 Διαγραμματική απεικόνιση D1 ανά κατεύθυνση

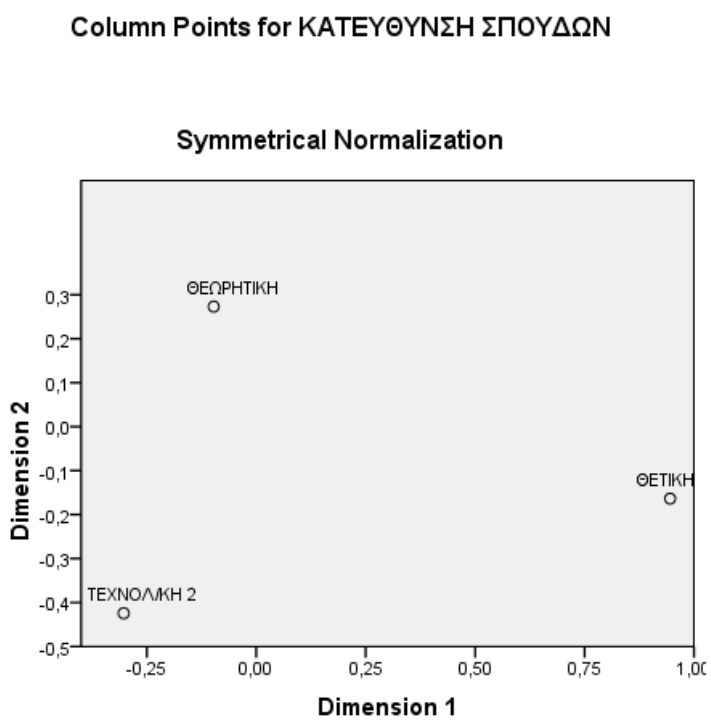


Διάγραμμα 4.9 Διαγραμματική απεικόνιση D2 ανά κατεύθυνση σπουδών

Τα παραπάνω δύο διαγράμματα (4.8, 4,9) δείχνουν τα scores για κάθε για κάθε κατεύθυνση σπουδών για τους παράγοντες 1 και 2.

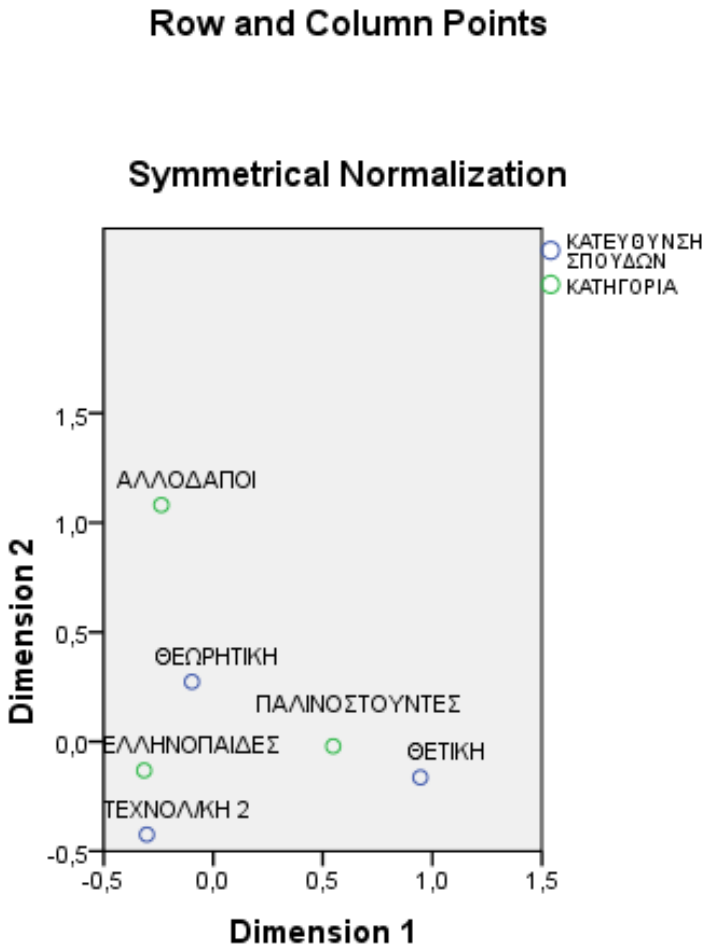


Διάγραμμα 4.10 Διαγραμματική απεικόνιση των Scores των D1 D2 ανά κατηγορία



Διάγραμμα 4.11 Διαγραμματική απεικόνιση των Scores των D1 D2 ανά κατεύθυνση σπουδών

Τα παραπάνω δύο διαγράμματα (4.10, 4.11) δείχνουν τα scores για κατηγορία και κάθε κατεύθυνση σπουδών για τους παράγοντες 1 και 2 ταυτόχρονα.



Διάγραμμα 4.12. Biplot

Τέλος, ο biplot correspondence map (διάγραμμα 4.12) δείχνει το score στις δύο διαστάσεις ταυτόχρονα και για την κατηγορία και την κατεύθυνση σπουδών.

Τώρα μπορούμε να δούμε τη χρησιμότητα των scores για τα μέτρα της απόστασης με την ερμηνεία των δύο διαστάσεων του μοντέλου μας, που είναι 100% μιας και $R = 2$.

Τα scores μας επιτρέπουν να συγκρίνουμε κατηγορία και κατεύθυνση σπουδών διαμέσου των παραγόντων σε χώρο δύο διαστάσεων.

Σημειωτέον ότι η συσχέτιση είναι μία τυποποιημένη μέτρηση της σχέσης ανάμεσα σε δύο συνήθως συνεχείς μεταβλητές.

Η correspondence είναι μία τυποποιημένη μέτρηση μεταξύ των παραγόντων (στην περίπτωση μας δύο) σε πολλαπλές μεταβλητές .

Αξίζει να σημειωθεί ότι οι παράγοντες (Dimensions) παράγουν εμπειρικά άξονες ή ιδιοδιανύσματα και όχι μόνο τις μεταβλητές που εισάγονται στην ανάλυση.

Συμπεράσματα της correspondence analysis.

Έτσι όπως προκύπτει από τον biplot οι Ελληνόπαιδες φαίνεται να προτιμούν την θεωρητική και την τεχνολογική κατεύθυνση περισσότερο από την θετική (πλησιέστερες αποστάσεις) ενώ οι αλλοδαποί είναι πλησιέστερα στην θεωρητική σε σχέση με την θετική και την τεχνολογική κατεύθυνση.

Όμως επειδή η συνολική inertia 3,8% όπως προέκυψε από τους παραπάνω πίνακες δεν διαφέρει στατιστικά σημαντικά από το 0 (τιμή της X^2 και τιμή του p-value) σημαίνει πως οι παραπάνω τάσεις δεν είναι σε τόσο έντονο βαθμό ώστε να μας δίνουν αρκετή πληροφορία/στοιχεία για γενίκευσή τους στον πληθυσμό και για διαφοροποίηση στις προτιμήσεις επιλογών των μαθητών.

ΠΑΡΑΤΗΜΑ Α

Βασικές έννοιες Γραμμικής Άλγεβρας και Αναλυτικής Γεωμετρίας.

Όπως αναφέρεται από την συγγραφική ομάδα των Δημήτριου Βάρσου, Δημήτρη Δεριζιώτη, Μιχάλη Μαλιάκα, Σταύρου Γ Παπασταυρίδη, Ευάγγελου Ράπτη και Ολυμπίας Ταλέλλη καθώς και από τον Θανάση Χρυσάκη προκύπτουν οι παρακάτω ορισμοί και παραδείγματα που αφορούν στη Γραμμική Άλγεβρα και Αναλυτική Γεωμετρία.

Βασικές έννοιες Γραμμικής Άλγεβρας.

1.Ορισμός του ίχνους (trace) ενός πίνακα A **ή** $tr(A)$

Ορίζουμε ως ίχνος (trace) ενός τετραγωνικού ($n \text{ x } n$) πίνακα A το άθροισμα των στοιχείων της κύριας διαγωνίου του δηλ: $tr(A) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + ... + a_{nn}$.

2.Ορισμός εσωτερικού γινομένου

Εσωτερικό γινόμενο δύο διανυσμάτων $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_n]$ και $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_n]$ ορίζουμε τον πραγματικό αριθμό $\boldsymbol{\alpha} \cdot \boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i \beta_i = \alpha_1 \beta_1 + \alpha_2 \beta_2 + ... + \alpha_n \beta_n$.

Αν πρόκειται για διανύσματα στηλών τότε το εσωτερικό τους γινόμενο προκύπτει αν πολλαπλασιασθεί το ανάστροφο του ενός με το άλλο και μετά υπολογισθεί η ορίζουσα δηλ: $\boldsymbol{\alpha} \cdot \boldsymbol{\beta} = \det(\boldsymbol{\alpha}^T \cdot \boldsymbol{\beta})$.

3.Ορισμός της τάξης ενός πίνακα(Rank of a Matrix) με αντίστοιχα παραδείγματα.

Έστω ένας $r \text{ x } c$ πίνακας. Μπορούμε να θεωρήσουμε τον πίνακα αυτόν σαν ένα σύνολο r διανυσμάτων γραμμών: $\boldsymbol{r}_i = (r_{i1}, ..., r_{ic})$, $i = 1, ..., r$ καθένα από τα οποία περιέχει c στοιχεία ή σαν ένα σύνολο c διανυσμάτων στηλών: $\boldsymbol{c}_j = (c_{1j}, ..., c_{rj})^T$, $j = 1, ..., c$ καθένα από τα οποία περιέχει r στοιχεία.

Η τάξη του πίνακα (Rank) ορίζεται ως:

α) Ο μεγαλύτερος αριθμός των γραμμικά ανεξάρτητων διανυσμάτων στηλών ή

β) Ο μεγαλύτερος αριθμός των γραμμικά ανεξάρτητων διανυσμάτων γραμμών.

Και οι δύο ορισμοί είναι ισοδύναμοι.

Για ένα r x c πίνακα:

α) Αν το r είναι μικρότερο του c τότε τάξη του πίνακα είναι το r

β) Αν το c είναι μικρότερο του r τότε τάξη του πίνακα είναι το c

Η τάξη ενός πίνακα μπορεί να γίνει μηδέν μόνο αν ο πίνακας δεν περιέχει στοιχεία.

Αν περιέχει έστω και ένα στοιχείο τότε η τάξη είναι μεγαλύτερη από το μηδέν.

Ένας πίνακας ονομάζεται πλήρους τάξης (Full Rank Matrix) αν όλες οι γραμμές ή οι στήλες είναι γραμμικά ανεξάρτητες.

Παραδείγματα εύρεσης της τάξης ενός πίνακα.

Παράδειγμα 1.

Έστω ο πίνακας $A=\begin{pmatrix}1 & 2 & 4 & 4\\ 3 & 4 & 8 & 0\end{pmatrix}$ με $r=2$ και $c=4$. Επειδή $r<c$ η μέγιστη τάξη του ισούται με το μέγιστο αριθμό των γραμμικά ανεξάρτητων γραμμών.

Επειδή η r_1 και η r_2 είναι γραμμικά ανεξάρτητες η τάξη του είναι 2.

Είναι πίνακας πλήρους τάξης.

Παράδειγμα 2.

Έστω ο πίνακας $B=\begin{pmatrix}1 & 2 & 3\\ 2 & 3 & 5\\ 3 & 4 & 7\\ 4 & 5 & 9\end{pmatrix}$ με $r=4$ και $c=3$. Επειδή $c<r$ η μέγιστη τάξη του ισούται με το μέγιστο αριθμό των γραμμικά ανεξάρτητων στηλών.

Η c_1 και η c_2 είναι γραμμικά ανεξάρτητες όμως η c_3 είναι γραμμικά εξαρτημένη διότι $c_{i3}=c_{i1}+c_{i2}, i=1,...,4$. Επομένως η τάξη του είναι 2 και δεν είναι πλήρους τάξης.

5.Ορισμός ιδιοδιανυσμάτων και ιδιοτιμών ενός τετραγωνικού πίνακα.

Έστω ο τετραγωνικός πίνακας A. Ο $M=I_n-\lambda A$ λέγεται χαρακτηριστικός πίνακας του A.

Η $\det M=\det(I_n-\lambda A)$ είναι πολυώνυμο ως προς λ και βαθμού n και λέγεται χαρακτηριστικό πολυώνυμο του A. Συμβολίζεται με $\Delta(\lambda)$.

Π.χ για 2x2 πίνακα είναι: $\Delta(\lambda)=\lambda^2-(trA)\lambda+\det A$.

Έστω πίνακας A και έστω ότι για κάποιο διάνυσμα u και κάποιο $\lambda\in\mathbb{R}$ είναι:

$$Au = \lambda u \quad (1).$$

Σ΄ αυτή την περίπτωση θα λέμε ότι το u είναι ένα ιδιοδιάνυσμα του A και λ η αντίστοιχη ιδιοτιμή.

Οι ιδιοτιμές βρίσκονται από την εξίσωση $\Delta(\lambda)=0$ ενώ τα ιδιοδιανύσματα από την εξίσωση (1).

ΒΙΒΛΙΟΓΡΑΦΙΑ

- **Wolfgang, H. Leopold, S. (2003)** "*Applied Multivariate Statistical Analysis*". Springer –Verlang Berlin, Heidelberg New York.
- **Abdi, H. (2007)** "*Discriminant correspondence analysis*." In: N.J. Salkind (Ed.): *Encyclopedia of Measurement and Statistic*. Thousand Oaks (CA): Sage. pp. 270–275
- **Ahdesmäki, M.; Strimmer K. (2010)** "*Feature selection in omics prediction problems using cat scores and false nondiscovery rate control"*, *Annals of Applied Statistics*, 4 (1), 503–519.
- **Fisher, R. A. (1936).** "*The Use of Multiple Measurements in Taxonomic Problems"*. *Annals of Eugenics* **7** (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x. hdl:2440/15227.
- **Friedman, J. H. (1989).** "*Regularized Discriminant Analysis*". *Journal of the American Statistical Association* (American Statistical Association) **84** (405): 165–175. doi:10.2307/2289860. JSTOR 2289860. MR0999675. http://www.slac.stanford.edu/cgi-wrap/getdoc/slac-pub-4389.pdf.
- **Martinez, A. M.; Kak, A. C. (2004).** "*PCA versus LDA*". *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23** (2): 228–233. doi:10.1109/34.908974. http://www.ece.osu.edu/~aleix/pami01.pdf.
- **McLachlan, G. J. (2004).** "*Discriminant Analysis and Statistical Pattern Recognition*. "Wiley Interscience. ISBN 0471691151. MR1190469.
- **Perriere, G.; & Thioulouse, J. (2003).** "*Use of Correspondence Discriminant Analysis to predict the subcellular location of bacterial proteins*", *Computer Methods and Programs in Biomedicine*, 70, 99–105.
- **Venables, W. N.; Ripley, B. D. (2002).** *"Modern Applied Statistics with S*"(4th ed.). Springer Verlag. ISBN 0-387-95457-0.
- **Yu, H.; Yang, J. (2001)**. "*A direct LDA algorithm for high-dimensional data — with application to face recognition"*, *Pattern Recognition*, 34 (10), 2067–2069.
- **Sathyanarayana, Shashi (2010*).*** *"Pattern Recognition Primer II"*. *Wolfram Demonstrations Project*. .

- **Mahalanobis, P C (1936).** "*On the generalised distance in statistics*". *Proceedings of the National Institute of Sciences of India* **2** (1): 49–55. http://www.insa.ac.in/insa_pdf/20005b8c_49.pdf. Retrieved 2008-11-05.
- **Παπαγεωργίου Ι. (2010)** Πανεπιστημιακές παραδόσεις ΟΠΑ
- **Murray R. Spiegel (1975)** *"Πιθανότητες και Στατιστική"* Mc Graw-Hill New York ΕΣΠΙ, Αθήνα.
- **Βάρσος Δ. Δεριζιώτης Δ. Μαλιάκας Μ. Παπασταυρίδης Σ. Ράπτης Ε. Ταλέλλη Ο. (2003)** *"Εισαγωγή στη Γραμμική Άλγεβρα"* Εκδόσεις σοφία , Θεσσαλονίκη.
- **Χρυσάκης Θ. (1998)** *"Γραμμική Άλγεβρα και Αναλυτική Γεωμετρία*" Εκδόσεις THANASSIS CHRYSSAKIS Αθήνα.

ΔΙΑΔΙΚΤΥΑΚΟΙ ΤΟΠΟΙ

http://www.cs.uu.nl/docs/vakken/wo2/SPSS/spss6.pdf

http://www.unt.edu/