Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΑΛΥΣΗ ΕΡΩΤΗΜΑΤΟΛΟΓΙΟΥ ΜΕΣΩ ΑΝΑΛΥΣΗΣ ΑΝΤΙΣΤΟΙΧΙΩΝ ΚΑΙ ΜΟΝΤΕΛΩΝ ΣΥΝΑΦΕΙΑΣ

Βασίλειος Ι. Καδδίτης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Δεκέμβριος 2008

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ..…… συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- …… (Επιβλέπων)
- ……
- ……

Η έγκριση της Διπλωματικής Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

QUESTIONNAIRE ANALYSIS THROUGH CORRESPONDENCE ANALYSIS AND ASSOCIATION MODELS

By

Vasilios J. Kadditis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Pireaus, Greece
December 2008

Η εργασία αυτή είναι αφιερωμένη στην οικογένεια μου για την ανυστερόβουλη στήριξη όλα αυτά τα χρόνια

Περίληψη

Η παρούσα εργασία πραγματεύεται πολυδιάστατα κατηγορικά δεδομένα όπως αυτά προκύπτουν από συλλογή μέσω ερωτηματολογίου. Για να αναλυθεί όμως ένα ερωτηματολόγιο το οποίο περιλαμβάνει πλήθος ερωτήσεων-μεταβλητών και να εξάγουμε ορισμένα χρήσιμα συμπεράσματα θα πρέπει, πρώτα από όλα, να του επιβάλουμε δομή στις περιπτώσεις εκείνες που η δομή απουσιάζει. Εφόσον η δομή αυτή επιτευχθεί, τότε χρησιμοποιώντας στατιστικές τεχνικές και μεθόδους κατάλληλες για κατηγορικά δεδομένα μπορούμε πιο εύκολα να μελετήσουμε τις σχέσεις μεταξύ των μεταβλητών για τις οποίες έχουμε δει ότι παρουσιάζουν το μεγαλύτερο ενδιαφέρον για περαιτέρω ανάλυση. Με τον τρόπο αυτό, μπορεί κανείς να πει ότι, η εργασία χωρίζεται σε δύο μεγάλες θεματικές ενότητες˙ το πρώτο μέρος αφορά τον τρόπο με τον οποίο δημιουργούμε τη δομή βρίσκοντας ομοειδείς ομάδες μεταβλητών και το δεύτερο μέρος αφορά τη μελέτη και ανάλυση μιας ομάδας ή υποσυνόλου μικρότερης διάστασης.

Για τον σκοπό αυτό, παρουσιάζουμε και αναλύουμε τη θεωρία της Παραγοντικής Ανάλυσης και της Ανάλυσης Κατά Συστάδες, δίνοντας ιδιαίτερη έμφαση στον τρόπο με τον οποίο ομαδοποιούμε κατηγορικές μεταβλητές. Ύστερα, συγκρίνουμε τα αποτελέσματα που δίνουν οι δύο διαφορετικές μέθοδοι όταν τις χρησιμοποιήσουμε για την ανάλυση του ερωτηματολογίου. Τα συμπεράσματα που προκύπτουν έχουν ιδιαίτερο ενδιαφέρον.

Επιπλέον, για την περαιτέρω ανάλυση ενός υποσυνόλου μεταβλητών βασιζόμαστε στα μοντέλα συνάφειας. Συγκεκριμένα, παρουσιάζουμε αναλυτικά τα μοντέλα συνάφειας για διδιάστατους και τρισδιάστατους πίνακες και δείχνουμε πώς μπορούν να χρησιμοποιηθούν από τους ερευνητές για τη μελέτη διατάξιμων αλλά και ονομαστικών κατηγορικών μεταβλητών. Η δυναμικότητα των μοντέλων αυτών παρουσιάζεται μέσα από μία εφαρμογή από τον χώρο των κοινωνικών επιστημών σε θέματα που αφορούν τις οικογένειες.

Τέλος για τη περαιτέρω ανάλυση μιας μικρότερης ομάδας μεταβλητών παρουσιάζουμε μια εναλλακτική μέθοδο των μοντέλων συνάφειας, την Ανάλυση Αντιστοιχιών και τα Μοντέλα

Συσχέτισης ενώ δεν παραλείπουμε να αναφερθούμε και στην Πολλαπλή Ανάλυση Αντιστοιχιών. Στις μεθόδους αυτές δίνεται έμφαση κυρίως στα γραφικά αποτελέσματα αλλά και στις εκτιμήσεις των σκορ των κατηγοριών των μεταβλητών.

Όλα τα παραπάνω τα συγκρίνουμε κριτικά μεταξύ τους στη θεωρία και στη πράξη έτσι ώστε ο ενδιαφερόμενος αναγνώστης να κατανοήσει περισσότερο τις μεθόδους αυτές και να αποκομίσει όσο το δυνατόν περισσότερες πληροφορίες που θα τον βοηθήσουν για την εφαρμογή τους.

Abstract

This dissertation deals with multivariate categorical data of a raw data set produced by a questionnaire designed for a research purpose. However, in order to analyze a questionnaire and extract some fruitful results, that includes a great number of questions-variables, we must first impose a structure on it especially on situations this specific structure is missing. Whenever the structure is imposed, by using statistical techniques and methods designed for categorical data, we can then study more efficiently the relations among the variables in concern for further analyses. In this way, it is better to view this dissertation as having two thematic sections, one that deals with the procedures to find and impose this structure to the whole set of variables and the other that involves the study and further analysis of the relations in a smaller group of variables.

For that purpose we present and develop the theoretical issues of Factor Analysis and Cluster Analysis, by focusing our attention in the approach of finding similar groups of categorical variables that behave in a similar manner. At the end we compare those results obtained from the two different methods by implementing them on a research questionnaire. These results show some very interesting findings.

Moreover, for the analysis of a smaller subset of categorical variables we use association models. We present association models for two-way and three-way contingency tables in an analytical and exploratory way and show how these models are very useful and help researchers clarify the relations among variables of ordinal but also nominal scales. Their utility is exhibited through an example of a tree-way contingency table taken from real data regarding family issues.

Finally, the analysis on a smaller subset is further explored by describing the issues of Correspondence Analysis, Correlation Models and Multiple Correspondence Analysis. In all of those methods, we focus on the interpretation of the results on the graphical displays of the data but also on the estimated category scores of the variables.

The above methods described in this dissertation and the results after implementing them are all critically compared with each other at each chapter. This gives to the interesting reader the possibility to fully understand them and to obtain additional information on their implementation.

Πίνακας Περιεχομένων

Περίληψη **vii**

Abstract **ix**

1. ΕΙΣΑΓΩΓΗ **1**

2. ΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ **5**

2.1 Εισαγωγή 5

2.2 Ορθογώνιο παραγοντικό μοντέλο 5

2.2.1 Μέθοδοι Εκτίμησης του Παραγοντικού Μοντέλου 7

2.3 Έλεγχος των Συσχετίσεων-Αριθμός των Παραγόντων 11

2.4 Περιστροφή των Παραγόντων 12

2.5 Σκορ Παραγόντων 14

2.6 Μη ορθογώνια Παραγοντική Ανάλυση 16

2.7 Σχόλια και Συμπεράσματα 16

2.8 Εφαρμογή της Μεθόδου 17

3. ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ **23**

3.1 Εισαγωγή 23

3.2 Ομοιογένεια Μεταβλητών 24

3.3 Μέτρα Ομοιότητας και Απόστασης για τις Μεταβλητές 26

3.3.1 Οι Μεταβλητές είναι Ποσοτικές 27

3.3.2 Οι Μεταβλητές είναι Ποιοτικές 28

3.3.3 Οι Συσχετίσεις ως Μέτρα Ομοιότητας ή Ανομοιότητας 30

3.4 Αλγόριθμοι και Μέθοδοι Ομαδοποίησης 31
3.4.1 Μέθοδος της Απλής Συνένωσης 32
3.4.2 Μέθοδος της Πλήρους Συνένωσης 33
3.4.3 Μέθοδος της Μέσης Ομάδας ή των μη Σταθμισμένων Μέσων 33
3.4.4 Μέθοδος του Ward 34
3.4.5 Διαιρετικές Μέθοδοι 35
3.4.6 Σχόλια Περί των Μεθόδων 36

3.5 Αξιολόγηση της Μεθόδου και Επιλογή Συστάδων 37

3.6 Εφαρμογή 39

4. ΜΟΝΤΕΛΑ ΣΥΝΑΦΕΙΑΣ ΓΙΑ ΠΙΝΑΚΕΣ 47

4.1 Μοντέλα Συνάφειας για Διδιάστατους Πίνακες 47
4.1.1 Εισαγωγή 47
4.1.2 Μοντέλο Συνάφειας Τύπου RC(M) 48

4.2 Μοντέλα Συνάφειας για Τρισδιάστατους Πίνακες 51
4.2.1 Εισαγωγή 51
4.2.2 Λογαριθμογραμμικά Μοντέλα 52
4.2.3 Αναλύοντας μόνο τις Αλληλεπιδράσεις $2^{ης}$ Τάξης 55
4.2.4 Αλληλεπίδραση τριών Παραγόντων 57
4.2.4.1 Διάφορα log-Trilinear Μοντέλα 58

4.3 Εφαρμογή 62

5. ΑΝΑΛΥΣΗ ΑΝΤΙΣΤΟΙΧΙΩΝ ΚΑΙ ΜΟΝΤΕΛΑ ΣΥΣΧΕΤΙΣΗΣ 71

5.1 Γενικά 71

5.2 Ανάλυση Αντιστοιχιών για Πίνακες Συνάφειας 2 Διαστάσεων 72
5.2.1 Εισαγωγή 72
5.2.2 Μάζες και Προφίλ 73
5.2.3 Μέσο Προφίλ και Απόκλιση από την Ανεξαρτησία 74
5.2.4 Μέτρα Απόστασης-Σχέση με το χ^2 τεστ του Pearson 74
5.2.5 Ολική Αδράνεια και χ^2 Απόσταση 76
5.2.6 Σύνδεση της SVD με την Ολική Αδράνεια-Συντεταγμένες των Προφίλ 77
5.2.7 Απόλυτες Συνεισφορές και Σχετικές Συνεισφορές 80
5.2.8 Σχέση με τη Γενικευμένη SVD του Πίνακα Αντιστοιχιών 81

5.3 Πολλαπλή Ανάλυση Αντιστοιχιών 82
5.3.1 Πίνακας Δείκτης και Πίνακας Burt 83
5.3.2 Προσαρμοσμένες Βασικές Αδράνειες 84
5.4 Μοντέλα Συσχέτισης 85
5.4.1 Εισαγωγή 85
5.4.2 Μοντέλα Κανονικής Συσχέτισης και Μοντέλα Ανάλυσης Αντιστοιχιών 86
5.4.2.1 Canonical Correlation Ανάλυση 86
5.4.2.2 Μοντέλα Συσχέτισης 87
5.5 Εφαρμογή $1^{η}$ 91
5.6 Εφαρμογή $2^{η}$ 98

ΠΑΡΑΡΤΗΜΑ Α 107

ΠΑΡΑΡΤΗΜΑ Β 113

ΠΑΡΑΡΤΗΜΑ Γ 119

ΒΙΒΛΙΟΓΡΑΦΙΑ 123

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Στην πράξη, τα δεδομένα που έχει ένας αναλυτής στα χέρια του σχεδόν πάντα είναι πολυμεταβλητά. Έτσι τίθεται στη διακριτική ικανότητα και ευχέρεια του αναλυτή, η αναγκαιότητα χρησιμοποίησης όλων των δεδομένων για την αποκομιδή της μεγαλύτερης δυνατής πληροφορίας από αυτά. Ωστόσο χρησιμοποιώντας μεθόδους της πολυμεταβλητής στατιστικής ο αναλυτής μπορεί να μελετήσει καλύτερα τα φαινόμενα που τον απασχολούν χωρίς να χάσει μεγάλο μέρος της πληροφορίας των αρχικών δεδομένων.

Στην εργασία αυτή θα παρουσιάσουμε και θα χρησιμοποιήσουμε κάποιες πολυμεταβλητές τεχνικές για να αναλύσουμε πολυδιάστατα κατηγορικά δεδομένα. Οι μέθοδοι και τεχνικές που θα αναπτύξουμε πρώτα σε θεωρητικό επίπεδο, στη συνέχεια θα χρησιμοποιηθούν για την ανάλυση ενός ερωτηματολογίου της Τράπεζας *ABC* που έχει ως στόχο τη διερεύνηση της αποτελεσματικότητας του συστήματος αξιολόγησης των εργαζομένων της. Το εν λόγω ερωτηματολόγιο παρατίθεται στο Παράρτημα Α, ενώ για λόγους δέσμευσης προς τους ανθρώπους που διενέργησαν την έρευνα αυτή να μην δημοσιευθεί το όνομα της τράπεζας και **να μην δοθούν τα αρχικά δεδομένα**, το πραγματικό όνομα δεν δίνεται. Η δε έρευνα ήταν ανώνυμη. Το ερωτηματολόγιο αυτό, θα αποτελεί τη βασική μας εφαρμογή και παράδειγμα επίδειξης των μεθοδολογιών που θα ασχοληθούμε στη συνέχεια.

Στο 2^{ο} Κεφάλαιο αναπτύσσουμε την μέθοδο της Παραγοντικής Ανάλυσης. Με την μέθοδο προσπαθούμε να βρούμε και να ερμηνεύσουμε κοινούς παράγοντες οι οποίοι είναι αφανείς και πιστεύουμε ότι μπορούν να ερμηνεύσουν τις συσχετίσεις μεταξύ των παρατηρούμενων μεταβλητών. Επιπλέον, οι κοινοί παράγοντες χρησιμοποιούνται ως νέες μεταβλητές για την μείωση της διάστασης των δεδομένων αφού είναι έτσι κατασκευασμένοι ώστε να διατηρούν όσο γίνεται την αρχική πληροφορία.

Στο 3^{ο} Κεφάλαιο αναφερόμαστε στην τεχνική της Ανάλυσης κατά Συστάδες. Η τεχνική αυτή είναι πολύ χρήσιμη για την ομαδοποίηση των αντικειμένων ενός πολυδιάστατου πίνακα σε ομοιογενείς ομάδες είτε τα αντικείμενα είναι οι παρατηρήσεις είτε είναι οι μεταβλητές

(χαρακτηριστικά). Όμως στο κεφάλαιο αυτό για την ανάλυση κατά συστάδες θα δώσουμε το πλαίσιο στο οποίο χρησιμοποιείται για την ομαδοποίηση των μεταβλητών μόνο ενός συνόλου δεδομένων.

Πριν προχωρήσουμε όμως με τα επόμενα κεφάλαια πρέπει να πούμε ότι ο στόχος της εργασίας είναι να αναδείξουμε τα οφέλη από τις διάφορες τεχνικές και μεθόδους για τη μελέτη πολυδιάστατων κατηγορικών δεδομένων. Για τον λόγο αυτό, το ερωτηματολόγιο του Παραρτήματος Α όταν το αναλύουμε με τις παραπάνω μεθόδους, έχουμε σκοπό να δημιουργήσουμε μικρότερες ομάδες μεταβλητών ίσως για μια διεξοδικότερη μελέτη τους. Αυτός είναι ο λόγος που χρησιμοποιούμε την παραγοντική ανάλυση και την ανάλυση κατά συστάδες για την ομαδοποίηση των μεταβλητών. Να μειώσουμε δηλαδή τη διάσταση του προβλήματος φτιάχνοντας ομάδες μικρότερης διάστασης τις οποίες μπορούμε να μελετήσουμε ευκολότερα αλλά και να κάνουμε αν χρειαστεί επιπλέον υποθέσεις. Θα πρέπει να τονιστεί ότι επειδή έχουμε κατηγορικά δεδομένα, η χρήση των παραπάνω μεθόδων απαιτεί ιδιαίτερη προσοχή. Έτσι για το $2^{ο}$ και $3^{ο}$ Κεφάλαιο όταν στο τέλος αναλύουμε το ερωτηματολόγιο, κάνουμε την παραδοχή ότι αυτό είναι χωρίς δομή (δεν είναι χωρισμένο σε ενότητες) και βάσει της μεθοδολογία που περιγράφουμε προσπαθούμε να του «επιβάλλουμε» μία μέσω δύο διαφορετικών προσεγγίσεων. Αν τα αποτελέσματα που θα δώσουν οι δύο μέθοδοι είναι περίπου ίδια τότε είναι σίγουρο ότι η δομή που βρήκαμε είναι η σωστή. Επίσης για να συγκρίνουμε τα αποτελέσματα της παραγοντικής ανάλυσης με εκείνα της ανάλυσης των συστάδων, στο τέλος του $3^{ου}$ Κεφαλαίου χρησιμοποιούμε την *meta-analysis* ή *Q-factor analysis*. Η διαδικασία αυτή συνιστά αρχικά την χρήση της παραγοντικής ανάλυσης για την εύρεση των κοινών παραγόντων και στη συνέχεια την χρησιμοποίηση των φορτίων των παραγόντων για την ομαδοποίηση και τη δημιουργία των συστάδων. Η μεθοδολογία αυτή έχει συζητηθεί αρκετά από διάφορους συγγραφείς όπως ο *Cattell* (1965), *Milligan* and *Cooper* (1987), *Urban* and *McDaniel* (1990), *Cheung* and *Chan* (2005), να είναι ορισμένα από τα ονόματα που έχουν ασχοληθεί με τις δύο αυτές μεθόδους σε παραλληλία. Με αυτό τον τρόπο εάν οι συστάδες «υπάρχουν» στις νέες διαστάσεις που κατασκευάσαμε, τότε η κλασική παραγοντική ανάλυση θα είναι σε θέση να εντοπίσει σωστά την δομή των μεταβλητών. Τέλος συγκρίνουμε τη δομή που δημιούργησαν τα δεδομένα με εκείνη που *a priori* είχε θέσει ο ερευνητής και ανάλογα την επιβεβαιώνουμε ή όχι.

Στο $4^{ο}$ Κεφάλαιο παρουσιάζουμε τα μοντέλα συνάφειας για διδιάστατους και τρισδιάστατους πίνακες, μαζί με μια σύντομη αναφορά στα λογαριθμογραμμικά μοντέλα. Στο

κεφάλαιο αυτό δείχνουμε πως τα μοντέλα αυτά χρησιμοποιούνται για τη μελέτη των σχέσεων των κατηγορικών μεταβλητών με διατάξιμη κλίμακα αλλά και στις περιπτώσεις εκείνες που τα μοντέλα μεταχειρίζονται τις μεταβλητές του πίνακα σαν ονομαστικές. Στο τέλος του κεφαλαίου υπάρχει και μια ενδιαφέρουσα εφαρμογή. Η σύνδεση του 4^{ου} Κεφαλαίου με τα δύο προηγούμενα είναι η εξής: Για παράδειγμα, μετά την ομαδοποίηση που δημιουργήσαμε ή μετά την εύρεση των παραγόντων, μπορούμε να αναλύσουμε τις σχέσεις των κατηγορικών μεταβλητών μέσα στην ομάδα της μικρότερης διάστασης από αυτές που έχουμε ήδη βρει φτιάχνοντας πίνακες συνάφειας τους οποίους μελετάμε με τα μοντέλα συνάφειας. Όμως οι ερωτήσεις-μεταβλητές του ερωτηματολογίου δεν χρησιμοποιήθηκαν για την ανάλυση αυτή γιατί το μέγεθος του δείγματος δεν ήταν επαρκές για να φτιάξουμε τουλάχιστον τρισδιάστατους πίνακες συνάφειας. Αλλά ας γίνουμε περισσότερο σαφής επί του θέματος αυτού:

Για να αναλύσουμε πολυδιάστατους πίνακες χρησιμοποιώντας λογαριθμογραμμικά μοντέλα και μοντέλα συνάφειας θα πρέπει πρωτίστως να έχουμε στη διάθεση μας ένα ικανοποιητικό μέγεθος δείγματος. Ο Cochran (1964) πρότεινε ότι τουλάχιστον 80% των κελιών πρέπει να έχουν αναμενόμενες συχνότητες πάνω από 5.0 και αυτές θα πρέπει να ξεπερνούν το 1.0 για όλα τα κελιά ενός πίνακα. Βάσει λοιπόν αυτού του πρακτικού θέματος τίθεται ένα μείζων πρόβλημα για την κατασκευή πολυδιάστατου πίνακα συνάφειας χρησιμοποιώντας τις μεταβλητές του ερωτηματολογίου της Τράπεζας *ABC*. Του ερωτηματολογίου οι μεταβλητές είναι σε πενταβάθμια κλίμακα *Likert* ενώ το μέγεθος του δείγματος δεν ξεπερνάει τους 150. Συνεπώς ακόμα και για την κατασκευή πίνακα συνάφειας 3 διαστάσεων οι αναμενόμενες συχνότητες των κελιών δεν θα ξεπερνούν το 1.2 και πολλά κελιά στην πράξη θα είναι με μηδέν συχνότητες. Αν τύχει να αναλύσουμε περισσότερες μεταβλητές θα διογκωθεί ακόμα πιο πολύ το πρόβλημα των μηδενικών κελιών. Ακόμα και αν συμπτύξουμε τις ακραίες κατηγορίες της κλίμακας γνωρίζοντας ότι θα χάσουμε πληροφορία ώστε να δημιουργήσουμε τριτοβάθμια κλίμακα για τις ερωτήσεις, στον νέο πίνακα συνάφειας 3 διαστάσεων που θα προκύψει, ο μέσος όρος της συχνότητας ενός κελιού θα είναι σχεδόν 5 παρατηρήσεις ενώ αρκετά από τα κελιά του πίνακα πάλι δεν θα έχουν συχνότητες! Για τα μοντέλα συνάφειας μεγαλύτερης διάστασης του τρία απαιτείται ο πίνακας να μην είναι αραιός αλλιώς τα αποτελέσματα δεν θα είναι ικανοποιητικά και στην πράξη δεν θα μπορέσουν να εκτιμηθούν οι αλληλεπιδράσεις των παραγόντων και να προβούμε σε συμπερασματολογία. Για τον λόγο αυτό την ανάλυση της σχέσης των μεταβλητών από ένα

υποσύνολο μικρότερης διάστασης την εφαρμόζουμε στο 4ο Κεφάλαιο σε ένα διαφορετικό σύνολο δεδομένων διαθέσιμο στη βιβλιογραφία. Άλλωστε αν έχουμε στη διάθεση μας ένα σύνολο δεδομένων με πολλές μεταβλητές και ένα καλό μέγεθος δείγματος τότε με βάση την μεθοδολογία που αναπτύσσουμε στο 2ο και 3ο Κεφάλαιο δημιουργούμε πίνακες συνάφειας διασταυρώνοντας τις μεταβλητές που θέλουμε να μελετήσουμε από τα υποσύνολα των μεταβλητών που δημιουργήσαμε – ο τρόπος της ανάλυσης του πίνακα αυτού σε γενικές γραμμές είναι ίδιος με την εφαρμογή που θα παρουσιάσουμε.

Τέλος στο 5ο Κεφάλαιο μιλάμε για την Ανάλυση Αντιστοιχιών και την Πολλαπλή Ανάλυση Αντιστοιχιών. Η μέθοδος της ανάλυσης αντιστοιχιών είναι πολύ ισχυρή για την ανάλυση κατηγορικών μεταβλητών σε μορφή διδιάστατου πίνακα συνάφειας. Επίσης γίνεται σύγκριση της μεθόδου με τα μοντέλα συσχέτισης. Από την άλλη η πολλαπλή ανάλυση αντιστοιχιών είναι ιδιαίτερα χρήσιμη για πολυδιάστατα κατηγορικά δεδομένα όπως τα ερωτηματολόγια. Στο τέλος του κεφαλαίου αναλύουμε την εφαρμογή του 4ου Κεφαλαίου μέσω της ανάλυσης αντιστοιχιών. Επιπλέον, εφαρμόζουμε την ανάλυση αντιστοιχιών στα δεδομένα του ερωτηματολογίου μας, δηλαδή τις υποομάδες των μεταβλητών όπου ερευνούμε την αποτελεσματικότητα του υπάρχοντος συστήματος αξιολόγησης από τα δεδομένα που έχουμε με την βοήθεια της πολλαπλής ανάλυσης αντιστοιχιών. Για τον σκοπό αυτό έχει δοθεί ιδιαίτερη έμφαση στη γραφική απόδοση τους.

Εν κατακλείδι στόχος της εργασίας αυτής είναι να δείξει τον τρόπο που χειριζόμαστε πολυδιάστατα κατηγορικά δεδομένα και τα οποία προέρχονται από έρευνες μέσω ερωτηματολογίων ή συναφών πηγών και ζητάμε να απομονώσουμε ορισμένα χαρακτηριστικά τα οποία θέλουμε να μελετήσουμε γιατί ικανοποιούν ορισμένες σημαντικές προϋποθέσεις όπως εκείνης της συσχέτισης ή συνάφειας (μεταξύ των μεταβλητών και μεταξύ των παρατηρήσεων).

ΚΕΦΑΛΑΙΟ 2

Παραγοντική Ανάλυση

2.1 Εισαγωγή

Συχνά συμβαίνει να ενδιαφερόμαστε να μετρήσουμε ορισμένες ποσότητες όπως για παράδειγμα την ευφυΐα, το κοινωνικό και οικονομικό *status*, την πολιτική ιδεολογία, το στρες, οι οποίες είναι αδύνατον να ποσοτικοποιηθούν άμεσα και να μετρηθούν. Όμως είναι δυνατόν να μετρήσουμε κάποιες άλλες «ποσότητες» οι οποίες εκφράζουν τις μεταβλητές για τις οποίες ενδιαφερόμαστε. Για να γίνει αυτό θεωρούμε για τις μεταβλητές που απαρτίζουν τον πίνακα δεδομένων $\mathbf{X}_{n\times p}$, ότι οι συσχετίσεις που υπάρχουν μεταξύ τους είναι «προϊόν» κάποιων αφανών (latent) παραγόντων που στην πραγματικότητα δεν υπάρχουν και πρέπει να εκτιμηθούν. Η παραγοντική ανάλυση είναι μια στατιστική τεχνική η οποία προσπαθεί να ερμηνεύσει τις συσχετίσεις μεταξύ των παρατηρούμενων μεταβλητών ενός πίνακα $\mathbf{X}_{n\times p}$ υποθέτοντας ότι οφείλονται αποκλειστικά στην ύπαρξη κάποιων κοινών παραγόντων οι οποίοι δεν είναι άμεσα εμφανείς. Με τον τρόπο αυτό αφενός πετυχαίνουμε μείωση της διάστασης των δεδομένων, αφού αντί να δουλεύουμε με τις αρχικές μεταβλητές τώρα δουλεύουμε με τους παράγοντες που «κατασκευάσαμε» και αφετέρου με έναν υποκειμενικό τρόπο να τους αναγνωρίσουμε ως κάποιες μη μετρήσιμες μεταβλητές. Ιδιαίτερα στην δεύτερη περίπτωση η παραγοντική ανάλυση βρίσκει ευρεία εφαρμογή στις κοινωνικές επιστήμες, την ψυχολογία και την ψυχομετρία, το marketing και στην έρευνα αγοράς.

2.2 Ορθογώνιο Παραγοντικό Μοντέλο

Έστω $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ είναι το $(p\times 1)$ διάνυσμα των παρατηρούμενων μεταβλητών του πίνακα $\mathbf{X}_{n\times p}$. Το πρόβλημα που προσπαθεί η παραγοντική ανάλυση να λύσει είναι κατά κάποιο τρόπο όμοιο με εκείνο της παλινδρόμησης για την πρόβλεψη ενός μεγέθους, μόνο που τώρα η σχέση αυτή αντιστρέφεται και προσπαθούμε να μάθουμε για τους παράγοντες όταν

γνωρίζουμε τις παρατηρούμενες μεταβλητές. Επιπλέον οι παράγοντες είναι άγνωστοι (latent), αλλά πιστεύεται ότι ο πίνακας των συνδιακυμάνσεων ή συσχετίσεων των παρατηρούμενων μεταβλητών περιέχει όλη την πληροφορία που χρειαζόμαστε για να εκτιμήσουμε την σχέση αυτή ώστε να συμπεράνουμε ότι οι μεταβλητές εξαρτώνται από έναν μικρότερο αριθμό μη παρατηρήσιμων (αφανών) μεταβλητών. Για τον λόγο αυτό, υποθέτουμε ότι οι p μεταβλητές $X_1, X_2, \ldots, X_p$ με μέσο $\boldsymbol{\mu}_{\mathbf{p\times 1}}$ και πίνακα συνδιακύμανσης $\boldsymbol{\Sigma}_{p\times p}$, μπορούν να γραφούν σαν γραμμικός συνδυασμός των m κοινών παραγόντων $F_1, F_2, \ldots, F_m$ και του σφάλματος $\boldsymbol{\varepsilon}_{p\times 1}$, με $m \ll p$. Η μορφή που θα έχουν τα δεδομένα είναι η εξής

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \ldots + \ell_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \ldots + \ell_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \ldots + \ell_{pm}F_m + \varepsilon_p \end{aligned} \tag{1.1}$$

Το μοντέλο (1.1) μπορεί να γραφτεί με μορφή πινάκων ως

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} \tag{1.2}$$

Οι συντελεστές ℓ_{ij} στην (1.1) λέγονται φορτία (loadings) τα οποία χρησιμεύουν για σταθμά ενώ κάτω από κατάλληλες προϋποθέσεις υποδηλώνουν τη σπουδαιότητα που έχει ο $j-$παράγοντας στην $i-$μεταβλητή και χρησιμοποιούνται συνήθως για την ερμηνεία των F_j. Τα πιο υψηλά φορτία σε απόλυτη τιμή για τον F_j παράγοντα με $j = 1, 2, \ldots, m$, συνδέουν τον παράγοντα αυτόν με τις μεταβλητές X_i, $i = 1, 2, \ldots, p$ και με αυτόν τον τρόπο εκτιμάται ότι θα εξηγηθούν οι συσχετίσεις των μεταβλητών αντιστοιχίζοντας τους παράγοντες σε αυτές.

Είναι όμως σημαντικό για το παραγοντικό μοντέλο να γίνουν ορισμένες υποθέσεις για την αναγνωρισιμότητα του. Υποθέτουμε λοιπόν ότι οι κοινοί παράγοντες F_j έχουν μέσο μηδέν και διακύμανση ένα και ότι είναι αμοιβαία ασυσχέτιστοι, $\operatorname{cov}(F_j, F_k) = 0$, $j \neq k$. Για το σφάλμα ε υποθέτουμε ότι έχει μέσο μηδέν και διακύμανση ψ_i^2 και ότι είναι επίσης αμοιβαία ασυσχέτιστο, $\operatorname{cov}(\varepsilon_i, \varepsilon_k) = 0$. Επιπλέον κάνουμε την υπόθεση ότι κάθε κοινός παράγοντας δεν σχετίζεται με το σφάλμα, $\operatorname{cov}(F_j, \varepsilon_i) = 0$, $\forall i, j$. Οι υποθέσεις αυτές μας οδηγούν σε μια απλούστερη μορφή για τη διακύμανση των X_i την οποία μπορούμε να γράψουμε ως $\operatorname{var}(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2 + \psi_i$ και η οποία έχει ιδιαίτερη σημασία. Παράλληλα, με την

υπόθεση $\text{cov}(\varepsilon_i, \varepsilon_k) = 0$, σημαίνει ότι μόνο οι παράγοντες και μόνο αυτοί, εξηγούν τις συσχετίσεις που υπάρχουν στα δεδομένα.

Επίσης παρατηρούμε για τη διακύμανση των X_i ότι μπορεί να χωριστεί σε δύο μέρη, στην ποσότητα $\ell_{i1}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2$ που ονομάζεται εταιρικότητα (communality) και συμβολίζεται με h_i^2 και στο ψ_i που ονομάζεται ιδιαιτερότητα (uniqueness, specificity). Άρα μπορούμε να γράψουμε την διακύμανση και ως εξής

$$\sigma_{ii} = \sigma_i^2 = \text{var}\left(X_i\right) = h_i^2 + \psi_i \tag{1.3}$$

Η σχέση (1.3) μας δείχνει ότι τη διακύμανση και κατά συνέπεια τη συνδιακύμανση των μεταβλητών του πίνακα $\mathbf{X}(n \times p)$ μπορούμε να την εκφράσουμε με τους όρους των φορτίων και των ιδιαιτεροτήτων. Αυτό έχει σαν αποτέλεσμα ο θεωρητικός πίνακας συνδιακύμανσης $\mathbf{\Sigma}(p \times p)$ να εκφράζεται μέσω των $\mathbf{L}(p \times m)$ και $\text{diag}(\mathbf{\Psi})$. Από την (1.2) και αφού το $\boldsymbol{\mu}_{p\times 1}$ δεν επηρεάζει τη διακύμανση, για τη συνδιακύμανση των μεταβλητών θα ισχύει ότι

$$\mathbf{\Sigma} = \text{cov}\left(\mathbf{X}\right) = \text{cov}\left(\mathbf{LF} + \boldsymbol{\varepsilon}\right)$$

και επειδή οι $\mathbf{LF}, \boldsymbol{\varepsilon}$ είναι ασυσχέτιστοι τότε θα είναι

$$\begin{aligned} \mathbf{\Sigma} &= \text{cov}\left(\mathbf{LF}\right) + \text{cov}\left(\boldsymbol{\varepsilon}\right) \\ &= \mathbf{L}\text{cov}\left(\mathbf{F}\right)\mathbf{L}' + \mathbf{\Psi} \end{aligned}$$

Όμως από τις αρχικές υποθέσεις είναι $\text{cov}\left(\mathbf{F}\right) = \mathbf{I}$ άρα

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi} \tag{1.4}$$

με τον πίνακα $\mathbf{\Psi}$ να είναι διαγώνιος. Συνεπώς για να μπορέσουμε να ερμηνεύσουμε τις συνδιακυμάνσεις των μεταβλητών κάτω από το παραγοντικό μοντέλο (θεωρούμε ότι μπορεί να γραφούν όπως στην 1.4), θα πρέπει να εκτιμήσουμε τον πίνακα των φορτίων και των ιδιαιτεροτήτων καθότι ο θεωρητικός πίνακας συνδιακύμανσης στην (1.4) διαμερίζεται σε δύο μέρη, στο κομμάτι εκείνο που ερμηνεύουν οι κοινοί παράγοντες δηλαδή την εταιρικότητα και στο κομμάτι που οφείλεται στους μοναδικούς παράγοντες την ιδιαιτερότητα, η οποία δεν ερμηνεύεται από το μοντέλο.

2.2.1 Μέθοδοι Εκτίμησης του Παραγοντικού Μοντέλου

Υπάρχουν διάφορες μέθοδοι για να εκτιμήσουμε τους συντελεστές του παραγοντικού μοντέλου αλλά οι περισσότερο διαδεδομένες είναι η μέθοδος των κυρίων συνιστωσών και η

μέθοδος της μέγιστης πιθανοφάνειας. Πρώτα θα αναφερθούμε στην μέθοδο των κυρίων συνιστωσών.

Στην πράξη τον θεωρητικό πίνακα συνδιακύμανσης στην (1.4) τον αντικαθιστούμε με τον δειγματικό πίνακα συνδιακύμανσης $\mathbf{S}_{p\times p}$ και εκτιμούμε τα φορτία και την εταιρικότητα έτσι ώστε να ισχύει ότι

$$\mathbf{S} \cong \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}} \tag{1.5}$$

Καθώς ο πίνακας συνδιακύμανσης είναι συμμετρικός, χρησιμοποιούμε το θεώρημα της φασματικής ανάλυσης (spectral decomposition theorem) για να τον παραγοντοποιήσουμε και να βρούμε την προσέγγιση του. Με τη μέθοδο των κυρίων συνιστωσών αγνοούμε τον $\mathbf{\Psi}$, άρα θα πάρουμε την προσέγγιση του $\mathbf{S}_{p\times p}$ μόνο από το $\hat{\mathbf{L}}\hat{\mathbf{L}}'$. Συνεπώς για τον $\mathbf{S}$ θα έχουμε ότι

$$\mathbf{S} = \mathbf{C}\mathbf{D}\mathbf{C}' \tag{1.5.α}$$

όπου στην (1.5.α) ο $\mathbf{C}$ είναι ένας ορθογώνιος πίνακας με κανονικοποιημένα ιδιοδιανύσματα $\mathbf{c}: \left\{c_i^T c_j = 0, \|c_i\| = 1\right\}$ και ο $\mathbf{D}$ είναι ο διαγώνιος πίνακας με τις ιδιοτιμές $\theta_1, \theta_2, \ldots, \theta_p$ του $\mathbf{S}$. Επιπλέον αφού ο $\mathbf{S}$ είναι θετικά ημιορισμένος $\mathbf{S} \geq 0$, μπορούμε να γράψουμε τον $\text{diag}(\mathbf{D})$

$$\mathbf{D} = \mathbf{D}^{1/2}\mathbf{D}^{1/2}$$

όπου $\mathbf{D}^{1/2} = \text{diag}\left(\sqrt{\theta_1}, \sqrt{\theta_2}, \ldots, \sqrt{\theta_p}\right)$, έτσι ώστε η παραγοντοποίηση του $\mathbf{S}$ να γίνει

$$\mathbf{S} = \mathbf{C}\mathbf{D}\mathbf{C}' = \mathbf{C}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{C}' = \left(\mathbf{C}\mathbf{D}^{1/2}\right)\left(\mathbf{C}\mathbf{D}^{1/2}\right)' \tag{1.6}$$

Η σχέση (1.6) είναι της μορφής $\mathbf{S} = \hat{\mathbf{L}}\hat{\mathbf{L}}'$, αλλά καθώς ο $\mathbf{C}\mathbf{D}^{1/2}$ είναι $p \times p$ πίνακας και ο $\hat{\mathbf{L}}$ είναι ένας $p \times m$ πίνακας θέτουμε, $\mathbf{D}_1 = \text{diag}\left(\theta_1, \theta_2, \ldots, \theta_m\right)$ με $m \ll p$, $\theta_1 > \theta_2 > \ldots > \theta_m$ και $\mathbf{C}_1 = \left(c_1, c_2, \ldots, c_m\right)$. Έτσι μετά την τροποποίηση αυτή θα πάρουμε

$$\hat{\mathbf{L}}_1\left(p \times m\right) = \mathbf{C}_1\mathbf{D}_1^{1/2} = \left(\sqrt{\theta_1}c_1, \sqrt{\theta_2}c_2, \ldots, \sqrt{\theta_m}c_m\right) \tag{1.7}$$

Η (1.7) σε πίνακες γράφεται στην μορφή

$$\hat{\mathbf{L}}_{1(p\times m)} = \begin{pmatrix} \hat{\ell}_{11} & \hat{\ell}_{12}\cdots & \hat{\ell}_{1m} \\ \vdots & \vdots & \vdots \\ \hat{\ell}_{p1} & \hat{\ell}_{p2}\cdots & \hat{\ell}_{pm} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12}\cdots & c_{1m} \\ \vdots & \vdots & \vdots \\ c_{p1} & c_{p2}\cdots & c_{pm} \end{pmatrix} \begin{pmatrix} \sqrt{\theta_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\theta_m} \end{pmatrix}.$$

Ωστόσο χρειάζεται να υπολογίσουμε και τα $\hat{\psi}_i$ ειδικά στην περίπτωση που $m \ll p$. Επειδή για τον πίνακα $\hat{\mathbf{L}}_1\hat{\mathbf{L}}_1'\,(p\times p)$ το άθροισμα της διαγωνίου του ισούται με $\sum_{j=1}^{m}\hat{\ell}_{ij}^2$, ορίζουμε ότι $\hat{\psi}_i = s_i^2 - \sum_{j=1}^{m}\hat{\ell}_{ij}^2$ και πλέον έχουμε την προσέγγιση του $\mathbf{S}$ στην (1.5).

Οι εκτιμημένες εταιρικότητες $\tilde{h}_i^2$, δείχνουν το ποσοστό της διακύμανσης της κάθε μεταβλητής που εξηγείται από τους παράγοντες που βρήκαμε και όσο μεγαλύτερη είναι η τιμή αυτή τόσο καλύτερα ερμηνεύεται η μεταβλητή X_i από το παραγοντικό μοντέλο. Η ποσότητα $\sum\tilde{h}_i^2 \big/ \mathrm{tr}(\mathbf{S})$ δείχνει το συνολικό ποσοστό της δομής που ερμηνεύεται από τους m παράγοντες. Στην πράξη αν κάποιες μεταβλητές έχουν χαμηλά $\tilde{h}_i^2$, για παράδειγμα < 0.30, τότε το παραγοντικό μοντέλο που προσαρμόσαμε δεν καταφέρνει να εξηγήσει μεγάλο μέρος της διακύμανσης των μεταβλητών αυτών και πιθανώς να χρειάζεται να ερευνήσουμε για επιπλέον παράγοντες ώστε να αυξήσουμε την προσαρμοστικότητα του μοντέλου στο σύνολο των μεταβλητών. Αν πάλι χρησιμοποιούμε τον δειγματικό πίνακα συσχετίσεων $\mathbf{R}\,(p\times p)$ για να ερμηνεύσουμε τις συσχετίσεις των μεταβλητών, αντίστοιχα, η ποσότητα $\sum\tilde{h}_i^2 \big/ p$ μας δείχνει το συνολικό ποσοστό της διακύμανσης που ερμηνεύεται από τους m παράγοντες.

Οι εκτιμήσεις των ιδιαιτεροτήτων $\hat{\psi}_i$ υπολογίζονται από τη σχέση $\hat{\psi}_i = 1 - \tilde{h}_i^2$ και όπως είπαμε είναι το κομμάτι εκείνο της διακύμανσης της κάθε μεταβλητής που δεν μπορεί να εξηγηθεί από το παραγοντικό μοντέλο. Ιδανικά θέλουμε να ισχύει $\tilde{h}_i^2 \approx 1$ και $\hat{\psi}_i \approx 0$. Η ποσότητα $\mathbf{E} = \mathbf{S} - \left(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}}\right)$ αναφέρεται ως ο εκτιμημένος πίνακας υπολοίπων και για την μέθοδο των κυρίων συνιστωσών αποτελεί εσωτερικό κριτήριο της επιτυχίας της. Όταν τα στοιχεία του πίνακα $\mathbf{E}$ είναι μικρά σε απόλυτη τιμή τότε η προσαρμογή θεωρείται καλή.

Στην περίπτωση που για την εκτίμηση του παραγοντικού μοντέλου χρησιμοποιηθεί η μέθοδος της μέγιστης πιθανοφάνειας, τότε αρχικά υποθέτουμε ότι τα σφάλματα ακολουθούν την πολυμεταβλητή κανονική κατανομή και το ίδιο υποθέτουμε για τους παράγοντες. Έτσι προκύπτει ότι και το διάνυσμα των μεταβλητών $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ προέρχεται από την πολυμεταβλητή κανονική κατανομή. Αν λοιπόν $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ο λογάριθμος της συνάρτησης πιθανοφάνειας (log-likelihood) του τυχαίου δείγματος και των παραμέτρων είναι:

$$\log L\left(\vec{X};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)=\sigma\tau\alpha\theta\varepsilon\rho\acute{\alpha}-\frac{n}{2}\log\left|2\pi\boldsymbol{\Sigma}\right|-\frac{n}{2}tr\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)-\frac{n}{2}\left(\bar{X}-\boldsymbol{\mu}\right)\boldsymbol{\Sigma}^{-1}\left(\bar{X}-\boldsymbol{\mu}\right)^{'} \quad (1.8)$$

Αν αντικαταστήσουμε τον αμερόληπτο εκτιμητή του $\hat{\boldsymbol{\mu}}=\bar{X}$, η (1.8) είναι η log-likelihood του $\boldsymbol{\Sigma}$ μόνο:

$$\log L\left(\vec{X};\boldsymbol{\Sigma}\right)=-\frac{n}{2}\left[\log\left|2\pi\boldsymbol{\Sigma}\right|+\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\right] \quad (1.9)$$

όπου στην (1.9) αντικαταστούμε το $\boldsymbol{\Sigma}$ με $\boldsymbol{\Sigma}=\mathbf{L}\mathbf{L}^{\mathrm{T}}+\boldsymbol{\Psi}$.

Οι εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\mathbf{L}}$ και $\hat{\boldsymbol{\Psi}}$ θα προκύψουν όταν μεγιστοποιήσουμε την συνάρτηση (1.9) ως προς $\mathbf{L}$ και $\boldsymbol{\Psi}$ με τον περιορισμό ότι ο $\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L}$ είναι διαγώνιος. Ωστόσο η μεγιστοποίηση της παραπάνω συνάρτησης δεν έχει αναλυτική λύση και πρέπει να λυθεί επαναληπτικά χρησιμοποιώντας κάποιον αλγόριθμο (για λεπτομέρειες βλ. Harman, 1976-203:204 και Knafl and Grey, 2007). Επιπλέον, δεν είναι βέβαιο ότι ο επαναληπτικός αλγόριθμος τελικά θα συγκλίνει με αποτέλεσμα να βρεθούν λύσεις με αρνητικά στοιχεία στον πίνακα $\hat{\boldsymbol{\Psi}}$ ή να εκτιμηθούν οι εταιρικότητες των μεταβλητών με τιμές ίσες ή μεγαλύτερες της μονάδας. Το τελευταίο συνήθως ονομάζεται *Heywood case*.

Ένα σημαντικό πλεονέκτημα της συγκεκριμένης μεθόδου εκτίμησης έναντι εκείνης των κυρίων συνιστωσών είναι το γεγονός ότι μπορούμε να κάνουμε ελέγχους καλής προσαρμογής της καταλληλότητας του ορθογώνιου παραγοντικού μοντέλου. Δοθέντος της εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\mathbf{L}}$ και $\hat{\boldsymbol{\Psi}}$, ο έλεγχος συγκρίνει το $\hat{\boldsymbol{\Sigma}}$ με το $\frac{(n-1)}{n}\mathbf{S}$ που είναι ο εκτιμητής μέγιστης πιθανοφάνειας του $\boldsymbol{\Sigma}$ όταν δεν υπάρχουν περιορισμοί στο πίνακα (δηλ. η H_{α} – εναλλακτική υπόθεση). Η ελεγχοσυνάρτηση με τη διόρθωση του *Bartlett* δίνεται από τη σχέση

$$\left[n-\left(\frac{2p+11}{6}\right)-\frac{2m}{3}\right]\ln\left|\frac{n\hat{\boldsymbol{\Sigma}}}{(n-1)\mathbf{S}}\right|$$

και η οποία ασυμπτωτικά έχει X_{ν}^{2} κατανομή όταν η μηδενική (H_0) υπόθεση αληθεύει με $\nu=\frac{1}{2}\left[(p-m)^2-(p+m)\right]$ βαθμοί ελευθερίας. Μια συνήθης προσέγγιση με τη συγκεκριμένη μέθοδο εκτίμησης είναι να ξεκινάμε το παραγοντικό μοντέλο με τη λύση $m=1$ και σταδιακά να αυξάνουμε τους παράγοντες έως ότου ο έλεγχος καλής προσαρμογής γίνει αποδεκτός ή μέχρις ότου οι βαθμοί ελευθερίας γίνουν αρνητικοί.

2.3 Έλεγχος των Συσχετίσεων - Αριθμός των Παραγόντων

Σε πραγματικά δεδομένα εκείνο που ο αναλυτής πρέπει πρώτα να κάνει είναι να αποφασίσει με ποιον πίνακα θα δουλέψει. Ανάλογα με τη φύση των δεδομένων, θα πρέπει διαλέξει ανάμεσα στον πίνακα συνδιακυμάνσεων **S** ή στον πίνακα συσχετίσεων **R** καθώς και ποια μέθοδο εκτίμησης θα χρησιμοποιήσει για τους παράγοντες. Για παράδειγμα, με τη μέθοδο της μέγιστης πιθανοφάνειας η οποία είναι ανεξάρτητη κλίμακας (scale invariant), δεν παίζει ρόλο ποιο πίνακα θα χρησιμοποιήσουμε για να εκτιμήσουμε το μοντέλο, αλλά με μεθόδους που δεν είναι scale invariant όπως συμβαίνει με τη μέθοδο των κυρίων συνιστωσών τα αποτελέσματα που θα πάρουμε αν βασιστούμε στον πίνακα συνδιακυμάνσεων θα είναι πολύ διαφορετικά από τον πίνακα συσχετίσεων.

Αν τελικά δουλέψουμε με τον πίνακα συσχετίσεων **R** είναι σημαντικό να υπάρχουν συσχετίσεις αρκετά μεγάλες, γιατί αυτές τις συσχετίσεις θα προσπαθήσουμε να εξηγήσουμε. Θεωρητικά συσχετίσεις με τιμές $>|0.40|$ είναι ευπρόσδεκτες, σε αντίθετη περίπτωση δεν έχει έννοια να συνεχίσουμε αφού δεν θα μπορέσουμε να βρούμε κοινούς παράγοντες. Αν πάλι υπάρχουν μεταβλητές που είναι ασυσχέτιστες με τις υπόλοιπες, καλό είναι να τις αγνοήσουμε διότι θα προκύψουν από μόνες τους ως ένας ξεχωριστός παράγοντας. Κάποιοι συγγραφείς έχουν προτείνει ότι ο $\mathbf{R}^{-1}$ θα πρέπει να είναι σχεδόν διαγώνιος πίνακας ώστε να βρεθεί ικανοποιητικό παραγοντικό μοντέλο. Για να εκτιμηθεί πόσο κοντά είναι ο $\mathbf{R}^{-1}$ σε διαγώνιο πίνακα, ο Kaiser (1970) πρότεινε ένα μέτρο δειγματικής καταλληλότητας (MSA) το οποίο για την i – μεταβλητή υπολογίζεται ως

$$\mathrm{MSA_i} = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} q_{ij}^2}$$

όπου r_{ij}^2 είναι η ρίζα ενός στοιχείου του πίνακα **R** και q_{ij}^2 είναι η ρίζα ενός στοιχείου από τον πίνακα $\mathbf{Q} = \mathbf{D}\mathbf{R}^{-1}\mathbf{D}$, με $\mathbf{D} = \left[\left(\mathrm{diag}\mathbf{R}^{-1}\right)^{1/2}\right]^{-1}$. Καθώς το $\mathrm{MSA_i}$ πλησιάζει την τιμή 1, αυτό είναι ένδειξη της καταλληλότητας της μεταβλητής ενώ στην πράξη θα πρέπει να ξεπερνάει το 0.8 για να υπάρξουν ικανοποιητικά αποτελέσματα.

Ένα ακόμη μέτρο που είναι πολύ σημαντικό για τον έλεγχο των δεδομένων πριν την παραγοντική ανάλυση είναι η τιμή του στατιστικού KMO (Kaiser-Meyer-Olkin)

$$\text{KMO} = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} a_{ij}^2}$$

όπου r_{ij}^2 και a_{ij}^2 είναι οι συντελεστές δειγματικής και μερικής συσχέτισης αντίστοιχα. Το KMO συγκρίνει το σχετικό μέγεθος των συντελεστών συσχέτισης σχετικά με τους μερικούς συντελεστές συσχέτισης και αν η τιμή του είναι μεγάλη τότε τα δεδομένα είναι κατάλληλα (Καρλής, 2005). Έχει υπολογιστεί ότι τιμές μικρότερες του 0.5 είναι πολύ κακές ενώ πάνω από 0.8 θεωρούνται καλές για να προχωρήσουμε με την ανάλυση.

Αφού λοιπόν γίνουν οι απαραίτητοι έλεγχοι και προχωρήσουμε σε παραγοντική ανάλυση θα πρέπει να αποφασίσουμε για τον αριθμό των παραγόντων m που θα κρατήσουμε. Κάποια κριτήρια για το σκοπό αυτό είναι τα εξής:

- Επιλέγουμε τόσους παράγοντες m ώστε να επιτύχουμε κάποιο επιθυμητό ποσοστό της συνολικής διακύμανσης
- Επιλέγουμε παράγοντες m ίσους με τον αριθμό των ιδιοτιμών της διάσπασης του **S** ή του **R** που έχουν τιμή μεγαλύτερη της μονάδας
- Από το scree plot των ιδιοτιμών αποφασίζουμε για τον αριθμό m, από τη στιγμή που το γράφημα αρχίζει να αλλάζει κλίση
- Κάνουμε ελέγχους για τον αριθμό των παραγόντων. Για τη μέθοδο εκτίμησης της μέγιστης πιθανοφάνειας, μέχρι το κριτήριο καλής προσαρμογής γίνει αποδεκτό ή οι βαθμοί ελευθερίας γίνουν αρνητικοί. Για τη μέθοδο των κυρίων συνιστωσών, ελέγχουμε τον εκτιμημένο πίνακα (reproduced matrix) και κοιτάμε τις αποκλίσεις του από τον πραγματικό να είναι μικρές

Μια σημαντική διαφορά ανάμεσα στις δύο μεθόδους εκτίμησης είναι το γεγονός ότι με τη μέθοδο των κυρίων συνιστωσών από την στιγμή που από m παράγοντες πάμε σε $m+1$ στο παραγοντικό μοντέλο, οι τιμές των φορτίων των προηγούμενων παραγόντων δεν αλλάζουν άρα δεν αλλάζει και η ερμηνεία τους. Κάτι αντίστοιχο όμως δεν ισχύει με τη μέθοδο της μέγιστης πιθανοφάνειας.

2.4 Περιστροφή των Παραγόντων

Με την περιστροφή των παραγόντων προσπαθούμε να κάνουμε τους παράγοντες πιο ερμηνεύσιμους. Η περιστροφή είναι ανεξάρτητη της μεθόδου εκτίμησης και βοηθάει τον

αναλυτή στο να ξεχωρίσει τους παράγοντες, δηλαδή να μπορέσει πιο εύκολα να τους δώσει φυσική ερμηνεία αφού ο σκοπός της είναι να διαχωριστούν οι μεταβλητές που εξηγούνται από τους παράγοντες.

Με την περιστροφή δεν αλλάζουν κάποια από τα χαρακτηριστικά του μοντέλου όπως η προσαρμοστικότητα και το συνολικό ποσοστό της διακύμανσης που ερμηνεύεται από το μοντέλο, ο εκτιμημένος πίνακας συσχετίσεων ή συνδιακυμάνσεων και οι εταιρικότητες, παρά μόνο οι τιμές των φορτίων και το ποσοστό της διακύμανσης που εξηγείται από κάθε παράγοντα.

Γενικά, αν $\mathbf{L}(p \times m)$ είναι ο πίνακας με τα φορτία των παραγόντων και $\mathbf{T}$ είναι ένας ορθογώνιος πίνακας έτσι ώστε $\mathbf{TT}' = \mathbf{I} = \mathbf{T}'\mathbf{T}$, τότε ισχύει ότι

$$\hat{\mathbf{L}}\mathbf{T}\left(\hat{\mathbf{L}}\mathbf{T}\right)' = \hat{\mathbf{L}}\mathbf{T}\mathbf{T}'\hat{\mathbf{L}} = \hat{\mathbf{L}}\hat{\mathbf{L}}'$$

επομένως ο $\mathbf{T}$ ορίζει έναν ορθογώνιο μετασχηματισμό και ο πίνακας $\hat{\mathbf{L}}\mathbf{T}$ μπορεί να θεωρηθεί ως ο νέος πίνακας των φορτίων. Αν συμβολίσουμε με $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}'$ τα φορτία μετά την περιστροφή, τότε παίρνουμε την ίδια εκτίμηση του πίνακα συνδιακύμανσης $\mathbf{S}$ αφού

$$\mathbf{S} \cong \hat{\mathbf{L}}^*\hat{\mathbf{L}}^* + \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}\mathbf{T}\mathbf{T}'\hat{\mathbf{L}} + \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}}.$$

Από γεωμετρικής άποψης τα φορτία της $i-$γραμμής του πίνακα $\hat{\mathbf{L}}$ είναι οι συντεταγμένες για την μεταβλητή X_i, $i = 1,\ldots,p$. Σύμφωνα με τον Harman (1976), με την περιστροφή στοχεύεται να επιτευχθεί αυτό που ονομάζει ως *simple structure*. Αυτό σημαίνει ότι οι μεταβλητές που έχουν υψηλά φορτία σε ένα παράγοντα θα πρέπει να έχουν μικρή ή σχεδόν μηδενική συνεισφορά στους υπόλοιπους παράγοντες. Σχηματικά αυτό μπορεί να δειχτεί ως εξής για τρεις παράγοντες και πέντε μεταβλητές:

	F_{i1}	F_{i2}	F_{i3}
X_1	+	-	-
X_2	+	-	-
X_3	-	+	-
X_4	-	-	+
X_5	-	+	-

Κάτι τέτοιο στην πράξη δεν είναι βέβαιο ότι θα συμβεί, αλλά όταν αυτό επιτευχθεί τότε η ερμηνεία των παραγόντων γίνεται πολύ πιο εύκολα.

Υπάρχουν στη βιβλιογραφία διάφοροι μετασχηματισμοί για να περιστρέψουμε τους παράγοντες. Οι περισσότερο διαδεδομένοι ορθογώνιοι μετασχηματισμοί είναι η *varimax* περιστροφή όπως αυτή προτάθηκε από τον Kaiser (1985) και η οποία προσπαθεί να ελαχιστοποιήσει τον αριθμό των μεταβλητών που έχουν μεγάλα φορτία σε κάθε παράγοντα. Δηλαδή προσπαθεί να κάνει κάτι ανάλογο του *simple structure*. Στην πιο απλή μορφή όπου $m = 2$ παράγοντες, ο πίνακας μετασχηματισμού $\mathbf{T}$ για δοθείσα γωνία φ υπολογίζεται από τον πίνακα

$$T_{2\times 2} = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

και οι νέες τιμές των φορτίων μετά την περιστροφή από την σχέση $\hat{\mathbf{L}}^{*}_{p\times 2} = \hat{\mathbf{L}}\mathbf{T}$. Η δεύτερη πιο διαδεδομένη μέθοδος ορθογώνιας περιστροφής των αξόνων είναι η *quartimax* και η οποία προσπαθεί να ελαχιστοποιήσει τον αριθμό των παραγόντων που εξηγούν μια μεταβλητή. Συνήθως με την *quartimax* οι μεταβλητές μαζεύονται στους δύο πρώτους παράγοντες αν $m \geq 2$.

Υπάρχουν όμως περιπτώσεις που η ορθογώνια περιστροφή δεν ξεχωρίζει καλά τους παράγοντες και χρειάζεται να προσαρμόσουμε μη-ορθογώνια περιστροφή διότι τυγχάνει οι παράγοντες να είναι συσχετισμένοι μεταξύ τους. Για παράδειγμα, ο παράγοντας που εκφράζει το επαγγελματικό *status* του παιδιού θα πρέπει, φυσιολογικά, να σχετίζεται με εκείνο του πατέρα του αντί οι δυο παράγοντες να είναι ανεξάρτητοι (ασυσχέτιστοι). Κάποια επιπλέον κριτήρια μη-ορθογώνιων (oblique) περιστροφών που χρησιμοποιούνται συχνά από τους αναλυτές είναι το *oblimin* και το *promax* και για μια αναλυτικότερη παρουσίαση τους, προτείνεται στον αναγνώστη να κοιτάξει στον Harman (1976) και στον Jobson (1992).

2.5 Σκορ Παραγόντων

Έχουμε ήδη αναφέρει ότι ένας από τους σκοπούς που γίνεται παραγοντική ανάλυση είναι η μείωση της αρχικής διάστασης των δεδομένων (δηλαδή των μεταβλητών). Για να γίνει αυτό εκφράζουμε τους παράγοντες σαν γραμμική συνάρτηση των αρχικών μεταβλητών έτσι ώστε αντί να δουλεύουμε με τις p μεταβλητές του πίνακα $\mathbf{X}_{n\times p}$, να δουλεύουμε με τους m παράγοντες ($m \ll p$) στους οποίους έχουμε δώσει μια φυσική ερμηνεία και μπορούν να θεωρηθούν σαν «καινούργιες» μεταβλητές. Για να μπορέσουμε όμως να χρησιμοποιήσουμε τους «κατασκευασμένους» παράγοντες είτε σε περαιτέρω αναλύσεις, όπως στην MANOVA ή

στην πολλαπλή παλινδρόμηση είτε για διαγνωστικούς σκοπούς σχετικά με τη συμπεριφορά των παρατηρήσεων σε αυτούς είτε για τη δημιουργία δεικτών, πρέπει να δημιουργήσουμε τιμές για τις n γραμμές (παρατηρήσεις) του αρχικού πίνακα $\mathbf{X}_{n\times p}$. Οι τιμές αυτές ονομάζονται σκορ των παραγόντων.

Έχοντας εκτιμήσει ένα παραγοντικό μοντέλο με $\hat{\mathbf{L}}$, $\hat{\mathbf{\Psi}}$, $\hat{\mathbf{L}}^*$ και $\hat{\mathbf{\Psi}}^*$ αντίστοιχα, να είναι οι εκτιμήσεις των παραμέτρων του πριν και μετά την περιστροφή, εκτιμούμε τα σκορ των παραγόντων, $\hat{\mathbf{f}} = \left(\hat{f}_{i1}, \hat{f}_{i2}, \ldots, \hat{f}_{im}\right)'$ για $i = 1, \ldots, n$. Υπάρχουν διάφορες μέθοδοι εκτίμησης των παραγοντικών σκορ όπως, η *regression* μέθοδος, η μέθοδος *Bartlett* και η μέθοδος του *Anderson*. Εδώ θα αναφερθούμε στην μέθοδο *regression* ή αλλιώς στις *Thomson* εκτιμήσεις. Για τις υπόλοιπες μεθόδους παραπέμπουμε στον Harman (1976, §16). Ωστόσο αναφέρουμε πως και οι τρεις μέθοδοι δίνουν παράγοντες με μέση τιμή μηδέν $\mathrm{E}\left(f_i\right) = 0$.

Regression μέθοδος

Κάθε παράγοντας f_j γράφεται σαν γραμμικός συνδυασμός των μεταβλητών X_i με κεντροποίηση στην μορφή

$$\begin{aligned} f_1 &= \beta_{11}\left(X_1 - \bar{X}_1\right) + \beta_{12}\left(X_2 - \bar{X}_2\right) + \ldots + \beta_{1p}\left(X_p - \bar{X}_p\right) + \xi_1 \\ &\vdots \\ f_m &= \beta_{m1}\left(X_1 - \bar{X}_1\right) + \beta_{m2}\left(X_2 - \bar{X}_2\right) + \ldots + \beta_{mp}\left(X_p - \bar{X}_p\right) + \xi_m \end{aligned} \tag{1.10}$$

Η (1.10) σε μορφή πινάκων γράφεται ως

$$\mathbf{F} = \mathbf{B}_1'\left(\mathbf{X} - \bar{\mathbf{X}}\right) + \boldsymbol{\xi} \tag{1.11}$$

Ο πίνακας των συντελεστών που χρειάζεται για να πάρουμε τις εκτιμήσεις των παραγόντων είναι ο $\mathbf{B}_1$ και λέγεται factor score coefficient matrix. Αποδεικνύεται ότι για να βρούμε τον πίνακα των συντελεστών των σκορ πρέπει να υπολογίσουμε τις ποσότητες $\hat{\mathbf{B}}_1 = \mathbf{S}^{-1}\hat{\mathbf{L}}$ ή $\hat{\mathbf{B}}_1 = \mathbf{R}^{-1}\hat{\mathbf{L}}$ αντίστοιχα. Οι εκτιμήσεις ή σκορ των παραγόντων δίνονται με αντικατάσταση για το $\hat{\mathbf{B}}_1$ στην (1.11), $\hat{\mathbf{F}} = \hat{\mathbf{B}}_1'\left(\mathbf{X} - \bar{\mathbf{X}}\right) + \boldsymbol{\xi}$, δηλαδή από τις σχέσεις

$$\hat{\mathbf{F}} = \mathbf{Y}_c\mathbf{S}^{-1}\hat{\mathbf{L}} \text{ ή } \hat{\mathbf{F}} = \mathbf{Y}_S\mathbf{R}^{-1}\hat{\mathbf{L}} \tag{1.12}$$

όπου στην (1.12) $\mathbf{Y}_S$ είναι ο πίνακας των τυποποιημένων μεταβλητών $\left(x_{ij} - \bar{x}_j\right)/s_j$. Συνήθως ζητάμε να υπολογίσουμε τα σκορ για τους παράγοντες που βρήκαμε μετά την περιστροφή, οπότε στην περίπτωση αυτή στις παραπάνω σχέσεις χρησιμοποιούμε αντί του $\hat{\mathbf{L}}$ τον $\hat{\mathbf{L}}^*$.

2.6 Μη – ορθογώνια Παραγοντική Ανάλυση

Το ορθογώνιο παραγοντικό μοντέλο βασίστηκε στην υπόθεση ότι οι παράγοντες είναι ασυσχέτιστοι μεταξύ τους. Όμως σε αρκετές περιπτώσεις η υπόθεση αυτή δεν είναι ρεαλιστική και οι παράγοντες θα πρέπει να συσχετίζονται. Δηλαδή αντί της υπόθεσης $\mathrm{Cov}(F) = \mathbf{I}$, τώρα υποθέτουμε ότι $\mathrm{Cov}(F) = \Omega$ άρα $\mathbf{\Sigma} = \mathbf{L\Omega L}^{\mathrm{T}} + \mathbf{\Psi}$ και επιπλέον στη σχέση θα πρέπει να εκτιμήσουμε τον πίνακα $\mathbf{\Omega}$. Στην πράξη συνήθως αυτό που γίνεται όταν θέλουμε οι παράγοντες να είναι συσχετισμένοι, απλώς εφαρμόζουμε μια μη-ορθογώνια περιστροφή των αξόνων, π.χ. *oblimin* μέθοδο και οδηγούμαστε σε παράγοντες που είναι συσχετισμένοι μεταξύ τους.

2.7 Σχόλια και Συμπεράσματα

Η παραγοντική ανάλυση θεωρείται *model-based* μέθοδος. Σε σχέση με άλλες τεχνικές, προσπαθεί περισσότερο να ερμηνεύσει τη δομή παρά την μεταβλητότητα. Από τον τρόπο ορισμού του μοντέλου χρησιμοποιείται κυρίως για συνεχή δεδομένα ενώ το ίδιο υποθέτουμε και για τους παράγοντες που κατασκευάζουμε. Όταν στα δεδομένα υπάρχουν κατηγορικές μεταβλητές με δύο ή περισσότερες κατηγορίες, όπως για παράδειγμα όταν οι μεταβλητές αφορούν ερωτήσεις στην κλίμακα *Likert*, τότε η παραγοντική ανάλυση θα πρέπει να γίνεται με προσοχή. Σύμφωνα με τους Bishop (1977), Bartholomew et al. (2002), στις περιπτώσεις που οι μεταβλητές είναι κατηγορικές με μεγάλο αριθμό κατηγοριών, περισσότερες από 5 ή 6 τότε μπορούμε να θεωρήσουμε τα δεδομένα σε συνεχή κλίμακα και να προχωρήσουμε στην εύρεση μοντέλου. Τονίζουν όμως ότι ενώ το ίδιο ισχύει και για κατηγορικές μεταβλητές με λιγότερες κατηγορίες, οι εκτιμήσεις των φορτίων μάλλον θα είναι μεροληπτικές.

Στις περιπτώσεις που έχουμε κατηγορικά δεδομένα και θέλουμε να προχωρήσουμε σε παραγοντική ανάλυση, για τον υπολογισμό των συσχετίσεων χρησιμοποιούνται άλλοι συντελεστές όπως του Kruskal ο gamma, του Somer ο d, ο συντελεστής του Kendall tau ή του Spearman ο r_S, αντί του συντελεστή r_p του Pearson εκτός και αν οι συσχετίσεις που υπολογίζονται από τους συντελεστές αυτούς είναι πολύ κοντά $(\cong 1.00)$ με τον r_p.

Τέλος, μπορεί να γίνει παραγοντική ανάλυση έχοντας μόνο τον πίνακα συνδιακυμάνσεων ή συσχετίσεων των μεταβλητών και όχι τα πλήρη δεδομένα. Τότε είναι ξεκάθαρο ότι τα σκορ των παραγόντων δεν μπορούν να υπολογιστούν.

2.8 Εφαρμογή της Μεθόδου

Σε έρευνα που πραγματοποιήθηκε για λογαριασμό της Τράπεζας *ABC*, οι εργαζόμενοι και τα στελέχη απάντησαν σε ένα ερωτηματολόγιο που είχε ως στόχο να ερευνήσει την αποτελεσματικότητα του υπάρχοντος συστήματος αξιολόγησης ώστε μελλοντικά να υπάρξει βελτίωση του. Το ερωτηματολόγιο διατίθεται στο Παράρτημα Α στο τέλος της εργασίας. Οι ερωτήσεις εμφανίζονται σε πενταβάθμια κλίμακα *Likert* με τις εξής κατηγορίες:
1 = Διαφωνώ Απολύτως, 2 = Διαφωνώ, 3 = Ούτε συμφωνώ Ούτε διαφωνώ, 4 = Συμφωνώ,
5 = Συμφωνώ Απολύτως.

Το ερωτηματολόγιο αποτελείται από 72 ερωτήσεις που για την συνέχεια της ανάλυσης θα τις μεταχειριζόμαστε ως κατηγορικές μεταβλητές ή απλά μεταβλητές p, ενώ το μέγεθος του δείγματος είναι $n = 149$ εργαζόμενοι. Σκοπός της ανάλυσης είναι να ομαδοποιήσουμε κατάλληλα τις p μεταβλητές έτσι ώστε να δημιουργήσουμε ομοιογενής ομάδες μικρότερης διάστασης με συγκεκριμένες ιδιότητες και ταυτότητα. Με τον τρόπο αυτό η παραγοντική ανάλυση θα χρησιμοποιηθεί σαν τεχνική μείωσης (ελάττωσης) της αρχικής διάστασης των δεδομένων του $\mathbf{X}(n \times p)$, δηλαδή του ερωτηματολογίου, σε μικρότερα υποσύνολα ώστε οι μεταβλητές στην κάθε ομάδα (group) να έχουν μεγάλες συσχετίσεις μεταξύ τους, ενώ παράλληλα οι μεταβλητές των διαφορετικών group να είναι σχετικά ασυσχέτιστες. Επιπλέον με την εύρεση των παραγόντων (group) θα μπορέσουμε να ελέγξουμε και να επιβεβαιώσουμε την αρχική δομή του ερωτηματολογίου έτσι όπως αυτή δόθηκε σε εμάς από τον ερευνητή. Όπως θα δούμε και στο επόμενο κεφάλαιο, την μείωση της διάστασης ενός συνόλου δεδομένων μπορούμε να την επιτύχουμε επίσης χρησιμοποιώντας την ανάλυση κατά συστάδες αλλά η προσέγγιση στο πρόβλημα είναι διαφορετική.

Για να έχουμε όσο το δυνατόν αμερόληπτες εκτιμήσεις των παραμέτρων, αφαιρέθηκαν τα μη πλήρη δεδομένα με *listwise deletion* παίρνοντας τελικά δείγμα $n = 143$ εργαζομένων. Χρησιμοποιώντας για την εύρεση των παραγόντων τον πίνακα συσχετίσεων $\mathbf{R}_s$ ελέγχουμε πρώτα τον βαθμό των συσχετίσεων μεταξύ των κατηγορικών μεταβλητών στον πίνακα. Είναι πολύ σημαντικό ο πίνακας των συσχετίσεων να έχει μεγάλες τιμές, θεωρητικά μεγαλύτερες του $|0.40|$, αλλιώς δεν έχει νόημα να προχωρήσει η ανάλυση αφού δεν πρόκειται να βρεθεί ένα ικανοποιητικό παραγοντικό μοντέλο για τα δεδομένα. Άλλωστε τις συσχετίσεις αυτές θα προσπαθήσουμε να εξηγήσουμε με τους παράγοντες.

Από τον πίνακα συσχετίσεων - *δεν δίνεται λόγω του μεγέθους του* - για όλα τα ζεύγη μεταβλητών, παρατηρείται μια μέτρια θετική συσχέτιση των μεταβλητών αλλά αυτό είναι κάπως αναμενόμενο αφού έχουμε κατηγορικές μεταβλητές να μετρήσουμε. Οι χαμηλές συσχετίσεις εμφανίζονται κυρίως για τις μεταβλητές $\{\alpha 3, \alpha 5, \alpha 8, \alpha 14, c11, \varepsilon 1\}$ σε σχέση με τις υπόλοιπες (και με αντίστοιχα μεγάλα p-values), οπότε τις εξαιρούμε από την συνέχεια της ανάλυσης αφού όπως όλα δείχνουν θα προκύψουν από μόνες τους σαν ένας ξεχωριστός παράγοντας (Θα δούμε ύστερα ότι στην ανάλυση κατά συστάδες θα προκύψουν σαν μία συστάδα μεταβλητών). Στη συνέχεια, ελέγχουμε να δούμε αν οι τιμές του στατιστικού KMO και του MSA των μεταβλητών είναι ικανοποιητικές. Συγκεκριμένα η τιμή του KMO=0.916 είναι υψηλή που σημαίνει ότι οι συσχετίσεις είναι ικανοποιητικές, ενώ όλες οι τιμές για τα MSA των 66 μεταβλητών είναι πάνω από 0,840 οπότε μπορούν να χρησιμοποιηθούν για την ανάλυση και δεν χρειάζεται να διώξουμε επιπλέον άλλες.

Το πρόβλημα της επιλογής αριθμού παραγόντων είναι άμεσα συνδεδεμένο με τη μέθοδο εκτίμησης του μοντέλου. Τους παράγοντες θα τους εκτιμήσουμε με τη μέθοδο των κύριων συνιστωσών αφού η μέθοδος της μέγιστης πιθανοφάνειας δεν πρέπει να χρησιμοποιείται διότι τα δεδομένα δεν είναι προέρχονται από κανονική κατανομή (οι μεταβλητές-ερωτήσεις είναι κατηγορικές). Άρα αντί να κάνουμε ελέγχους καλής προσαρμογής για το μοντέλο που θα προσαρμόσουμε και για τον αριθμό των παραγόντων που θα κρατήσουμε, μπορούμε να ελέγξουμε το μοντέλο βάσει του ποσοστού της διακύμανσης που εξηγείται από τους παράγοντες ή από τα κατάλοιπα του εκτιμημένου πίνακα συσχετίσεων $\hat{\mathbf{R}}_s$. Επίσης για να επιλέξουμε αριθμό παραγόντων, μπορούμε να χρησιμοποιήσουμε διάφορα κριτήρια που χρησιμοποιούνται επίσης στην ανάλυση σε κύριες συνιστώσες (PCA) και που βασίζονται στις ιδιοτιμές του πίνακα συσχετίσεων $\mathbf{R}_s$. Επομένως μπορεί κανείς να χρησιμοποιήσει το κριτήριο του Kaiser, το scree plot ή την ποσότητα $\sum \tilde{h}_i^2 / p$. Στην περίπτωση αυτή θα πρέπει να προσαρμόσουμε αρκετά μοντέλα και να κρατήσουμε αυτό που θεωρούμε καλύτερο με βάση κάποιο από τα παραπάνω κριτήρια. Παρόλα αυτά ένα μοντέλο με πολλούς παράγοντες που ικανοποιεί τα παραπάνω κριτήρια μπορεί να χάνει ερμηνείας και ένα άλλο που δεν είναι τόσο ικανοποιητικό και έχει λιγότερους παράγοντες να έχει καλύτερη φυσική ερμηνεία.

Όσον αφορά το ερωτηματολόγιο, στην Εικόνα 3 που υπάρχει στο Παράρτημα Β μπορούμε να δούμε τις ιδιοτιμές και το ποσοστό της διακύμανσης που κάθε ιδιοτιμή ερμηνεύει. Με βάση το κριτήριο του Kaiser, υπάρχουν 10 ιδιοτιμές μεγαλύτερες της μονάδας οπότε θα

πρέπει να επιλέξουμε μοντέλο με 10 παράγοντες. Ωστόσο η $1^{η}$ ιδιοτιμή είναι 28.988 και η $10^{η}$ μόλις 1.079. Το scree plot της Εικόνα 4 ωστόσο, δείχνει ότι πρέπει να επιλέξουμε 3 ή 4 παράγοντες το πολύ. Επίσης από τον πίνακα της Εικόνας 3 βλέπουμε ότι μετά από την $4^{η}$ ιδιοτιμή $\lambda_4 = 1.985$, ο ρυθμός μείωσης της διακύμανσης είναι αρκετά αργός. Το γεγονός αυτό ενισχύει την υποψία μας ότι το πολύ τέσσερεις παράγοντες χρειάζεται να επιλέξουμε. Επίσης, αθροιστικά, οι 4 παράγοντες εξηγούν το 61.085% της συνολικής διακύμανσης ενώ οι 10 παράγοντες με το κριτήριο Kaiser το 73.279%.

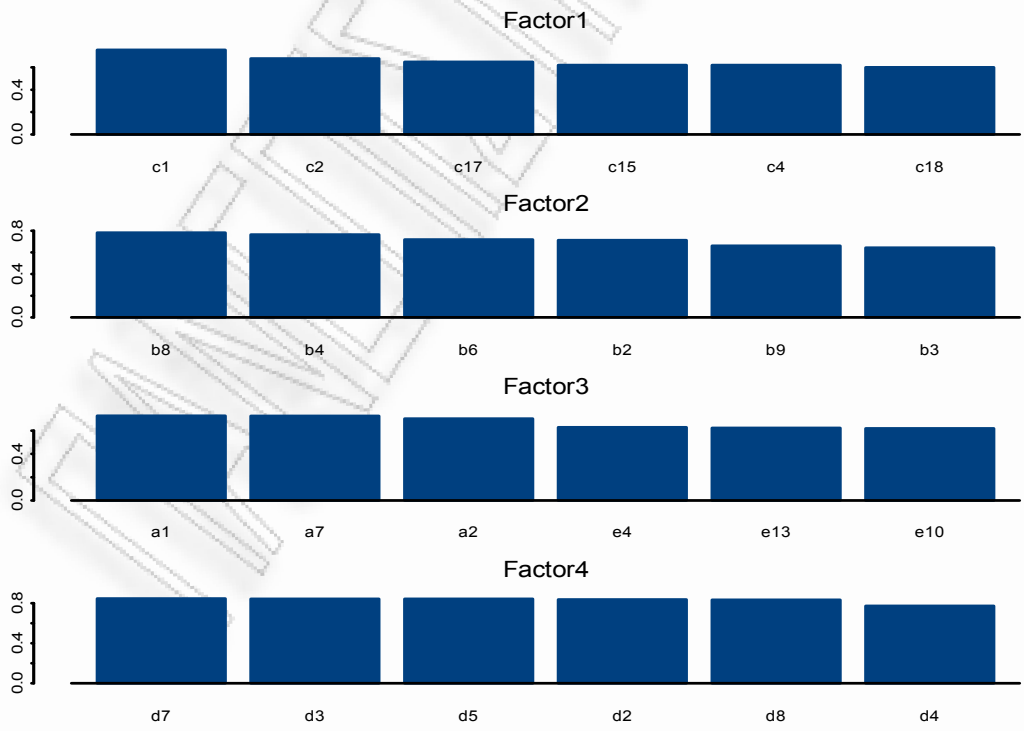
Για την τελική λύση διάφορα παραγοντικά μοντέλα εξετάστηκαν μεταξύ τριών και έξι παραγόντων καθώς και εκείνο με τους δέκα παράγοντες. Τελικά επιλέχθηκε το μοντέλο με τέσσερεις παράγοντες ως το πιο κατάλληλο για λόγους φυσικής ερμηνείας των παραγόντων αλλά και βάσει των παραπάνω κριτηρίων. Από όλα αυτά τα διαφορετικά μοντέλα παρατηρήθηκε, ότι οι μεταβλητές είχαν την «τάση» να δημιουργήσουν 4 group μεταβλητών. Επίσης, τις παραμέτρους του παραγοντικού μοντέλου τις εκτιμήσαμε με την μέθοδο των κύριων συνιστωσών όπως προαναφέραμε, ενώ για να δούμε καλύτερα την υποκείμενη δομή των παραγόντων χρησιμοποιήσαμε την ορθογώνια περιστροφή *varimax* των αξόνων.

Στην Εικόνα 1 του Παραρτήματος Β υπάρχει ο πίνακας $\hat{\mathbf{L}}^{*}$, δηλαδή τα φορτία μετά την ορθογώνια περιστροφή, όπου με έντονο χρώμα είναι τα φορτία μεγαλύτερα από **0.45** και που θα μας χρησιμεύσουν για την αναγνώριση των παραγόντων. Επίσης στην Εικόνα 2 δίνεται ο πίνακας με τις εταιρικότητες των μεταβλητών για το παραγοντικό μοντέλο. Από αυτόν τον πίνακα μπορούμε να δούμε τις συνολικές διακυμάνσεις για τις μεταβλητές που εξηγούν οι παράγοντες για το μοντέλο που εκτιμήσαμε. Η συνολική εικόνα δείχνει να εξηγείται ένα καλό ποσοστό για τις μεταβλητές εκτός από τις α10, α15 και ε7 που το μοντέλο με τέσσερεις παράγοντες ερμηνεύει το 39.5%, 39.3% και 39.8% της διακύμανσης τους. Αυτό ίσως να σημαίνει ότι περισσότεροι παράγοντες χρειάζονται για να αυξηθούν τα ποσοστά αυτά αλλά αυτό δεν είναι απαραίτητο ότι θα συμβεί. Πράγματι αν προσθέσουμε κι άλλον παράγοντα τότε παίρνουμε το 44% για την α10, το 44.7% για την α15 και το 44.5% για την ε7, δηλαδή καμία ουσιαστική διαφορά και το μοντέλο με πέντε παράγοντες δεν έχει καλή ερμηνεία. Οπότε δεν υπάρχει λόγος να μην δεχτούμε το μοντέλο με τέσσερεις παράγοντες. Επίσης ένας ακόμη λόγος για να αποδεχτούμε το συγκεκριμένο παραγοντικό μοντέλο είναι ότι στον εκτιμημένο πίνακα καταλοίπων $\hat{\mathbf{E}}$ οι διαφορές είναι αρκετά μικρές.

Αν και τα φορτία της Εικόνας 1 δεν φαίνεται να ικανοποιούν πλήρως το *simple structure* ωστόσο φαίνεται ότι έχουν ένα ερμηνεύσιμο μοτίβο (pattern). Από τα φορτία του πίνακα $\hat{\mathbf{L}}^{*}$

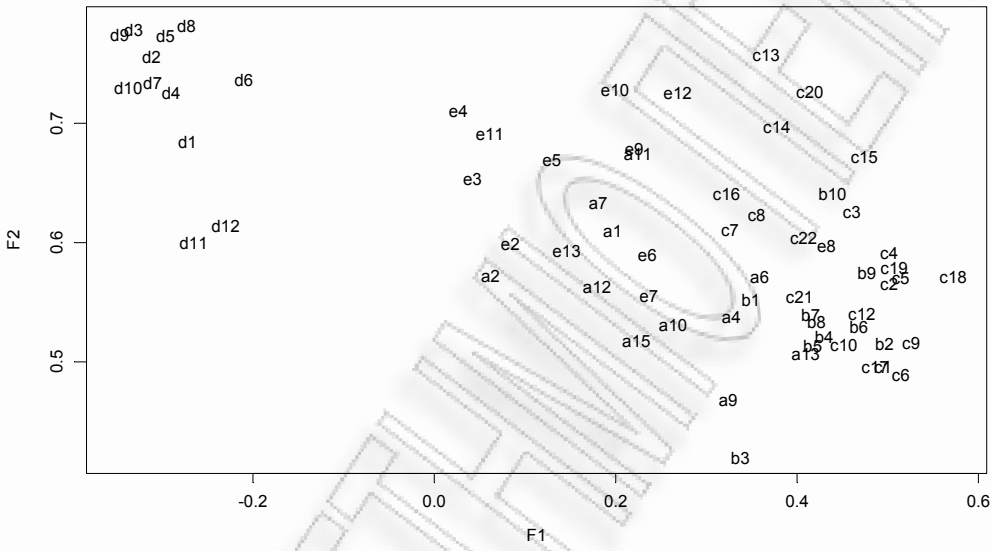
παρατηρούμε ότι στη διαμόρφωση του $1^{ου}$ παράγοντα συνεισφέρουν περισσότερο οι μεταβλητές {α4, α6, α9, α10, α13, α15, c1, c2, c3, c4, c5, c9, c10, c12, c13, c14, c15, c16, c17, c18, c19, c20, c21, c22}, στον $2^{ο}$ παράγοντα οι μεταβλητές {β1, β2, β3, β4, β5, β6, β7, β8, β9, β10, c6, c7, c8, ε8}, στον $3^{ο}$ παράγοντα οι {α1, α2, α7, α11, α12, ε2, ε3, ε4, ε5, ε6, ε7, ε9, ε10, ε11, ε12, ε13} και στον $4^{ο}$ παράγοντα οι {d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11, d12}. Αυτό σημαίνει ότι 66 μεταβλητές μπορούν να μελετηθούν ανά κατηγορίες όπου κάθε κατηγορία αντιστοιχεί σε ένα παράγοντα. Παραπέρα το ερωτηματολόγιο της έρευνας μπορεί να δομηθεί με 4 ενότητες και η κάθε ενότητα να περιέχει συγκεκριμένες ερωτήσεις. Έτσι δημιουργείται ομοιογένεια και δίνεται η δυνατότητα στον ερευνητή να βγάλει ευκολότερα τα συμπεράσματα του. Επίσης αφού οι ερωτήσεις-μεταβλητές μέσα σε κάθε κατηγορία είναι αμοιβαίως συσχετισμένες, είναι δυνατόν σε νέα έρευνα κάποιες να παραληφθούν και το νέο ερωτηματολόγιο να είναι πιο περιεκτικό και ενδεχόμενα να αποτυπώσει καλύτερα το σκοπό για τον οποίο κατασκευάστηκε. Ακόμα μπορούμε να δώσουμε και μια ερμηνεία ή μια «επικεφαλίδα» για κάθε παράγοντα ή ενότητα του ερωτηματολογίου, ώστε να γνωρίζουμε τι αντιπροσωπεύει κάθε group μεταβλητών. Σε αυτό ίσως να μας βοηθήσει το παρακάτω γράφημα που δείχνει τις μεταβλητές με τα υψηλότερα φορτία σε κάθε παράγοντα.

Γράφημα 1: *Οι μεταβλητές με τα μεγαλύτερα φορτία για κάθε παράγοντα για το μοντέλο.*



Θα μπορούσαμε λοιπόν να ονομάσουμε τον 1ο παράγοντα ως «*Συνάντηση αξιολόγησης και ικανότητα των αξιολογητών-προϊσταμένων*», τον 2ο παράγοντα ή ενότητα «*Γενική εικόνα των προϊσταμένων*», την 3η ενότητα (κατηγορία) «*Στήριξη από την επιχείρηση και ποιότητα των κριτηρίων αξιολόγησης*» και τον 4ο παράγοντα ως «*Χρησιμότητα και αξιοπιστία του συστήματος αξιολόγησης*».

Γράφημα 2: *Γραφική απεικόνιση των rotated loadings για m=2.*

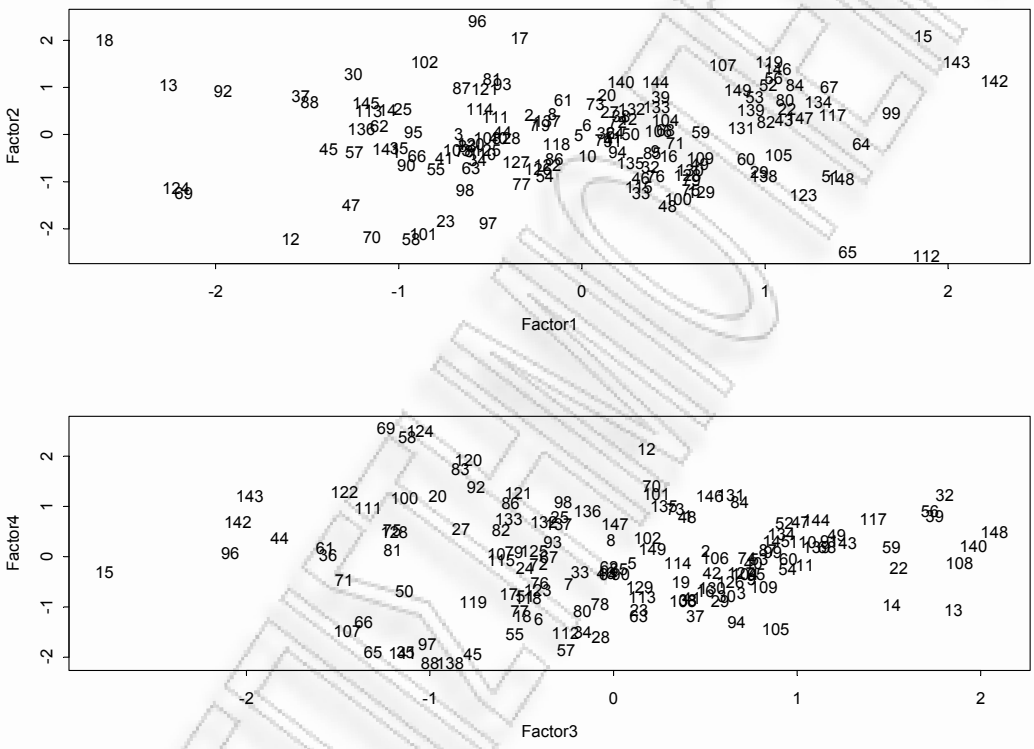


Επιπλέον, με γνώμονα τις 4 ενότητες μπορεί ο ερευνητής να δημιουργήσει νέες ερωτήσεις και να ελέγξει το πού θα τις τοποθετήσει. Ένα άλλο πλεονέκτημα που του δίνεται είναι ότι μπορεί να θεωρήσει τους παράγοντες ως συνεχείς «καινούργιες» μεταβλητές και να εκτιμά σκορ για τους νυν και μελλοντικούς ερωτηθέντες στις μεταβλητές αυτές. Αυτό ίσως τον βοηθήσει να δει πως ομαδοποιούνται οι ερωτώμενοι και μαζί με κάποια άλλα χαρακτηριστικά όπως για παράδειγμα το φύλο, η ηλικία, το μορφωτικό επίπεδο, να δει αν οι παράγοντες έχουν διακριτική ικανότητα μέσα στις διάφορες υποομάδες. Οι τιμές των σκορ των ερωτηθέντων επίσης μπορούν να αποτελέσουν κριτήριο για να ξεχωρίσουν κάποιοι εργαζόμενοι από το σύστημα αξιολόγησης, υπό την έννοια ότι μεγαλύτερα σκορ σε κάποιους ή σε όλους τους παράγοντες μπορεί να σημαίνει την αποτελεσματικότερη αξιολόγηση τους.

Στην Εικόνα 5 του Παραρτήματος μπορεί κανείς να δει τις εκτιμήσεις του πίνακα των συντελεστών των σκορ. Από τον πίνακα αυτόν μπορούμε να βρούμε τα σκορ και για τους 143 εργαζόμενους. Όταν $m > 2$ είναι δύσκολο σε ένα γράφημα να δούμε την απεικόνιση τους και να βγάλουμε συμπεράσματα. Μια λύση είναι να απεικονίσουμε ανά δύο τους παράγοντες.

Στο Γράφημα 3 έχουμε αναπαραστήσει τα σκορ των παρατηρήσεων παίρνοντας ανά ζεύγη τους άξονες (παράγοντες). Έτσι για παράδειγμα βλέπουμε ότι στους 2 πρώτους άξονες η παρατήρηση 18 έχει την μικρότερη τιμή (σκορ) και η παρατήρηση 142 την μεγαλύτερη. Στον $3^{ο}$ και $4^{ο}$ άξονα μαζί, η παρατήρηση 15 έχει την μικρότερη τιμή και η παρατήρηση 148 την μεγαλύτερη.

Γράφημα 3: *Απεικόνιση των σκορ για όλες τις παρατηρήσεις παίρνοντας ανά ζεύγη τους παράγοντες.*



ΚΕΦΑΛΑΙΟ 3

Ανάλυση Κατά Συστάδες

3.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζουμε την ανάλυση κατά συστάδες (cluster analysis), τη δημοφιλή μέθοδο η οποία χρησιμοποιείται από τους αναλυτές για την ομαδοποίηση των παρατηρήσεων ή των μεταβλητών, γενικά των αντικειμένων, ενός πολυδιάστατου συνόλου δεδομένων. Στην παρούσα εργασία όμως, δεν θα χρησιμοποιήσουμε την ανάλυση κατά συστάδες για την ομαδοποίηση των παρατηρήσεων ως προς τις μεταβλητές. Το αποτέλεσμα θα είναι ότι το κεφάλαιο αναφέρεται εξολοκλήρου στον τρόπο που εφαρμόζουμε την μέθοδο για να ομαδοποιούμε μεταβλητές. Έτσι, οι έννοιες της απόστασης και της ομοιότητας καθώς και οι αλγόριθμοι ομαδοποίησης που θα αναλύσουμε αναφέρονται αποκλειστικά στον σκοπό αυτό, χωρίς αυτό να σημαίνει ότι οι ίδιοι αλγόριθμοι δεν χρησιμοποιούνται για την ομαδοποίηση των παρατηρήσεων. Άλλωστε οι πιο ανεπτυγμένες μέθοδοι ομαδοποίησης και αλγόριθμοι υλοποίησης τους έχουν σχεδιαστεί για να βρίσκουν ομοιογενείς ομάδες παρατηρήσεων καθώς η ομαδοποίηση μεταβλητών μπορεί να γίνει και με άλλες μεθόδους.

Για να γίνει κατανοητό πως η ανάλυση κατά συστάδες μπορεί να χρησιμοποιηθεί σαν μέθοδος εύρεσης ομοιογενών ομάδων πρέπει πρώτα να δοθεί μια συγκεκριμένη μεθοδολογία που χρειάζεται κανείς να ακολουθήσει. Συνήθως η διαδικασία που απαιτείται για να έχει πραγματικά αποτελέσματα η ανάλυση κατά συστάδες αφορά: (1) Την ομοιογένεια των μεταβλητών του συνόλου δεδομένων, (2) Την επιλογή ενός μέτρου απόστασης ή ομοιότητας, (3) Τη χρήση κάποιου αλγορίθμου και μεθόδου ομαδοποίησης και (4) Την αξιολόγηση της μεθόδου και τον τελικό αριθμό των συστάδων.

Τα τέσσερα σημεία ανάλογα με τη φύση των δεδομένων και τις απαιτήσεις της έρευνας θα πρέπει κάθε φορά να τροποποιούνται για να έχουμε επιθυμητά αποτελέσματα. Αυτό σημαίνει ότι για παράδειγμα, αν για το (1) οι μεταβλητές θα πρέπει για την ομοιογένεια τους να γίνουν όλες δίτιμες και αυτό είναι χρονοβόρο και δύσκολο, τότε μπορούμε να επιλέξουμε κάποια άλλη κλίμακα μέτρησης για να τις μετασχηματίσουμε. Παρόμοια συμπεράσματα ισχύουν για

το (2) και (3), δηλαδή αν επιλέξουμε για την ομαδοποίηση τη μέθοδο της απλής συνένωσης και δεν πάρουμε μία ποιοτική ομαδοποίηση μπορούμε να χρησιμοποιήσουμε εναλλακτικά κάποιον άλλον αλγόριθμο ή κάποια άλλη μέθοδο. Για το (4), η αξιολόγηση μπορεί να γίνει είτε με κάποια «εσωτερικά» κριτήρια είτε με «εξωτερικά» κριτήρια, ενώ ο αριθμός των συστάδων μπορεί να καθοριστεί από την αρχή ή μπορεί να αποφασιστεί αφότου ολοκληρωθεί ο αλγόριθμος ομαδοποίησης. Άρα συμπεραίνουμε ότι η μέθοδος ενέχει την υποκειμενικότητα του αναλυτή αλλά και την ευελιξία καθώς μπορεί εύκολα και χωρίς κόπο (αφού δεν χρειάζεται να προσαρμόσει κάποιο μοντέλο και να κάνει ελέγχους) να κάνει διαφορετικές επιλογές με σκοπό να επιλέξει την καλύτερη για τα δεδομένα του και τις ανάγκες του προβλήματος που έχει να αντιμετωπίσει ομαδοποίηση. Ωστόσο όπως και ο Anderberg (1973) αναφέρει «Η ανάλυση κατά συστάδες δεν είναι εύκολη διαδικασία ακόμα και όταν *διαισθητικά* η ομαδοποίηση μας είναι ορατή».

3.2 Ομοιογένεια των Μεταβλητών

Γενικά, οι μεταβλητές ταξινομούνται στις ποσοτικές (quantitative) και στις ποιοτικές ή κατηγορικές (qualitative/categorical) ανάλογα με την κλίμακα μέτρησης τους. Οι κλίμακες μέτρησης διακρίνονται σε τέσσερεις κατηγορίες: στην ονομαστική (nominal) κλίμακα, στην διατάξιμη (ordinal) κλίμακα, στην διαστηματική (interval) κλίμακα και στην ratio κλίμακα.

Στις ποσοτικές μεταβλητές ανήκει η διαστηματική και η ratio κλίμακα. Η ratio κλίμακα έχει τις ιδιότητες (α) του λόγου (αναλογίας) δύο τιμών, (β) της απόστασης ανάμεσα σε δύο αντικείμενα της κλίμακας και (γ) της διάταξης των αντικειμένων κατά μήκος της κλίμακας. Για την διαστηματική κλίμακα μπορούμε να πούμε μόνο ότι ικανοποιεί τις ιδιότητες (β) και (γ) της ratio κλίμακας. Για τις ποιοτικές μεταβλητές μπορούμε να πούμε ότι: ανήκουν αυτές που είναι σε ονομαστική κλίμακα και που έχουν δύο ή περισσότερες τιμές για να ξεχωρίζουν οι κατηγορίες, για παράδειγμα το φύλο [άντρας (π.χ. σκορ 0) και γυναίκα (π.χ. σκορ 1)], η πολιτική ιδεολογία [δημοκρατικός (1), ανεξάρτητος (2), ρεπουμπλικάνος (3)], το χρώμα ματιών [μπλε (1), καστανό (2), πράσινο (3)] και αυτές που είναι σε διατάξιμη κλίμακα με κατηγορίες που εμφανίζουν διάταξη, όπως για παράδειγμα το ετήσιο οικονομικό εισόδημα [χαμηλό (1), μεσαίο (2), υψηλό (3)], η προτίμηση σε μια συγκεκριμένη μάρκα μπύρας (π.χ. με σκορ 1,……,5) ούτως ώστε όσο μεγαλύτερο είναι το σκορ τόσο μεγαλύτερος είναι ο βαθμός προτίμησης, ή το επίπεδο μόρφωσης [γυμνάσιο (1), λύκειο (2), πανεπιστήμιο (3)]. Στο πρώτο είδος μεταβλητών τα σκορ χρησιμεύουν για να ξεχωρίζουν διακριτά οι κατηγορίες

των μεταβλητών ενώ στο δεύτερο παίρνουν τον ρόλο μιας σειράς στην οποία ο ερωτώμενος αποκρίνεται βάσει της προτίμησης του. Συνεπώς από τις ποιοτικές μεταβλητές η διατάξιμη κλίμακα ικανοποιεί μόνο την (γ) ιδιότητα, ενώ η ονομαστική κλίμακα δεν ικανοποιεί καμία.

Οι παραπάνω κλίμακες μέτρησης κατατάσσονται ιεραρχικά από την ονομαστική στην ratio κλίμακα. Άρα κανείς μπορεί πάντα να πάει από υψηλότερη σε χαμηλότερη κλίμακα χάνοντας όμως πληροφορίες όπως και να ανέβει μία κλίμακα κάνοντας κάποιες επιπλέον υποθέσεις.

Όπως αναφέρθηκε νωρίτερα, για την ανάλυση κατά συστάδες θα πρέπει να υπάρχει ομοιογένεια της κλίμακας μέτρησης των μεταβλητών αλλιώς δεν θα είμαστε σίγουροι ότι ο αλγόριθμος που χρησιμοποιήσαμε για να κάνουμε την ομαδοποίηση έδωσε τα σωστά αποτελέσματα αφού οι αποστάσεις που θα υπολογιστούν δεν θα έχουν παντού την ίδια βαρύτητα. Στην περίπτωση που ο πίνακας δεδομένων, $\mathbf{Y} = [y_{ij}]$, περιέχει μεταβλητές μικτού τύπου τότε για τον αναλυτή μία επιλογή που έχει να κάνει είναι να αποφασίσει σε ποια κλίμακα επιθυμεί να συνεχίσει την ανάλυση και να μετασχηματίσει τις υπόλοιπες μεταβλητές σε αυτήν ώστε να δημιουργήσει ομοιογένεια των μεταβλητών για την κλίμακα μέτρησης. Ωστόσο υπάρχει και η εναλλακτική οδός της ομαδοποίησης με ομοειδής μεταβλητές. Αυτό σημαίνει ότι κάνουμε ομαδοποιήσεις για κάθε τύπο μεταβλητών ξεχωριστά ελπίζοντας ότι οι ομάδες που θα βρούμε θα είναι περίπου όμοιες. Στη πράξη όμως κάτι τέτοιο δεν είναι απαραίτητο να συμβεί και συνήθως τα αποτελέσματα οδηγούν σε διαφορετικές ταξινομήσεις.

Αν λοιπόν έχουμε ένα σετ δεδομένων με μεταβλητές σε μικτή μορφή και έστω ότι επιλέξουμε τον μετασχηματισμό τους στη διατάξιμη (ordinal) κλίμακα ώστε να υπάρξει ομοιογένεια, τότε σε γενικές γραμμές μπορούμε να κάνουμε τα εξής:

(1) Για τη περίπτωση των μεταβλητών με διαστηματική ή ratio κλίμακα μπορούμε να χωρίσουμε το διάστημα σε επίπεδα (κατηγορίες) και να ορίσουμε την διάταξη. Το πρόβλημα βέβαια είναι να οριστούν οι γειτονικές κατηγορίες ή να δημιουργηθούν οι κλάσεις ίσου ή άνισου πλάτους για την κλίμακα και να βρεθεί η απόσταση των γειτονικών κατηγοριών, έτσι ώστε οι παρατηρήσεις σε μια κατηγορία να έχουν την ίδια κατάταξη (rank) ενώ παράλληλα να διατηρείται η διαταξιμότητα ανάμεσα στις παρατηρήσεις των διαφορετικών κατηγοριών. Ένα από τα μειονεκτήματα της μεθόδου αυτής είναι πως οι διαφορές που υπάρχουν μεταξύ των αντικειμένων της ίδιας κατηγορίας χάνονται ή δεν λαμβάνονται υπόψη.

(2) Για τις μεταβλητές που είναι σε ονομαστική κλίμακα, όπως αναφέραμε νωρίτερα οι τιμές που τους δίνουμε δεν πρέπει να συγχέονται με κάποιο είδος διάταξης. Εδώ ο στόχος

είναι να επιβάλουμε μια διάταξη στις κατηγορίες τους - εάν αυτό είναι εφικτό - και να εκτιμήσουμε τα σκορ των κατηγοριών. Υπάρχουν διάφορες προσεγγίσεις. Μία ενδεδειγμένη λύση είναι να χρησιμοποιήσουμε μια μεταβλητή αναφοράς (reference-instrument variable) την οποία θα ταξινομήσουμε σε πίνακα με την μεταβλητή που μας ενδιαφέρει να ορίσουμε διάταξη. Για να γίνει αυτό, έστω Y μία διαστηματική μεταβλητή, πχ. η ηλικία, την οποία διακριτοποιούμε όπως και στο (1) σε j κατηγορίες ή επίπεδα και έστω X η ονομαστική μεταβλητή με i κατηγορίες. Τότε χρησιμοποιώντας μοντέλα συνάφειας για την υπό συνθήκη συνάφεια ανάμεσα στην X που μεταβάλλεται στα επίπεδα της Y μεταβλητής θα μας δώσουν εκτιμήσεις για τις κατηγορίες της (**Σημ.** Περισσότερα για το θέμα αυτό στα επόμενα δύο κεφάλαια). Για να γίνει όμως σωστά αυτό θα πρέπει η επιλογή της μεταβλητής αναφοράς να είναι σχετική με την ονομαστική μεταβλητή απόκρισης. Επιπλέον, στις δύο περιπτώσεις που αναφέραμε, τα σκορ των κατηγοριών μπορούν να χρησιμοποιηθούν για τον υπολογισμό του συντελεστή συσχέτισης r_{xy}.

3.3 Μέτρα Ομοιότητας και Απόστασης για τις Μεταβλητές

Οι έννοιες της απόστασης και της ομοιότητας είναι δύο έννοιες αντίθετες. Η απόσταση έχει σαν σκοπό να μετρήσει πόσο απέχουν δύο αντικείμενα. Συχνά οι δύο έννοιες περιγράφονται ενιαία και έτσι θα τις θεωρήσουμε και εμείς. Με την ανάλυση κατά συστάδες γενικά, σκοπός είναι να δημιουργήσουμε ομάδες μέσα στις οποίες τα αντικείμενα είτε αυτά είναι οι παρατηρήσεις είτε είναι οι μεταβλητές να απέχουν λίγο, ενώ τα αντικείμενα για τις διαφορετικές ομάδες να απέχουν αρκετά, δηλαδή να είναι ανόμοια.

Ο τρόπος με τον οποίο υπολογίζουμε την εγγύτητα (proximity) για όλες τις p μεταβλητές του πίνακα $\mathbf{Y} = [y_{ij}]$ έχει να κάνει πρωτίστως με το μέτρο που υπολογίζουμε την απόσταση. Στην πράξη κατασκευάζουμε έναν πίνακα ανομοιότητας (dissimilarity matrix) ή ομοιότητας (similarity matrix) αντίστοιχα, ο οποίος περιγράφει τον βαθμό της συνάφειας για όλα τα ζεύγη του πίνακα. Εάν με την εγγύτητα θέλουμε να δείξουμε το μέτρο της ομοιότητας, τότε η τιμή του θα είναι μεγάλη όταν δύο αντικείμενα είναι κοντά ή «όμοια» μεταξύ τους, ενώ εάν ζητείται η απόσταση και το μέτρο της ανομοιότητας, η τιμή του μέτρου αυτού θα είναι μικρή για τα αντικείμενα που είναι σχεδόν κοντά ή «όμοια». Δηλαδή βλέπουμε ότι μια ομοιότητα μπορεί εύκολα να μετασχηματιστεί σε ανομοιότητα ή απόσταση αν χρησιμοποιήσουμε το συμπληρωματικό της.

Ο πίνακας αυτός αποτελεί το σημείο εκκίνησης για την ανάλυση κατά συστάδες. Τα στοιχεία του πίνακα είναι οι τιμές για το μέτρο που θα υπολογίσουμε και οι αλγόριθμοι για να κάνουν την ομαδοποίηση θα υλοποιηθούν πάνω στα στοιχεία του πίνακα χωρίς να χρειαστεί ο αναλυτής να επιστρέψει ποτέ στον αρχικό πίνακα δεδομένων. Παράλληλα, αποτελεί και ένα εσωτερικό κριτήριο για την εγκυρότητα της μεθόδου ομαδοποίησης που θα χρησιμοποιήσει. Ο πίνακας ανομοιότητας είναι συμμετρικός $p \times p$ ή $n \times n$ όταν τα στοιχεία του είναι οι αποστάσεις ανάμεσα σε δύο παρατηρήσεις και η μορφή του είναι η εξής

$$\mathbf{D} = \begin{pmatrix} 0 & \cdots & d_{1p} \\ \vdots & \ddots & \vdots \\ d_{p1} & \cdots & 0 \end{pmatrix}$$

Όπως έχουμε πει θα αναφερθούμε μόνο στην περίπτωση της ομαδοποίησης των μεταβλητών του πίνακα $\mathbf{Y} = [y_{ij}]$. Για τον σκοπό αυτό χρειάζεται να οριστούν μέτρα ομοιότητας ή ανομοιότητας, που να χαρακτηρίζουν τις σχέσεις μεταξύ των στηλών του αρχικού πίνακα. Τα μέτρα αυτά αναγκαστικά θα πρέπει να είναι συμμετρικά. Έτσι, αν $A(X_1, X_2)$ είναι η συνάφεια ανάμεσα στις μεταβλητές X_1 και X_2 τότε θα πρέπει να ισχύει ότι: $A(X_1, X_2) = A(X_2, X_1)$. Διακρίνουμε τις παρακάτω περιπτώσεις:

3.3.1 Οι Μεταβλητές είναι Ποσοτικές

Αν συμβολίσουμε με $(y_{1f}, y_{2f}, ..., y_{nf})$ την $\mathbf{f}-$στήλη του πίνακα $\mathbf{Y} = [y_{ij}]$ και με $(y_{1g}, y_{2g}, ..., y_{ng})$ την $\mathbf{g}-$στήλη του ίδιου πίνακα τότε στην περίπτωση που έχουμε ποσοτικές μεταβλητές ως μέτρο ομοιότητας δύο μεταβλητών χρησιμοποιείται ο παραμετρικός συντελεστής συσχέτισης του Pearson, r_P. Έτσι αν οι f και g είναι δύο μεταβλητές του πίνακα, ο συντελεστής του Pearson υπολογίζεται από την σχέση

$$r_P = \frac{\mathrm{Cov}(f,g)}{\left[\mathrm{Var}(f)\mathrm{Var}(g)\right]^{1/2}} = \frac{\sum_{i=1}^{n}(y_{if} - \bar{y}_f)(y_{ig} - \bar{y}_g)}{\left(\left[\sum_{i=1}^{n}(y_{if} - \bar{y}_f)^2\right]\left[\sum_{i=1}^{n}(y_{ig} - \bar{y}_g)^2\right]\right)^{1/2}}$$

Ο συντελεστής συσχέτισης μετράει το μέγεθος τις γραμμικής συσχέτισης ανάμεσα στις μεταβλητές f και g, ενώ συνήθως για τον υπολογισμό του υποθέτουμε ότι ο πληθυσμός από

όπου προέκυψε το δείγμα είναι κανονικός. Το εύρος των τιμών του είναι μεταξύ του διαστήματος [-1,1] και είναι ανεξάρτητος των μονάδων μέτρησης.

Εναλλακτικά για την μέτρηση του βαθμού συσχέτισης δύο μεταβλητών, μπορούμε να χρησιμοποιήσουμε το συνημίτονο της γωνίας μεταξύ των δύο $(n\times 1)$ διανυσμάτων **f** και **g**. Τότε το συνημίτονο της γωνίας α δίνεται από τη σχέση

$$\cos\alpha_{fg} = \frac{\sum_{i=1}^{n} y_{if}\, y_{ig}}{\left(\left[\sum_{i=1}^{n} y_{if}^2\right]\left[\sum_{i=1}^{n} y_{ig}^2\right]\right)^{1/2}}$$

Όσο πιο παράλληλα είναι τα δύο διανύσματα **f** και **g**, τόσο μεγαλύτερη είναι η τιμή του συνημιτόνου αφού η γωνία που σχηματίζεται μεταξύ τους είναι μικρή. Όσον αφορά τη βασική διαφορά τους αυτή έχει να κάνει με τις κλίμακες μέτρησης των μεταβλητών. Το συνημίτονο της γωνίας κάνει χρήση της πληροφορίας από την ratio κλίμακα ενώ ο συντελεστής συσχέτισης χρησιμοποιείται μόνο για τη διαστηματική (interval) κλίμακα. Τα παραπάνω δύο μέτρα συσχέτισης ονομάζονται και *Q-type* μέτρα ομοιότητας.

3.3.2 Οι Μεταβλητές είναι Ποιοτικές (Κατηγορικές)

Όπως είπαμε νωρίτερα, οι κατηγορικές μεταβλητές διακρίνονται σε διατάξιμες και σε ονομαστικές (ή καλύτερα στην ονομαστική και στη διατάξιμη κλίμακα) στις οποίες συμπεριλαμβάνονται και οι δίτιμες μεταβλητές.

Όταν η κλίμακα είναι διατάξιμη, τότε δίνοντας συνήθως ακέραια σκορ για τις κατηγορίες τους θεωρούμε γνωστή τη διάταξη των κατηγοριών τους και προχωράμε στην ανάλυση με βάση τις τιμές αυτές. Όμως δεν μπορούμε να πούμε με βεβαιότητα ότι η απόσταση που υπάρχει ανάμεσα σε δύο γειτονικές κατηγορίες είναι ίση με την απόσταση που υπάρχει ανάμεσα στις δύο επόμενες κατηγορίες. Θα πρέπει να σημειωθεί ότι όταν στις διατάξιμες μεταβλητές δίνονται διακριτές τιμές $(\pi\chi.1,2,3,4,5)$ τότε μπορούμε απευθείας να τις θεωρήσουμε σαν να έχουν κατάταξη (ranks).

Με βάση τα παραπάνω, για να υπολογίσουμε ένα μέτρο ομοιότητας μεταξύ δύο κατηγορικών μεταβλητών σε διατάξιμη κλίμακα, αρκεί να βρούμε τις συσχετίσεις τους χρησιμοποιώντας τον μη-παραμετρικό συντελεστή του Spearman r_S, χρησιμοποιώντας τα ranks των κατηγοριών τους βάσει της σχέσης

$$r_S = 1 - 6\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)}$$

όπου $-1 \le r_s \le 1$ και n το μέγεθος του δείγματος. Όταν χρησιμοποιούνται τα ranks για τον υπολογισμό του συντελεστή r_S, τις τιμές που είναι ίσες (ties) συνήθως τις αντικαθιστούμε με την μέση τιμή τους. Ο τρόπος αυτός υπολογισμού των συσχετίσεων είναι αρκετά εύχρηστος για τον εντοπισμό ακραίων παρατηρήσεων. Ο συντελεστής r_S ερευνά κυρίως για την ύπαρξη μονότονης σχέσης των μεταβλητών. Στην ειδική περίπτωση που $r \cong r_S$, τότε υπάρχει μεγάλη γραμμική σχέση μεταξύ των μεταβλητών (Kuanli et al., 1996).

Επίσης, αντί του συντελεστή r_S μπορεί να χρησιμοποιηθεί ο συντελεστής gamma των Goodman and Kruskal (1963). Το μέτρο αυτό υπολογίζεται όπως ο tau του Kendall, μόνο που τα ties δεν συμπεριλαμβάνονται στον υπολογισμό. Για την μέτρηση της συσχέτισης μεταξύ ζεύγη αντικειμένων ο gamma ορίζεται ως

$$\gamma = \frac{C - D}{C + D}$$

όπου C αντιπροσωπεύει τον αριθμό για τα ζεύγη των συμφωνιών (concordant) και D τον αριθμό για τα ζεύγη των ασυμφωνιών (discordant). Όπως και με τον συντελεστή συσχέτισης ισχύει ότι $-1 \le \gamma \le 1$. Παρόμοια με τον συντελεστή r_S, μόνο η μονοτονία χρειάζεται μεταξύ δύο μεταβλητών για $|\gamma| = 1$.

Όταν οι κατηγορικές μεταβλητές είναι σε ονομαστική κλίμακα είναι λάθος να τις μεταχειριζόμαστε σαν διατάξιμες και να υπολογίζουμε απευθείας είτε τον r_S είτε τον tau. Εκείνο που πρέπει να γίνεται στην περίπτωση αυτή είναι να βρούμε τα σκορ για τις κατηγορίες τους όπως αναφέραμε νωρίτερα και να τα αντικαταστήσουμε με τα ranks τους. Είναι σημαντικό να γνωρίζουμε ότι η χρησιμοποίηση των ranks στις ονομαστικές μεταβλητές αφορά περισσότερο τη διατήρηση της κλίμακας τους αφού γίνεται μόνο η προσπάθεια να επιφέρουμε μια διάταξη και όχι της εύρεσης της απόστασης μεταξύ των κατηγοριών τους. Τότε και μόνο τότε μπορούν να χρησιμοποιηθούν οι παραπάνω συντελεστές.

Στην περίπτωση που ο πίνακας $\mathbf{Y} = [y_{ij}]$ περιέχει δίτιμες μεταβλητές, κατασκευάζουμε 2×2 πίνακες συνάφειας και στη συνέχεια υπολογίζουμε το στατιστικό K (Kaufman and Rousseeuw, 1990) βάσει της σχέσης:

$$K = \frac{(ad - bc)^2 n}{(a+b)(a+c)(b+d)(c+d)}$$

το οποίο προσεγγιστικά ακολουθεί X^2 κατανομή με 1 βαθμό ελευθερίας για κάθε πίνακα ξεχωριστά και $n = a + b + c + d$. Το στατιστικό αυτό χρησιμοποιείται για τον έλεγχο ανεξαρτησίας δύο μεταβλητών. Όμως η τιμή του K θεωρείται και σαν μέτρο συσχέτισης των μεταβλητών άρα μπορεί να χρησιμοποιηθεί ως μέτρο ομοιότητας. Ο λόγος είναι ότι το στατιστικό K συνδέεται άμεσα με τον συντελεστή συσχέτισης του Pearson από την σχέση $|r_P| = \sqrt{K / n}$. Αυτό είναι ένα πόρισμα των Lance and Williams (1965). Χαμηλές τιμές για το K υποδηλώνουν ότι οι μεταβλητές είναι ασυσχέτιστες συνεπώς, η μεταξύ τους απόσταση είναι μεγάλη.

3.3.3 Οι Συσχετίσεις ως Μέτρα Ομοιότητας ή Ανομοιότητας για τον Πίνακα D

Αφού δείξαμε τους τρόπους που μπορεί κανείς να υπολογίσει τις συσχετίσεις των μεταβλητών (παραμετρικές ή μη-παραμετρικές) για διάφορες περιπτώσεις, για να μπορέσουμε να τις χρησιμοποιήσουμε στον πίνακα ανομοιότητας $\mathbf{D} = \left[d_{fg} \right]$ θα πρέπει να τις μετατρέψουμε σε ανομοιότητες (dissimilarities), συμβολικά $d(f, g)$, ώστε βάσει του πίνακα αυτού να γίνει η ομαδοποίηση τους. Για να γίνει αυτό μία λογική θα είναι να χρησιμοποιήσουμε τη σχέση: $d(f, g) = (1 - R(f, g))$, όπου $R(f, g)$ συμβολίζει γενικά τον συντελεστή συσχέτισης. Βάσει της σχέσης αυτής οι μεταβλητές με μεγάλη θετική συσχέτιση θα έχουν συντελεστή ανομοιότητας κοντά στο μηδέν και θα θεωρούνται ότι είναι όμοιες, ενώ οι μεταβλητές με μεγάλη αρνητική συσχέτιση θα θεωρούνται πολύ ανόμοιες ή ότι απέχουν αρκετά. Αν θελήσουμε να χρησιμοποιήσουμε τις ομοιότητες (similarities) στον πίνακα $\mathbf{D} = \left[d_{fg} \right]$ πάλι θα πρέπει να μετασχηματίσουμε τα παραπάνω μέτρα αφού με τον τρόπο που οριστήκαν παίρνουν και αρνητικές τιμές. Αν βασιστούμε στη σχέση: $s(f, g) = (1 + R(f, g))$ τα ζεύγη μεταβλητών με μεγάλες τιμές για το $R(f, g)$ θα θεωρούνται όμοια στον νέο πίνακα ομοιότητας. Αυτό άλλωστε επιβεβαιώνεται και από τη σχέση: $d(f, g) = 1 - s(f, g)$.

Είναι γνωστό πως ο συντελεστής συσχέτισης του Pearson και το συνημίτονο της γωνίας μεταξύ δύο μεταβλητών μένουν αναλλοίωτα σε γραμμικούς μετασχηματισμούς. Ωστόσο τα μέτρα αυτά έχουν δεχτεί ορισμένες κριτικές. Σύμφωνα με τον Cox et al. (2001), οι συντελεστές ομοιότητας θα πρέπει να μένουν αναλλοίωτοι στους μετασχηματισμούς των

μεταβλητών. Για τον λόγο αυτό πρέπει να υπάρχει ένα εναλλακτικό μέτρο. Οι κλίμακες των μεταβλητών που μελετήθηκαν είναι η απόλυτη (absolute), η κλίμακα διαφορών, η ratio και η διαστηματική κλίμακα. Οι μεταβλητές μετασχηματίζονται στην ομοιόμορφη (uniformity) κλίμακα, κατά σειρά, από την απόλυτη, τη διαφορών, τη ratio κλίμακα και τη διαστηματική σύμφωνα με τον τύπο

$$u_{iz} = y_{iz}$$
$$u_{iz} = y_{iz} - \overline{y}_z$$
$$u_{iz} = \left(\frac{1}{n} \sum_i y_{is}^2 \right)^{-1/2} y_{iz}$$
$$u_{iz} = \left(\frac{1}{n-1} \sum_i \left(y_{is} - \overline{y}_s \right)^2 \right)^{-1/2} \left(y_{iz} - \overline{y}_z \right)$$

Έτσι ένα εναλλακτικό μέτρο ομοιότητας αντί του συντελεστή συσχέτισης όπως δείξαμε πριν είναι το παρακάτω

$$s(f,g) = \frac{2\sum_i u_{if} u_{ig}}{\left(\sum_i u_{if}^2 + \sum_i u_{ig}^2 \right)}$$

3.4 Αλγόριθμοι και Μέθοδοι Ομαδοποίησης

Στις προηγούμενες παραγράφους αναφερθήκαμε στο πως μπορούμε να χειριστούμε πολυμεταβλητά δεδομένα με διαφορετικές κλίμακες καθώς και στον τρόπο με τον οποίο υπολογίζουμε μέτρα ομοιότητας ή ανομοιότητας για τις στήλες του πίνακα $\mathbf{Y} = [y_{ij}]$ ώστε να κατασκευάσουμε έναν πίνακα ανομοιότητας $\mathbf{D} = \left[d_{fg} \right]$. Το επόμενο στάδιο της ανάλυσης κατά συστάδες έχει να κάνει με την πληροφορία που παίρνουμε από τον πίνακα ανομοιότητας την οποία χρησιμοποιούμε για να δημιουργήσουμε ομοιογενείς ομάδες. Το τελευταίο αφορά την επιλογή μεθόδου ομαδοποίησης και ενός κατάλληλου αλγόριθμου υλοποίησης της.

Οι μέθοδοι ομαδοποίησης των αντικειμένων κατατάσσονται σε δύο μεγάλες κατηγορίες: στις ιεραρχικές και στις μη-ιεραρχικές μεθόδους. Επιπλέον, για κάθε μέθοδο, υπάρχουν διάφοροι υπολογιστικοί αλγόριθμοι για να δημιουργήσουμε τις ομάδες. Ανάλογα με το είδος των δεδομένων και τον σκοπό που πραγματοποιούμε την ανάλυση είναι προτιμότερο να δοκιμάζουμε αρκετούς αλγόριθμους για να βρούμε την καλύτερη λύση. Συνοπτικά, οι μη-ιεραρχικές μέθοδοι ομαδοποίησης έχουν ως στόχο να δημιουργήσουν k συστάδες, με $k \le n$, όπου ο αριθμός των συστάδων είναι προκαθορισμένος από τον αναλυτή. Οι πιο διαδεδομένοι

αλγόριθμοι για τις μη-ιεραρχικές μεθόδους είναι ο k-means του MacQueen, ο k-medoid και η fuzzy analysis. Επιπλέον, οι μη-ιεραρχικές μέθοδοι είναι κατάλληλες μόνο για την ομαδοποίηση των γραμμών (παρατηρήσεις). Από την άλλη, οι ιεραρχικές μέθοδοι δεν κατασκευάζουν μια μόνο διαμέριση με k συστάδες αλλά παράγουν μια ακολουθία από συστάδες όπου στα διάφορα στάδια ο αριθμός k παίρνει όλες τις δυνατές τιμές.

Υπάρχουν δύο είδη ιεραρχικών μεθόδων: οι συσσωρευτικές (agglomerative) και οι διαιρετικές (divisive). Οι συσσωρευτικές μέθοδοι ταξινομούνται σε τρεις κατηγορίες: Συνένωσης (Linkage), Κέντρου βάρους (Centroid) και Ελαχιστοποίησης της Διακύμανσης (Error Variance Methods). Οι παραπάνω μέθοδοι χρησιμοποιούνται για την ομαδοποίηση των γραμμών. Μόνο η μέθοδος της Συνένωσης και του Ward (Variance Method ή Incremental Sums of Squares) είναι κατάλληλες για την ομαδοποίηση μεταβλητών (στήλες). Παρόμοια, οι διαιρετικές μέθοδοι μπορούν να χρησιμοποιηθούν και στις δύο περιπτώσεις ωστόσο η διαφορά τους με τις ιεραρχικές μεθόδους είναι στον τρόπο που τα στοιχεία ενώνονται και δημιουργούν την συστάδα.

Στα επόμενα θα αναφερθούμε μόνο στις ιεραρχικές μεθόδους ομαδοποίησης και κυρίως σε εκείνες που χρησιμοποιούνται για την ομαδοποίηση μεταβλητών. Για την ερμηνεία και ανάλυση των μη-ιεραρχικών μεθόδων προτείνεται στον αναγνώστη ο Lewis (1985) ενώ για σύγκριση των ιεραρχικών μεθόδων ο Romesburg (1984) και οι Punj and Stewart (1983).

3.4.1 Μέθοδος της Απλής Συνένωσης

Γενικά στις ιεραρχικές μεθόδους ζητάμε από τον αλγόριθμο να μας κατασκευάσει ένα δενδροδιάγραμμα το οποίο να περιέχει όλες τις τιμές για το k. Από την μία πλευρά του δενδροδιαγράμματος υπάρχουν k συστάδες που περιέχουν ένα μόνο αντικείμενο ($k = n$) ενώ από την άλλη υπάρχει μία συστάδα που περιέχει και τα n αντικείμενα ($k = 1$). Η μέθοδος της απλής συνένωσης είναι η παλαιότερη και η πιο απλή. Για τον αλγόριθμο υλοποίησης της η απόσταση ανάμεσα σε δύο συστάδες R και Q, που δεν είναι η ίδια με τον συντελεστή ανομοιότητας d παραπάνω, ορίζεται ως $d(R,Q) = \min d(i,j) \quad \forall i \in R, j \in Q$. Από τον ορισμό της φαίνεται ότι η απόσταση ανάμεσα σε δύο συστάδες αντιστοιχεί στην μικρότερη όλων των αποστάσεων των ζευγαριών. Όταν οι συστάδες A και B ενωθούν για να δημιουργήσουν μια νέα συστάδα R, όλες οι ανομοιότητες ανάμεσα στην R και σε άλλες συστάδες Q προκύπτουν από τον παρακάτω ανανεωτικό τύπο

$$d(R,Q) = \min\{d(A,Q), d(B,Q)\} = \frac{1}{2}\big(d(A,Q) + d(B,Q)\big) - \frac{1}{2}\big|d(A,Q) - d(B,Q)\big|$$

Μια σημαντική αδυναμία της μεθόδου είναι το γεγονός ότι όταν δύο συστάδες έρθουν πολύ κοντά σε κάποιο σημείο αναγκαστικά θα ενωθούν. Οι συστάδες αυτές δεν θα είναι κατά ανάγκη όμοιες. Αυτό ονομάζεται το φαινόμενο της αλυσίδας λόγω του γεγονότος ότι άσχημα διαχωριζόμενες συστάδες ενώνονται. Τότε οι συστάδες που δημιουργούνται έχουν ένα μακρύ σχήμα σχεδόν γραμμικό, αντί για το κλασσικό σχήμα μπάλας.

3.4.2 Μέθοδος της Πλήρους Συνένωσης

Η μέθοδος αυτή μπορεί να θεωρηθεί ότι λειτουργεί αντίθετα με την προηγούμενη μέθοδο. Τώρα η απόσταση ανάμεσα σε δύο συστάδες ορίζεται ως η μεγαλύτερη ανομοιότητα μεταξύ των στοιχείων της μιας συστάδας και μιας άλλης. Δηλαδή είναι: $d(R,Q) = \max d(i,j)$ $\forall i \in R, j \in Q$. Η ανανεωτική φόρμουλα για τον υπολογισμό όλων των αποστάσεων ή ανομοιοτήτων τώρα γίνεται

$$d(R,Q) = \max\{d(A,Q), d(B,Q)\} = \frac{1}{2}\big(d(A,Q) + d(B,Q)\big) + \frac{1}{2}\big|d(A,Q) - d(B,Q)\big|$$

Το μειονέκτημα της μεθόδου ως προς τον σχηματισμό των συστάδων είναι ότι φτιάχνει πολλές και συμπαγείς συστάδες με αποτέλεσμα να έχουν μικρή διάμετρο. Επιπλέον από τον ορισμό της, ενδέχεται κάποιες συστάδες να περιέχουν τουλάχιστον ένα απομακρυσμένο ζεύγος στοιχείων με αποτέλεσμα αυτές να ενωθούν πολύ αργότερα. Έτσι σχετικά όμοια στοιχεία θα μείνουν ασύνδετα για αρκετό χρόνο. Η μορφή των τελικών συστάδων δεν είναι πολύ καλή παρόλο που είναι σχετικά μπαλοειδής αφού δεν διαχωρίζονται εύκολα.

3.4.3 Μέθοδος της Μέσης Ομάδας ή των Μη Σταθμισμένων Μέσων

Αν και οι δύο προηγούμενες μέθοδοι μοιάζουν, η μέθοδος της μέσης ομάδας (group average) είναι πιο εύχρηστη διότι δεν χρησιμοποιεί καθόλου σταθμά. Ο αλγόριθμος υλοποίησης της λειτουργεί ως εξής: έστω ότι έχουμε δύο συστάδες R, Q και ότι $|R|, |Q|$ συμβολίζουν τον αριθμό των στοιχείων που καθεμιά τους περιέχει. Τότε ο συντελεστής ανομοιότητα $d(R,Q)$ ανάμεσα στις συστάδες R και Q είναι η μέση τιμή όλων των $d(i,j)$. Οπότε έχουμε τη σχέση:

$$d(R,Q) = \frac{1}{|R||Q|}\sum_{i,j} d(i,j)$$

Ο ανανεωτικός τύπος της μεθόδου έχει την μορφή

$$d(R,Q)=\frac{|A|}{|R|}d(A,Q)+\frac{|B|}{|R|}d(B,Q).$$

Στον παραπάνω ανανεωτικό τύπο οι $d(A,Q)$, $d(B,Q)$ βρίσκονται από τον πίνακα με τις προηγούμενες ανομοιότητες ενώ οι ανομοιότητες που δεν αφορούν την συστάδα R παραμένουν ίδιες. Μια σημαντική ιδιότητα της μεθόδου είναι η ύπαρξη μονοτονίας των ανομοιοτήτων για μια συστάδα που ενώνεται με μία άλλη και αυτό γιατί οι αποστάσεις των επόμενων ενώσεων είναι πάντοτε μεγαλύτερες από τις προηγούμενες ενώσεις.

Εκτός από τα δενδροδιαγράμματα που απεικονίζουν τις ενώσεις των συστάδων σε κάθε βήμα, τα αποτελέσματα της κάθε μεθόδου μπορούμε να τα δούμε και από τα *banner plots*. Το συνολικό πλάτος ενός *banner* έχει μεγάλη σημασία γιατί μας δίνει μια γενική εικόνα της κατασκευής από τον αλγορίθμου που χρησιμοποιήθηκε. Γενικά όσο πιο μακρύ είναι το *banner* για κάθε στοιχείο, συμβολικά $l(i)$ για το μήκος του, τόσο μεγαλύτερος είναι και ο συσσωρευτικός συντελεστής (AC) για τα δεδομένα. Ο AC μετριέται σε $0-1$ κλίμακα και ορίζεται ως

$$\mathrm{AC}=\frac{1}{n}\sum_{i=1}^{n}l(i)$$

Ωστόσο με σχετική επιφύλαξη πρέπει να δεχόμαστε την τιμή του, διότι τιμές κοντά στη μονάδα δεν αποδεικνύουν με βεβαιότητα ότι η κατάλληλη ομαδοποίηση έχει επιτευχθεί, αυτό θα το κρίνει κυρίως ο αναλυτής, αφού είναι αρκετά ευαίσθητος σε έκτροπες παρατηρήσεις με αποτέλεσμα να μεγαλώνουν αρκετά την τιμή του.

3.4.4 Μέθοδος του Ward

Με την μέθοδο αυτή ελαχιστοποιούμε την ποσότητα:

$$d\left(C_i,C_j\right)=\frac{\left(\overline{\mathbf{x}}_i-\overline{\mathbf{x}}_j\right)\left(\overline{\mathbf{x}}_i-\overline{\mathbf{x}}_j\right)'}{1/n_i+1/n_j}$$

όπου $\overline{\mathbf{x}}_i$ και $\overline{\mathbf{x}}_j$ είναι οι μέσοι των διανυσμάτων C_i, C_j και n_i, n_j είναι το πλήθος των στοιχείων των δύο διανυσμάτων. Σκοπός είναι να ενώσουμε τις ομάδες οι οποίες οδηγούν στην ελαχιστοποίηση του συνολικού αθροίσματος των αποστάσεων. Ως συστάδα ορίζεται μια ομάδα της οποίας η διακύμανση των στοιχείων της είναι σχετικά μικρή.

3.4.5 Διαιρετικές Μέθοδοι

Οι διαιρετικές μέθοδοι είναι στη φύση τους ιεραρχικές. Η κύρια διαφορά τους έγκειται στο ότι η διαδικασία εύρεσης ομάδων ξεκινάει από την αντίθετη κατεύθυνση. Αρχικά όλα τα στοιχεία αποτελούν μια ομάδα και έπειτα ο αλγόριθμος τα διαιρεί σε δύο συστάδες. Σε κάθε επόμενο βήμα κάθε συστάδα διαιρείται σε νέες συστάδες και η διαδικασία συνεχίζεται έως ότου δημιουργηθούν n συστάδες αποτελούμενες από ένα μόνο στοιχείο. Τα αποτελέσματα της μεθόδου παριστάνονται όπως και στις ιεραρχικές μεθόδους από το δενδροδιάγραμμα αλλά γενικά δεν αναμένονται να είναι ίδια.

Οι διαιρετικές μέθοδοι διακρίνονται σε δύο είδη: στις μονοθετικές και στις πολυθετικές. Στις μονοθετικές κάθε διαίρεση μιας συστάδας γίνεται βάσει μιας μεταβλητής σε αντίθεση με τις πολυθετικές όπου και οι p μεταβλητές συμμετέχουν ταυτόχρονα. Εμείς θα αναφερθούμε στις πολυθετικές και θα δώσουμε έναν αλγόριθμο για την υλοποίηση της μεθόδου. Οι διαιρετικές μέθοδοι έχουν δεχτεί κριτικές σχετικά με τον υπολογιστικό χρόνο που απαιτείται για την διαμέριση παρόλα αυτά ο αλγόριθμος των MacNaughton-Smith et al. (1964), παρέχει καλά αποτελέσματα χωρίς μεγάλο φόρτο χρόνου υλοποίησης.

Ο αλγόριθμος ξεκινάει διαιρώντας όλα τα στοιχεία που υπάρχουν στην ομάδα σε δύο νέες ομάδες χωρίς όμως να εξετάζονται όλες οι πιθανές διαιρέσεις αλλά κατά κάποιο τρόπο χρησιμοποιώντας μια επαναληπτική διαδικασία. Έστω ότι R είναι η αρχική συστάδα και A, B οι δύο επόμενες που προέρχονται από την διαίρεση του R. Αρχικά μετακινούμε ένα στοιχείο από το A στο B, με $A \equiv R$, για το οποίο ισχύει ότι είναι το περισσότερο ανόμοιο, δηλαδή έχει την μεγαλύτερη κατά μέσο όρο ανομοιότητα από όλα τα υπόλοιπα στοιχεία. Αυτό μπορεί να δειχτεί ως εξής

$$d\left(i, A\setminus\{i\}\right) = \frac{1}{|A|-1}\sum_{\substack{j\in A\\ j\neq i}} d(i,j) \;\Rightarrow\; A' = A\setminus\left\{i'\right\} \text{ και } B' = B\cup\left\{i'\right\}$$

Η διαδικασία συνεχίζεται και υπολογίζουμε την μέση ανομοιότητα για την μεγαλύτερη συστάδα και την συγκρίνουμε με τη μέση ανομοιότητα των στοιχείων της άλλης συστάδας με σκοπό να διώξουμε επιπλέον στοιχείο. Δηλαδή

$$d\left(i, A\setminus\{i\} - d(i,B)\right) = \frac{1}{|A|-1}\sum_{\substack{j\in A\\ j\neq i}} d(i,j) - \frac{1}{|B|}\sum_{h\in B} d(i,h)$$

Η διαδικασία αυτή συνεχίζεται μέχρι η διαφορά αυτή να γίνει αρνητική ή μηδέν. Τότε για να αποφασίσουμε ποια συστάδα ξανά θα διαμερίσουμε βασιζόμαστε στο κριτήριο της

διαμέτρου, δηλαδή της μεγαλύτερης ανομοιότητας ανάμεσα σε δύο στοιχεία. Άρα για κάθε συστάδα Q θα ισχύει ότι

$$\text{diam}(Q) = \max d(j,h)$$

Ο αλγόριθμος τερματίζεται όταν μείνουμε με συστάδες που να αποτελούνται από ένα μόνο στοιχείο. Όπως και στην περίπτωση των συσσωρευτικών μεθόδων μπορούμε να δούμε τον ποιότητα της ομαδοποίησης από το *banner plot* και τον αντίστοιχο συντελεστή. Για κάθε στοιχείο i μετράμε το μήκος $l(i)$ βάσει μιας τυποποιημένης κλίμακας στο $0-1$. Άρα και πάλι ο διαιρετικός συντελεστής (DC) ορίζεται από τη σχέση

$$\text{DC} = \frac{1}{n}\sum_{i=1}^{n} l(i)$$

3.4.6 Σχόλια περί των Μεθόδων

I. Οι ιεραρχικές μέθοδοι πάσχουν από το γεγονός ότι δεν μπορούμε να διορθώσουμε ότι έχει γίνει στα προηγούμενα βήματα,

II. Όταν οι συσσωρευτικές μέθοδοι ενώσουν δύο στοιχεία δεν είναι δυνατό να χωριστούν και όταν οι διαιρετικές μέθοδοι διαμερίσουν δύο στοιχεία δεν μπορούν ποτέ να ενωθούν,

III. Οι διαιρετικές μέθοδοι φτιάχνουν μπαλοειδής συστάδες περίπου στη μορφή που φτιάχνει και η μέθοδος της μέσης ομάδας οπότε τα αποτελέσματα των δύο μεθόδων μπορούν να συγκριθούν,

IV. Η μέθοδος της απλής συνένωσης αγνοεί τις έκτροπες παρατηρήσεις και τις μεταχειρίζεται σαν απομονωμένα στοιχεία μέχρι τα τελευταία βήματα της διαδικασίας. Για τον λόγο αυτό η μέθοδος της μέσης ομάδας και του Ward είναι προτιμότερες από τις δύο μεθόδους συνένωσης διότι είναι αρκετά ευαίσθητες στις ακραίες ή έκτροπες παρατηρήσεις. Ωστόσο οι δύο μέθοδοι της Συνένωσης λόγω της ιδιότητας τους να παραμένουν αναλλοίωτοι σε μονότονους μετασχηματισμούς των συντελεστών ανομοιότητας είναι αρκετά χρήσιμοι για δεδομένα με διατάξιμη κλίμακα,

V. Σχεδόν όλες οι γνωστές ιεραρχικές μέθοδοι ομαδοποίησης είναι παραλλαγές ενός ανανεωτικού τύπου των Lance and Williams (1967), που ονομάζεται μέθοδος της ευέλικτης στρατηγικής (Flexible Strategy).

3.5 Αξιολόγηση της Μεθόδου και Επιλογή Συστάδων

Όταν ο αναλυτής καταλήξει με την τελική ομαδοποίηση και αφού έχουν προηγηθεί τα βήματα που περιγράψαμε προηγουμένως, τότε η όλη διαδικασία που ακολουθήθηκε θα χρειαστεί να ελεγχθεί για τα αποτελέσματα που έδωσε. Δηλαδή χρειάζεται κατά κάποιο τρόπο να επιβεβαιώσουμε τον διερευνητικό χαρακτήρα της μεθόδου.

Στην παράγραφο αυτή θα παρουσιάσουμε κάποιες τεχνικές για την αξιολόγηση της ποιότητας των αποτελεσμάτων της μεθόδου ομαδοποίησης και προτάσεις για την επιλογή του κατάλληλου αριθμού των συστάδων. Στον αναγνώστη προτείνεται να μελετήσει τους Rohlf, Anderberg, Milligan and Cooper, Haldiki, Hennig, μεταξύ άλλων, οι οποίοι αναφέρουν λεπτομερώς πολλά στοιχεία για το θέμα αυτό. Κυρίως ενδιαφερόμαστε για:

- Υπολογισμό του συντελεστή συσχέτισης μεταξύ του αρχικού πίνακα ανομοιοτήτων (proximity matrix) και του πίνακα που προκύπτει από τις αποστάσεις μετά την ένωση των στοιχείων για τη κατασκευή του δενδροδιαγράμματος. Ο πίνακας αυτός ονομάζεται derived matrix. Ο συντελεστής που προκύπτει λέγεται cophenetic correlation και μας δίνει ένα μέτρο καλής προσαρμογής ανάμεσα στο αποτέλεσμα της ιεραρχικής ομαδοποίησης και στον αρχικό πίνακα αποστάσεων. Τιμές του συντελεστή κοντά στην μονάδα είναι δείγμα πολύ καλής ομαδοποίησης αν και η κλίμακα των δεδομένων επηρεάζει αισθητά την τιμή του. Άρα μπορεί να χρησιμοποιηθεί και ως μέτρο σύγκρισης για διαφορετικούς αλγόριθμους ομαδοποίησης για τα ίδια δεδομένα.
- Μια υπόθεση που γίνεται για την αξιολόγηση της μεθόδου συγκρίνει την πραγματική ομαδοποίηση εάν είναι γνωστή, με εκείνη που πήραμε εμείς. Αν όμως είναι άγνωστη η πραγματική ομαδοποίηση τότε μεταξύ δύο διαφορετικών αλγορίθμων υλοποίησης μπορεί να γίνει η σύγκριση και να βρεθεί ο βαθμός ομοιότητας τους. Αν ο αριθμός των συστάδων δύο διαφορετικών μεθόδων είναι ίδιος, τότε υπολογίζεται ο συντελεστής κ του Cohen για την συμφωνία τους, προσαρμοσμένος για τυχαίους παράγοντες. Αν οι δύο μέθοδοι δώσουν διαφορετικά αποτελέσματα τότε ο Rand index, I_{HA} χρησιμοποιείται για τον βαθμό συμφωνίας τους με τιμές μεταξύ 0 και 1, με $0 =$ καμία συμφωνία και $1 =$ πλήρης συμφωνία (Hubert and Arabie, 1985). Ανάμεσα στους δυο δείκτες ο I_{HA} είναι καλύτερος.
- Ένα ακόμη μέτρο είναι ο υπολογισμός του συντελεστή συσχέτισης ανάμεσα στον αρχικό πίνακα ανομοιοτήτων και στον πίνακα με τιμές 0 ή 1 του οποίου οι τιμές αυτές αφορούν

τα ζευγάρια των στοιχείων που πρέπει να ανήκουν στην ίδια συστάδα βάσει του αρχικού πίνακα αποστάσεων ή όχι. Ο συντελεστής αυτός λέγεται point-biserial ή Hubert gamma. Τιμές κοντά στη μονάδα δείχνουν ότι τα ζευγάρια που κωδικοποιήθηκαν με την τιμή 0 ήταν όντως περίπου όμοια και αντίστροφα.

- Παρόμοιο μέτρο με το προηγούμενο είναι και ο συντελεστής gamma ή gamma coefficient of concordance W. Τιμές κοντά στην μονάδα δείχνουν συμφωνία μεταξύ του αρχικού πίνακα ανομοιοτήτων και της τελικής ομαδοποίησης των στοιχείων.
- Ο αριθμός των συστάδων μπορεί να καθοριστεί *a priori* από τον αναλυτή, ειδικά στην περίπτωση που έχει κάποια σχετική γνώση για το αντικείμενο της έρευνας ή αργότερα μελετώντας το δενδροδιάγραμμα της μεθόδου.
- Μια γνωστή τεχνική για την επιλογή του αριθμού των συστάδων k είναι να χρησιμοποιηθεί η διαφορά των αποστάσεων σε κάθε συστάδα $\Delta S_t = S_{t-1} - S_t$. Εν ολίγης καθώς ο αριθμός των βημάτων αυξάνει από μία συστάδα για κάθε στοιχείο μέχρι μια τελική συστάδα με όλα τα στοιχεία μαζί, αυξάνεται παράλληλα και η απόσταση S. Οπότε όταν το k μειώνεται από $k = n$ σε $k = 1$ επιλέγεται τότε εκείνο όπου το ΔS γίνεται μέγιστο, δηλαδή η $k+1$ συστάδα. Επιπλέον, μπορούν να γίνουν και ορισμένα τεστ βασισμένα στον πίνακα ανομοιοτήτων για τον αριθμό των συστάδων αλλά δεν θα επεκταθούμε περισσότερο σε αυτά.

Κλείνοντας το κεφάλαιο αυτό θα πρέπει να αναφέρουμε πως η διαφορά της μεθόδου σε σχέση με άλλες τεχνικές της πολυμεταβλητής ανάλυσης, έγκειται στο ότι με την ανάλυση κατά συστάδες για να βρεθούν οι ομάδες δεν χρειάζεται να διακρίνουμε μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών οπότε δεν γίνεται λόγος για στατιστική συμπερασματολογία. Όμως είναι πολύ πιθανό να χρησιμοποιήσουμε κάποιο πιθανοθεωρητικό μοντέλο για την ομαδοποίηση (model-based clustering) και τότε η προσέγγιση είναι τελείως διαφορετική, ενώ για την επιλογή του αριθμού των συστάδων βασιζόμαστε σε κριτήρια πιθανοφάνειας αφού μπορούμε να προβούμε σε στατιστική συμπερασματολογία.

3.6 Εφαρμογή

Χρησιμοποιώντας για την εφαρμογή της μεθόδου το ερωτηματολόγιο της Τράπεζας ABC θα ομαδοποιήσουμε τις μεταβλητές-ερωτήσεις σε κατάλληλα υποσύνολα μικρότερης διάστασης κάνοντας ξανά την παραδοχή ότι το ερωτηματολόγιο δεν είναι δομημένο (οι ερωτήσεις είναι σκόρπιες) και θα δείξουμε με μια διαφορετική προσέγγιση τον τρόπο εύρεσης ομοιογενών ομάδων για τη δημιουργία δομής σε ένα σύνολο αντικειμένων. Επειδή η ανάλυση κατά συστάδες θα χρησιμοποιηθεί για την ομαδοποίηση των στηλών, τα αποτελέσματα θα τα συγκρίνουμε με εκείνα που πήραμε από την εφαρμογή του Κεφαλαίου 2 όταν χρησιμοποιήσαμε την παραγοντική ανάλυση για το ερωτηματολόγιο. Άρα πέρα από τις συστάδες που θα βρούμε, θα μπορέσουμε να πούμε με κάποια βεβαιότητα το κατά πόσο η παραγοντική ανάλυση και η ανάλυση κατά συστάδες δίνουν συγκρίσιμα αποτελέσματα. Το θέμα αυτό έχει επίσης συζητηθεί στο εισαγωγικό κεφάλαιο της εργασίας αυτής.

Ιδανικά θα είναι, να μπορέσουμε να δημιουργήσουμε τόσες συστάδες όσοι και οι παράγοντες που να περιέχουν τον ίδιο αριθμό στοιχείων, άρα και την ίδια φυσική ερμηνεία. Με βάση αυτό, γνωρίζουμε ότι η τελική ομαδοποίηση περιέχει τέσσερεις συστάδες και το οποίο μας βοηθάει να ελέγξουμε για την αξιοπιστία της ομαδοποίησης και των αλγόριθμων υλοποίησης. Επίσης, τις συστάδες μπορούμε να τις μελετήσουμε ευκολότερα σε σχέση με ένα ανομοιογενές πολυμεταβλητό σύνολο. Για παράδειγμα, αν ενδιαφερόμαστε για την μελέτη των σχέσεων των μεταβλητών μέσα στη συστάδα γνωρίζοντας πως τα στοιχεία που την απαρτίζουν έχουν μεγάλη συσχέτιση μεταξύ τους, μπορούμε να φτιάξουμε πίνακες και να κατασκευάσουμε μοντέλα για τον λόγο αυτό. Επίσης ως επακόλουθο της μεθόδου, είναι δυνατόν να γίνουν νέες υποθέσεις μεταξύ των συστάδων ως νέα υποσύνολα μεταβλητών μικρότερης διάστασης, δηλαδή σαν νέα σετ μεταβλητών, για τα οποία ενδιαφερόμαστε να μελετήσουμε τις μεταξύ τους σχέσεις μέσω της canonical correlation analysis.

Η ανάλυση αυτή εστιάζεται στις στήλες (ερωτήσεις) του ερωτηματολογίου διάστασης $\mathbf{Y}(149\times 72)$. Πριν ξεκινήσουμε ελέγχουμε αν η κλίμακα των μεταβλητών είναι ομοιογενής αλλιώς την επιβάλουμε εμείς, για να είμαστε σίγουροι ότι τα αποτελέσματα που θα πάρουμε από την ομαδοποίηση θα είναι ρεαλιστικά (αξιόπιστα). Όμως οι ερωτήσεις, όπως γνωρίζουμε, είναι όλες κατηγορικές μεταβλητές σε πενταβάθμια διατάξιμη κλίμακα κάτι που σημαίνει ότι δεν τίθεται θέμα αλλαγής της κλίμακας. Για να ξεκινήσει η διαδικασία ομαδοποίησης πρέπει να κατασκευάσουμε τον πίνακα ανομοιότητας $\mathbf{D}=\left[d_{fg}\right]$ βάσει στον οποίο θα υπολογίσουμε

ένα μέτρο ομοιότητας το οποίο θα το μετατρέψουμε κατάλληλα για να μπορεί να υπολογιστεί σαν απόσταση. Όπως έχει αναφερθεί, για τις μεταβλητές που είναι σε διατάξιμη κλίμακα μπορούμε να υπολογίσουμε μη παραμετρικούς συντελεστές συσχέτισης βασισμένους στα ranks αντί του συντελεστή του Pearson, εκτός και αν η συσχέτιση μεταξύ των τιμών τους είναι πολύ κοντά (περίπου στο 1.00). Στον Πίνακα 3.1 που ακολουθεί μπορούμε να δούμε για τρεις συντελεστές συσχέτισης, του Pearson (P), του Spearman (S) και του Kendall (K), το βαθμό συσχέτισης μεταξύ τους στις τιμές που υπολογίζουν για τα δεδομένα.

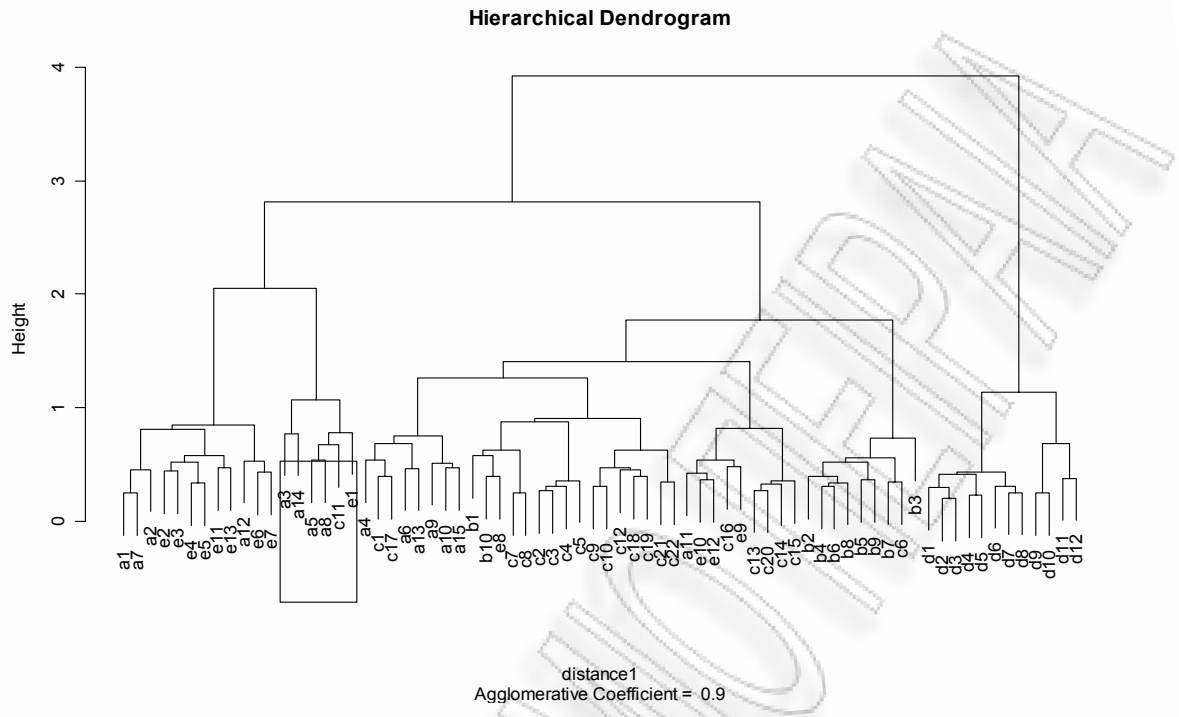
Πίνακας 3.1: *Συσχετίσεις μεταξύ των συντελεστών Pearson (P), Spearman (S) και Kendall (K).*

	Pearson (P)	Spearman (S)	Kendall (K)
Pearson (P)	1.0000000	0.9912212	0.9908747
Spearman (S)	0.9912212	1.0000000	0.9986683
Kendall (K)	0.9908747	0.9986683	1.0000000

Θα πρέπει να σημειώσουμε ότι για τον παραπάνω πίνακα υπολογισμού των συσχετίσεων αλλά και στη συνέχεια, αφαιρέσαμε με *listwise* δύο παρατηρήσεις διότι είχαν πολλές χαμένες απαντήσεις (missing values). Από τον Πίνακα 3.1 παρατηρούμε ότι μεταξύ τους οι τρεις συντελεστές είναι πάνω από 99% συσχετισμένοι όσον αφορά τα συγκεκριμένα δεδομένα, που σημαίνει ότι με όποιον συντελεστή από τους τρεις υπολογίσουμε τις συσχετίσεις οι αποκλίσεις στις τιμές θα είναι ασήμαντες. Άρα αφού υπολογίσουμε τις συσχετίσεις των μεταβλητών στη συνέχεια τις μετατρέπουμε σε συντελεστές ανομοιότητας $d(f,g)$ για τον πίνακα **D** και επιλέγουμε μέθοδο και αλγόριθμο ομαδοποίησης. Ύστερα εξετάζουμε την ισχύ της επιλεγμένης μεθόδου (εσωτερικά-εξωτερικά).

Όπως έχουμε αναφέρει για την ομαδοποίηση των μεταβλητών μπορούμε να διαλέξουμε ανάμεσα σε συσσωρευτικές και σε διαιρετικές μεθόδους. Για τα δεδομένα της ανάλυσης μας εξετάστηκαν προσεχτικά όλες οι ιεραρχικές μέθοδοι και τα καλύτερα αποτελέσματα τα πήραμε με την μέθοδο του Ward βασισμένη στις αποστάσεις του πίνακα ανομοιότητας **D**. Τα αποτελέσματα παρουσιάζονται στο δενδροδιάγραμμα του Γραφήματος 3.1 που ακολουθεί.

Γράφημα 3.1: *Δενδροδιάγραμμα για όλα τα δεδομένα με την μέθοδο του Ward.*



Με την ιεραρχική ομαδοποίηση όλη η διαδικασία αναπαρίσταται οπτικά με το παραπάνω δενδροδιάγραμμα. Το δενδροδιάγραμμα μας βοηθάει στο να επιλέξουμε αριθμό συστάδων ή την λύση που τελικά θα κρατήσουμε. Έτσι επιλέγουμε να πάρουμε επτά συστάδες μεταβλητών όπως φαίνεται και στον Πίνακα 3.2 όπου κάθε επίπεδο είναι και μία συστάδα.

Πίνακας 3.2: *Επτά συστάδες μεταβλητών από το δενδροδιάγραμμα του Γραφήματος 3.1.*

```
names(questions.ord[-c(4,116),])[cutagglomer1.1w1==1]
[1] "a1"  "a2"  "a7"  "a12" "e2"  "e3"  "e4"  "e5"  "e6"  "e7"  "e11" "e13"

names(questions.ord[-c(4,116),])[cutagglomer1.1w1==2]
[1] "a3"  "a5"  "a8"  "a14" "c11" "e1"

names(questions.ord[-c(4,116),])[cutagglomer1.1w1==3]
[1] "a4"  "a6"  "a9"  "a10" "a13" "a15" "c1"  "c17"

names(questions.ord[-c(4,116),])[cutagglomer1.1w1==4]
[1] "a11" "c13" "c14" "c15" "c16" "c20" "e9"  "e10" "e12"

names(questions.ord[-c(4,116),])[cutagglomer1.1w1==5]
[1] "b1"  "b10" "c2"  "c3"  "c4"  "c5"  "c7"  "c8"  "c9"  "c10" "c12" "c18"
"c19" "c21" "c22" "e8"
names(questions.ord[-c(4,116),])[cutagglomer1.1w1==6]
[1] "b2" "b3" "b4" "b5" "b6" "b7" "b8" "b9" "c6"

names(questions.ord[-c(4,116),])[cutagglomer1.1w1==7]
[1] "d1"  "d2"  "d3"  "d4"  "d5"  "d6"  "d7"  "d8"  "d9"  "d10" "d11" "d12"
```

Έχοντας υπόψη το Γράφημα 3.1 όπου ομαδοποιούνται όλες οι ερωτήσεις-μεταβλητές του ερωτηματολογίου έχουμε να σημειώσουμε τα εξής: Παρατηρήσαμε, πως όταν προσπαθήσαμε να ομαδοποιήσουμε και τις 72 μεταβλητές ανεξάρτητα με την μέθοδο που ακολουθήσαμε υπάρχει πάντοτε μία συστάδα η οποία παρουσιάζεται αναλλοίωτη. Οι μεταβλητές που είναι κυκλωμένες στο δενδροδιάγραμμα και με έντονο χρώμα στον Πίνακα 3.2 αποτελούν αυτή τη συστάδα και για τον λόγο αυτό θα πρέπει να τις εξετάσουμε. Κοιτώντας τις συσχετίσεις τους με τις υπόλοιπες μεταβλητές της ανάλυσης διαπιστώσαμε ότι είναι πολύ χαμηλές και δεν έχουν καθόλου διαχωριστική ισχύ, οπότε είναι προτιμότερο να τις αφαιρέσουμε. Αξίζει να θυμηθούμε ότι για την συγκεκριμένη ομάδα μεταβλητών ανάλογα είχαμε ενεργήσει και στην παραγοντική ανάλυση. Επιπλέον από το δενδροδιάγραμμα του Γραφήματος 3.1 κανείς θα μπορούσε να επιλέξει για τελική λύση τρεις συστάδες αποτελούμενες από τις μεταβλητές του Πίνακα 3.3 παρακάτω.

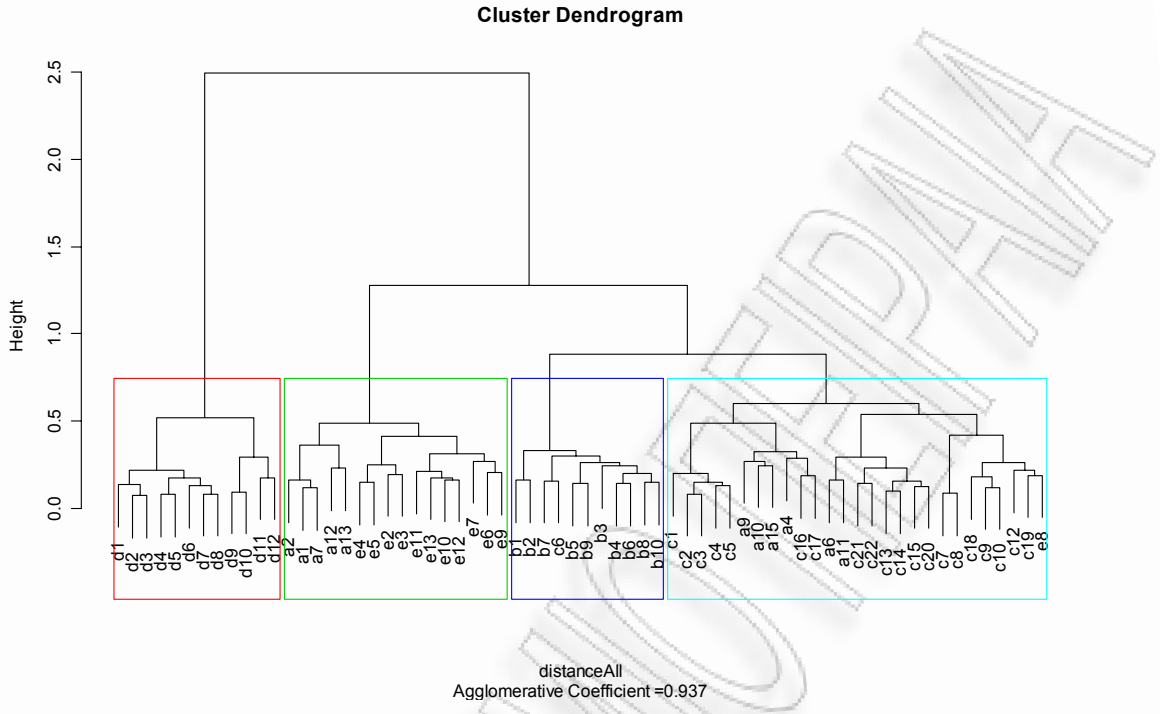
Πίνακας 3.3: *Τρεις συστάδες μεταβλητών από το δενδροδιάγραμμα του Γραφήματος 3.1.*

```
names(questions.ord[-c(4,116),])[cutagglomeRTria==1]
[1] "a1"  "a2"  "a3"  "a5"  "a7"  "a8"  "a12" "a14" "c11" "e1"  "e2"  "e3"
"e4"  "e5"  "e6"  "e7"  "e11" "e13"

names(questions.ord[-c(4,116),])[cutagglomeRTria==2]
[1] "a4"  "a6"  "a9"  "a10" "a11" "a13" "a15" "b1"  "b2"  "b3"  "b4"  "b5"
"b6" "b7"  "b8"  "b9"  "b10" "c1"  "c2" "c3"  "c4"  "c5"  "c6"  "c7"  "c8"
"c9" c10" "c12" "c13" "c14" "c15" "c16" "c17" "c18" "c19" "c20" "c21" "c22"
"e8"  "e9"  "e10" "e12"

names(questions.ord[-c(4,116),])[cutagglomeRTria==3]
[1] "d1"  "d2"  "d3"  "d4"  "d5"  "d6"  "d7"  "d8"  "d9"  "d10" "d11" "d12"
```

Βέβαια με την παραπάνω λύση για τον αριθμό των συστάδων που θα κρατήσουμε δεν φαίνεται να υπάρχει κάποια ιδιαίτερη ομαδοποίηση ώστε να μπορούμε να πούμε ότι έχουμε βρει ομοιογενής ομάδες μεταβλητών αφού δύσκολα θα μπορέσουμε να ερμηνεύσουμε τα αποτελέσματα. Η επιτυχία της μεθόδου θα είναι να «σπάσουμε» την μεγαλύτερη από τις τρεις συστάδες, δεύτερη, και αφού διώξουμε από την ανάλυση τις «προβληματικές» μεταβλητές $\{\alpha 3, \alpha 5, \alpha 8, \alpha 14, c11, \varepsilon 1\}$, να βρούμε τη δομή που υπάρχει στα δεδομένα. Σίγουρα όμως η τελική λύση θα είναι μεταξύ τεσσάρων και έξι ομάδων.

Η νέα ιεραρχική βέλτιστη ομαδοποίηση με τη χρήση της μεθόδου του Ward μας δίνει τα αποτελέσματα όπως παρουσιάζονται στο δενδροδιάγραμμα του Γραφήματος 3.2 και στον Πίνακα 3.4 όπου έχουμε επιλέξει ως τελική λύση τέσσερεις συστάδες.

Γράφημα 3.2 : *Δενδροδιάγραμμα της μεθόδου του Ward για την νέα ομαδοποίηση.*



Πίνακας 3.4: *Επιλογή ως λύση, τέσσερεις συστάδες μεταβλητών από το δενδροδιάγραμμα του Γραφήματος 3.2.*

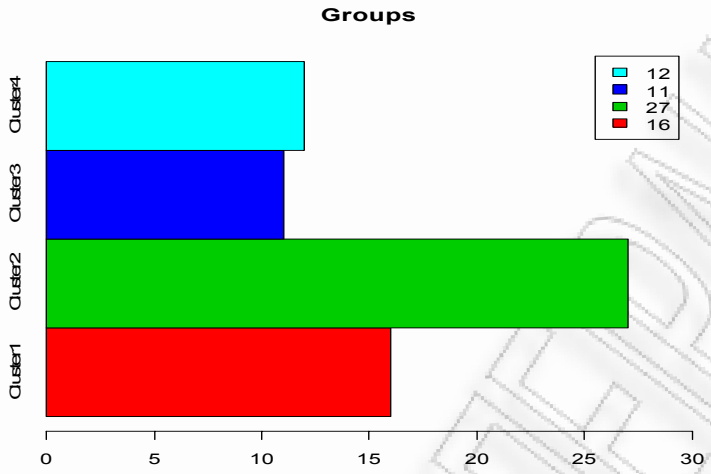
```
names(questions.ord[-c(4,116),-c(3,5,8,14,36,60)])[cuthierarchAll.1==1]
[1] "a1"  "a2"  "a7"  "a12" "a13" "e2"  "e3"  "e4"  "e5"  "e6"  "e7"  "e9"
"e10" "e11" "e12" "e13"

names(questions.ord[-c(4,116),-c(3,5,8,14,36,60)])[cuthierarchAll.1==2]
[1] "a4"  "a6"  "a9"  "a10" "a11" "a15" "c1"  "c2"  "c3"  "c4"  "c5"  "c7"
"c8"  "c9"  "c10" "c12" "c13" "c14" "c15" "c16" "c17" "c18" "c19" "c20"
"c21" "c22" "e8"

names(questions.ord[-c(4,116),-c(3,5,8,14,36,60)])[cuthierarchAll.1==3]
[1] "b1"  "b2"  "b3"  "b4"  "b5"  "b6"  "b7"  "b8"  "b9"  "b10" "c6"

names(questions.ord[-c(4,116),-c(3,5,8,14,36,60)])[cuthierarchAll.1==4]
[1] "d1"  "d2"  "d3"  "d4"  "d5"  "d6"  "d7"  "d8"  "d9"  "d10" "d11" "d12"
```

Από το Γράφημα 3.2 φαίνεται έχουμε επιλέξει ως λύση την επιλογή τεσσάρων συστάδων και αυτές παρουσιάζονται με τα στοιχεία τους στον Πίνακα 3.4. Η νέα ομαδοποίηση και η συγκεκριμένη λύση μοιάζουν να είναι ικανοποιητικές. Ένα δείγμα της ισχύς της συγκεκριμένης ομαδοποίησης εξασφαλίζεται από τον συντελεστή AC που στην περίπτωση μας είναι ίσος με 0.937, δηλαδή πολύ ισχυρή. Έτσι βλέπουμε ότι με 66 μεταβλητές αντί για 72, η ομαδοποίηση φαίνεται ότι βρήκε μια δομή για το ερωτηματολόγιο. Για τις τέσσερεις συστάδες έχουμε να πούμε τα εξής σε σχέση με την αρχική κατασκευή του ερωτηματολογίου:

- → Η αρχική κατασκευή του ερωτηματολογίου έχει πέντε ομάδες από ερωτήσεις. Εμείς δημιουργήσαμε τέσσερεις.
- → Δύο ομάδες ερωτήσεων τις εντόπισε η ανάλυση κατά συστάδες, την Β και Δ.
- → Από την Α ομάδα του «δομημένου» ερωτηματολογίου, «διώξαμε» τις 4 από τις 6 συνολικά ερωτήσεις. Συνολικά οι 6 ερωτήσεις αυτές, προφανώς, δεν ταιριάζουν με τις υπόλοιπες αφού έχουν σχεδόν μηδενικές συσχετίσεις και από πλευράς σχεδιασμού της έρευνας δεν προσφέρουν τίποτα περισσότερο αφού τις βρίσκουμε και αλλού με διαφορετική διατύπωση. Δηλαδή η πληροφορία που μας παρέχουν περιέχεται σε άλλες μεταβλητές και έτσι μπορεί να θεωρηθούν ότι είναι πλεονάζουσες.
- → Η ιεραρχική ομαδοποίηση χώρισε τις υπόλοιπες ερωτήσεις της Α ομάδας σε δύο συστάδες έτσι ώστε στη πρώτη συστάδα να ανήκουν οι ερωτήσεις που αφορούν τη στήριξη του συστήματος αξιολόγησης από την επιχείρηση και στη δεύτερη συστάδα οι ερωτήσεις της ομάδας Α που αφορούν περισσότερο διαδικαστικά θέματα της αξιολόγησης. Στην περίπτωση αυτή ίσως η α13 να ανήκει στη δεύτερη συστάδα.
- → Επίσης η ομαδοποίηση εντόπισε και την Γ κατηγορία του ερωτηματολογίου που είναι και η μεγαλύτερη και αφορά την συνάντηση αξιολόγησης. Σε αυτή την συστάδα (δεύτερη κατά σειρά στον Πίνακα 3.4) μπήκαν και κάποιες ερωτήσεις από την Α ομάδα που όπως είπαμε αφορούν κάποιες διαδικασίες που στηρίζεται η αξιολόγηση των εργαζομένων και μπορούν να προηγηθούν πριν τις ερωτήσεις της Γ ομάδας. Ένα θέμα είναι, η ερώτηση ε8 αν ανήκει σε αυτή τη συστάδα.
- → Επιπλέον, εντόπισε όλη την κατηγορία Ε και την ένωσε με τις υπόλοιπες από την Α για να φτιάξει τη συστάδα (πρώτη κατά σειρά στον Πίνακα 3.4) όπου συνδέει το σύστημα αξιολόγησης με τα κριτήρια στα οποία αξιολογούνται οι εργαζόμενοι.
- → Τέλος μπορούμε να πούμε ότι οι παραπάνω τέσσερεις συστάδες βρίσκονται σε αντιστοιχία με τους παράγοντες που κατασκευάσαμε στο Κεφάλαιο 2 σε μεγάλο βαθμό. Αυτό για την μέθοδο αποτελεί ένα κριτήριο της επιτυχίας της και δείχνει ότι στα δεδομένα υπάρχουν πραγματικά ομοιογενείς ομάδες τις οποίες οποιαδήποτε μέθοδος θα τις αναγνωρίσει.

Παρακάτω στο Γράφημα 3.3, δίνεται μια περιγραφική εικόνα του αριθμού των στοιχείων των συστάδων που η καθεμία περιέχει.

Γράφημα 3.3 : *Bar chart των συστάδων.*



Τώρα για την συγκεκριμένη μέθοδο ομαδοποίησης που εφαρμόσαμε θα δούμε ορισμένα κριτήρια για να επιβεβαιώσουμε την αξιοπιστία της. Κάποια από τα κριτήρια του Πίνακα 3.5 έχουν αναλυθεί νωρίτερα και όλα είναι εσωτερικά κριτήρια εκτός του Rand Index που είναι εξωτερικό.

Πίνακας 3.5 : *Εσωτερικά και εξωτερικά κριτήρια για την αξιοπιστία της μεθόδου και της επιλογής τεσσάρων συστάδων.*

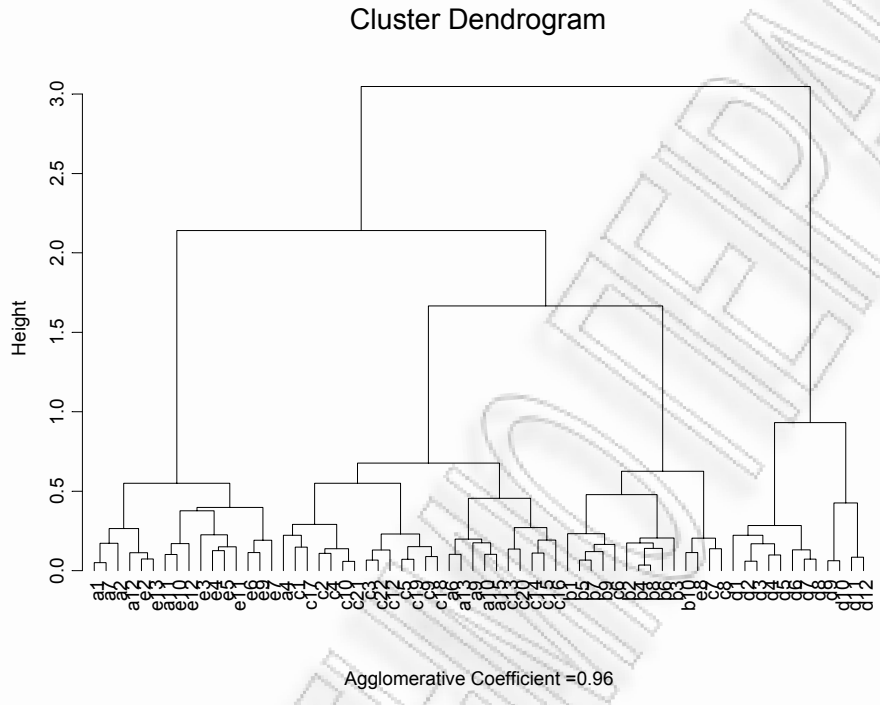
Cophenetic correlation	Gamma coefficient	Point-biserial (Hubert gam)	Δείκτης Dunn	Connectivity	Rand Index
0.677	0.778	0.575	0.386	20.416	0.792

Οι τιμές των παραπάνω κριτηρίων είναι ικανοποιητικές, αν αναλογιστούμε ότι έχουμε κατηγορικά δεδομένα για ανάλυση με αποτέλεσμα οι συσχετίσεις να μην είναι οι υψηλότερες σε απόλυτη τιμή. Δηλαδή μας δείχνουν ότι η μέθοδος ομαδοποίησης του Ward και η επιλογή τεσσάρων συστάδων είναι βέλτιστες για τα δεδομένα.

Θέλουμε τώρα να συγκρίνουμε άμεσα τα αποτελέσματα που βρήκαμε στην παραγοντική ανάλυση του προηγούμενου κεφαλαίου με αυτά που μας έδωσε η ανάλυση κατά συστάδες. Για τον λόγο αυτό χρησιμοποιούμε τα φορτία των παραγόντων από το ικανοποιητικό μοντέλο με τέσσερεις παράγοντες για την δημιουργία των συστάδων. Διαπιστώνουμε δηλαδή με έναν ακόμη τρόπο ότι η δομή που υπάρχει στα δεδομένα έχει αποκαλυφθεί αφού δεν υπάρχουν αντιφατικές λύσεις. Όλα συνιστούν ότι τέσσερεις ομάδες «φτιάχνουν» τα δεδομένα. Έτσι δείξαμε ότι η ανάλυση κατά συστάδες για την ομαδοποίηση μεταβλητών και γενικά για την μείωση της διάστασης των δεδομένων, μπορεί να χρησιμοποιηθεί ως συμπληρωματική της

παραγοντικής ανάλυσης. Τα αποτελέσματα παρουσιάζονται στο Γράφημα 3.4 και στον πίνακα που ακολουθεί.

Γράφημα 3.4 : *Δενδροδιάγραμμα της μεθόδου του Ward όταν η ομαδοποίηση έχει γίνει με βάση τους παράγοντες.*



Τα αποτελέσματα της ομαδοποίησης του Γραφήματος 3.4 συνιστούν ότι τέσσερεις συστάδες είναι μια λύση για τα δεδομένα. Επίσης φαίνεται ότι η ποιότητα της είναι πολύ καλή αφού ο AC=0.962 είναι πολύ υψηλός και οι ανομοιότητες μεταξύ των συστάδων είναι πολύ μεγαλύτερες από τις ανομοιότητες μεταξύ των στοιχείων εντός της συστάδας και το οποίο έχει σαν αποτέλεσμα οι γραμμές που ενώνουν τις συστάδες να είναι μακριές, ως εκ τούτου οδηγούμαστε στο συμπέρασμα ότι υπάρχουν τέσσερεις συστάδες καλά διακεκριμένες. Έτσι σε σύγκριση με το Γράφημα 3.2 οι διαφορές είναι ελάχιστες, ενώ οι τελικές συστάδες είναι ίδιες με τους παράγοντες όπως φαίνεται στον παρακάτω πίνακα και στο Κεφάλαιο 2.

```
names(questions.ord[-c(4,116),-c(3,5,8,14,36,60)])[cutagglom.fac.questnumAll==1]
[1] "d1"  "d2"  "d3"  "d4"  "d5"  "d6"  "d7"  "d8"  "d9"  "d10" "d11" "d12"

names(questions.ord[-c(4,116),-c(3,5,8,14,36,60)])[cutagglom.fac.questnumAll==2]
[1] "a4"  "a6"  "a9"  "a10" "a13" "a15" "c1"  "c2"  "c3"  "c4"  "c5"  "c9"
"c10" "c12" "c13" "c14" "c15" "c16" "c17" "c18" "c19" "c20" "c21" "c22"

names(questions.ord[-c(4,116),-c(3,5,8,14,36,60)])[cutagglom.fac.questnumAll==3]
[1] "b1"  "b2"  "b3"  "b4"  "b5"  "b6"  "b7"  "b8"  "b9"  "b10" "c6"  "c7"
"c8" "e8"

names(questions.ord[-c(4,116),-c(3,5,8,14,36,60)])[cutagglom.fac.questnumAll==4]
[1] "a1"  "a2"  "a7"  "a11" "a12" "e2"  "e3"  "e4"  "e5"  "e6"  "e7"  "e9"
"e10" "e11" "e12" "e13"
```

ΚΕΦΑΛΑΙΟ 4

Μοντέλα Συνάφειας Για Πίνακες

4.0 Γενικά

Αν και τα τελευταία χρόνια η ανάλυση πινάκων συνάφειας με τα λογαριθμογραμμικά μοντέλα θεωρείται μια πολύ ισχυρή μέθοδος για την αποκάλυψη των αλληλεπιδράσεων των μεταβλητών ιδιαίτερα για πολυδιάστατους πίνακες, ωστόσο, τα μοντέλα αυτά έχουν κάποιους σοβαρούς περιορισμούς. Οι περιορισμοί αυτοί αφορούν κυρίως τις κατηγορικές μεταβλητές του πίνακα και τον τρόπο με τον οποίο τις μεταχειρίζονται. Συγκεκριμένα, θεωρούν όλες τις μεταβλητές ονομαστικές στην κλίμακα και δεν λαμβάνουν υπόψη τη διαταξιμότητα που μπορεί να έχουν κάποιες ή όλες από αυτές. Με τον τρόπο αυτό περιορίζουν αρκετά την ανάλυση αλλά και τα συμπεράσματα που εξάγουμε. Την πληροφορία που παίρνουμε από τη φυσική διάταξη των κατηγοριών των μεταβλητών μπορούμε να την χρησιμοποιήσουμε και να κατασκευάσουμε μοντέλα πιο οικονομικά, που να μην είναι κορεσμένα και δίνουν δομή στις αλληλεπιδράσεις ενώ περιέχουν παραμέτρους συνάφειας για να την αναλύσουμε. Τα μοντέλα συνάφειας εμφανίζονται ιδιαίτερα στις περιπτώσεις που τα λογαριθμογραμμικά μοντέλα είναι κορεσμένα (Agresti, 2002).

4.1 Μοντέλα Συνάφειας για Διδιάστατους Πίνακες

4.1.1 Εισαγωγή

Το λογαριθμογραμμικό μοντέλο για διδιάστατους πίνακες όπως γνωρίζουμε μπορεί να πάρει μονάχα δύο μορφές: της ανεξαρτησίας και του κορεσμένου μοντέλου. Εάν το μοντέλο της ανεξαρτησίας απορριφθεί, τότε μένουμε με το κορεσμένο μοντέλο για να αναλύσουμε τον πίνακα το οποίο έχει τόσες παραμέτρους όσα και τα κελιά του πίνακα και δεν παρέχει επιπλέον πληροφορίες. Όμως, υπάρχουν οικονομικότερα (parsimonious) μοντέλα τα οποία είναι «ενδιάμεσα» του κορεσμένου μοντέλου και εκείνου της ανεξαρτησίας και τα οποία

μπορούν να μελετηθούν καθώς περιορίζουν τις παραμέτρους της αλληλεπίδρασης κατά ένα τρόπο. Τα μοντέλα αυτά περιγράφουν τη συνάφεια του πίνακα χρησιμοποιώντας παραμέτρους για τις γραμμές και τις στήλες αντί παραμέτρους για τα κελιά. Οι παράμετροι αυτοί μπορούν να ιδωθούν σαν σκορ που εκτιμάμε για τις κατηγορίες των μεταβλητών. Στην περίπτωση αυτή η δομή της αλληλεπίδρασης τίθεται υπό μελέτη, οπότε εάν η υπόθεση της ανεξαρτησίας απορριφθεί η απόκλιση από την ανεξαρτησία μπορεί να μελετηθεί χάρη στη δομή που δίνει το μοντέλο στον όρο της αλληλεπίδρασης.

4.1.2 Μοντέλο Συνάφειας Τύπου RC(M)

Η συγκεκριμένη οικογένεια μοντέλων έχει μελετηθεί εκτενώς από τον Goodman σε διάφορες εργασίες του καθώς ήταν και εκείνος που τα πρωτομελέτησε στις αρχές της δεκαετίας του ’80. Έκτοτε διάφοροι ερευνητές ασχολήθηκαν με την περαιτέρω θεμελίωση και ανάπτυξη τους όπως οι Becker, Glogg, Becker και Glogg, Haberman, Agresti, να είναι ορισμένα από τα σημαντικότερα ονόματα ερευνητών. Στη βιβλιογραφία που υπάρχει στο τέλος της εργασίας μπορεί ο ενδιαφερόμενος αναγνώστης να βρει ορισμένα από τα άρθρα τους. Η διατύπωση του μοντέλου μπορεί να γίνει ως εξής:

Σε έναν $I \times J$ πίνακα συνάφειας δύο μεταβλητών Α και Β, n_{ij} είναι η παρατηρούμενη συχνότητα για το (i, j) κελί και $\mathrm{E}(n_{ij}) = m_{ij}$ είναι η αναμενόμενη συχνότητα ενός μοντέλου για δειγματοληψία Poisson (**Σημ**. Ίδια αποτελέσματα ισχύουν και για άλλα δειγματοληπτικά σχέδια όπως όταν έχουμε πολυωνυμική κατανομή). Έστω η ποσότητα $P_{ij} = m_{ij} / N$ δηλώνει την αναμενόμενη πιθανότητα για το (i, j) κελί. Τότε η ανεξαρτησία μεταξύ των γραμμών και των στηλών του πίνακα μπορεί να δειχθεί ως

$$P_{ij} = P_{i\cdot} P_{\cdot j} \tag{1}$$

όπου $P_{i\cdot} = \sum_j P_{ij}$ και $P_{\cdot j} = \sum_i P_{ij}$ είναι οι κατανομές περιθωρίου. Το μοντέλο (1) μπορεί επίσης να γραφτεί με την παρακάτω πολλαπλασιαστική μορφή

$$P_{ij} = \alpha_i \beta_j \tag{2}$$

όπου τα a_i και β_j είναι θετικές παράμετροι και αντιστοιχούν στις κύριες επιδράσεις των γραμμών και των στηλών.

Η γενικότερη μορφή του μοντέλου (2) που προτάθηκε από τον Goodman είναι η εξής:

$$P_{ij} = \alpha_i \beta_j \exp\left(\sum_{m=1}^{M} \phi_m \mu_{im} \nu_{jm} \right) \tag{3}$$

Στο μοντέλο (3) τα $\alpha_i, \beta_j, \mu_{im}, \nu_{jm}$ και ϕ_m είναι παράμετροι για, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $m = 1, 2, \ldots, M$, με $1 \le M^* \le M$, ενώ $M = \min\left(I - 1, J - 1\right)$ είναι ο αριθμός των συνιστωσών ή η διάσταση της συνάφειας γραμμών-στηλών. Επίσης η ισοδύναμη λογαριθμογραμμική διατύπωση του είναι η

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \sum_{m=1}^{M} \phi_m^{AB} \mu_{im} \nu_{jm} \tag{4}$$

Το μοντέλο (4) ή αντίστοιχα το (3) δεν είναι λογαριθμογραμμικό μοντέλο αλλά λογαριθμο-πολλαπλασιαστικό (log-multiplicative) και για την προσδιορισιμότητα του πρέπει να θέσουμε περιορισμούς θέσης (location) και κλίμακας (scale) στις παραμέτρους του. Ο λόγος είναι ότι ο όρος διπλής αλληλεπίδρασης λ_{ij}^{AB} του κορεσμένου λογαριθμογραμμικού μοντέλου αναλύεται στο μοντέλο (4) από έναν πολλαπλασιαστικό όρο που περιλαμβάνει τρεις άγνωστες παραμέτρους που θα εκτιμηθούν ταυτόχρονα. Επιπλέον, το μοντέλο αυτό είναι πιο οικονομικό από άποψης παραμέτρων από το κορεσμένο μοντέλο για διδιάστατους πίνακες και μπορεί κάλλιστα να μελετήσει την συνάφεια μεταξύ των μεταβλητών Α και Β.

Στο μοντέλο (3) οι παράμετροι μ_{im} και ν_{jm} είναι τα τυποποιημένα σκορ των κατηγοριών των γραμμών και των στηλών που θα εκτιμηθούν από τα δεδομένα και η παράμετρος ϕ_m είναι ένα μέτρο της συνάφειας για τον $I \times J$ πίνακα που ονομάζεται «παράμετρος εσωτερικής συνάφειας» και δείχνει τον βαθμό και την κατεύθυνση της συνάφειας του πίνακα. Σύμφωνα με τον Goodman, το μοντέλο αυτό είναι το κορεσμένο RC μοντέλο συνάφειας γιατί η συνάφεια ή αλληλεπίδραση λ_{ij}^{AB} αναλύεται πλήρως σε M συνιστώσες ή όρους και έχει 0 βαθμούς ελευθερίας. Εάν αντικαταστήσουμε το M με το $M^* < \min\left(I - 1, J - 1\right)$, τότε παίρνουμε το μη-κορεσμένο RC μοντέλο αφού η συνάφεια αναλύεται με λιγότερες διαστάσεις/συνιστώσες (components) από τον μέγιστο της αριθμό. Για τον προσδιορισμό του μοντέλου τα σκορ θα πρέπει να ικανοποιούν ορισμένες συνθήκες. Οι Becker & Glogg (1989) πρότειναν διάφορα βάρη για τη στάθμιση τους ώστε να πάρουμε τυποποιημένα σκορ και στη περίπτωση που $M > 1$ τα σκορ αυτά να είναι και συσχετισμένα μεταξύ τους. Μπορούμε επίσης να χρησιμοποιήσουμε σκορ χωρίς στάθμιση ώστε αυτά να είναι κανονικοποιημένα και ορθογώνια. Για λόγους όμως σύγκρισης με τα μοντέλα συσχέτισης που θα δούμε στο επόμενο

κεφάλαιο θα δώσουμε τους περιορισμούς για τα σκορ όταν χρησιμοποιούνται ως σταθμά οι κατανομές περιθωρίου (**Σημ**. Ένας λόγος είναι ότι ο συντελεστής συσχέτισης ρ_{AB} εξαρτάται από τις περιθώριες κατανομές):

$$\sum_{i=1}^{I}\mu_{im}P_{i\bullet}=0,\qquad \sum_{j=1}^{J}\nu_{jm}P_{\bullet j}=0$$

$$\sum_{i=1}^{I}\mu_{im}^{2}P_{i\bullet}=1,\qquad \sum_{j=1}^{J}\nu_{jm}^{2}P_{\bullet j}=1$$

$$\sum_{i=1}^{I}\mu_{im}\mu_{im'}P_{i\bullet}=0,\ \sum_{j=1}^{J}\nu_{jm}\nu_{jm'}P_{\bullet j}=0 \qquad (5)$$

για $m\neq m'$. Όταν $M^{*}=1$ η τελευταία συνθήκη στην (5) είναι άσχετη.

Επίσης από το μοντέλο (3) μπορούμε να δούμε ότι η αλληλεπίδραση λ_{ij}^{AB} στο κορεσμένο λογαριθμογραμμικό μοντέλο αναλύεται σε M συνιστώσες που περιέχουν γραμμικούς όρους:

$$\lambda_{ij}=\log\left(P_{ij}/\alpha_i\beta_j\right)=\sum_{m=1}^{M}\phi_m\mu_{im}\nu_{jm} \qquad (6)$$

Για την ανάλυση στην (6) χρησιμοποιούμε την διάσπαση ιδιόμορφων τιμών (SVD) του πίνακα των αλληλεπιδράσεων λ_{ij}^{AB} του λογαριθμογραμμικού μοντέλου. Επίσης, τα παραπάνω σκορ μπορούμε να πούμε ότι μεγιστοποιούν το ϕ_m, $m=1,2,\ldots,M$, υπό τον περιορισμό ότι είναι ασυσχέτιστα όταν $m\neq m'$ (τελευταία συνθήκη στη 5).

Εάν $\theta_{ij.i'j'}$ είναι το τοπικό *odds ratio* για τον 2×2 υποπίνακα τότε από την (3) ή (4) παίρνουμε την ποσότητα

$$\log\left(\theta_{ij.i'j'}\right)=\sum_{m=1}^{M}\phi_m(\mu_{im}-\mu_{i'm})(\nu_{jm}-\nu_{j'm}) \qquad (7)$$

που είναι το log-*odds ratio* μεταξύ δύο διαδοχικών γραμμών και στηλών. Έτσι βλέπουμε ότι για την ερμηνεία του μοντέλου το log-*odds ratio* εκφράζεται σαν συνάρτηση του ϕ_m και των διαφορών των σκορ ανάμεσα στις διαδοχικές κατηγορίες για τις γραμμές και των γειτονικών ή διαδοχικών κατηγοριών των σκορ για τις στήλες. Άρα η παράμετρος ϕ_m αναφέρεται στη γενική συνάφεια όπως αυτή μετριέται από τα *odds ratios* και οι διαφορές ανάμεσα στα σκορ αντανακλούν τις διαφορές στα log-*odds ratios* στους διάφορους υποπίνακες. Επιπλέον βλέπουμε ότι το log-*odds ratio* αναλύεται σε M συνιστώσες, μία συνιστώσα για κάθε διάσταση.

Τέλος μπορούμε να πάρουμε μοντέλα που θέτουν περιορισμούς στα σκορ των γραμμών ή/και των στηλών. Έτσι όταν $\mu_{im} = \mu_{i'm}$ και $\nu_{jm} = \nu_{j'm}$, έχουμε ομοιογένεια δύο διαδοχικών γραμμών και δύο στηλών και αν το μοντέλο που θα προσαρμόσουμε είναι αποδεκτό μπορούμε να απλοποιήσουμε τον πίνακα συγχωνεύοντας τις κατηγορίες. Επίσης αν θεωρήσουμε ότι η απόσταση μεταξύ διαδοχικών κατηγοριών των μεταβλητών είναι γνωστή και σταθερή

$$\mu_{i+1} - \mu_i = \Delta, \text{ για } i = 1, 2, \ldots, I-1$$

$$\nu_{j+1} - \nu_j = \Delta^*, \text{ για } j = 1, 2, \ldots, J-1,$$

τότε παίρνουμε το μοντέλο ομοιόμορφης συνάφειας (*U*). Επίσης όταν ισχύει μόνο η $\mu_{i+1} - \mu_i = \Delta$ τότε παίρνουμε το μοντέλο επίδρασης στηλών (*C*), ενώ αντίστοιχα όταν ισχύει η $\nu_{j+1} - \nu_j = \Delta^*$ μόνο παίρνουμε το μοντέλο επίδρασης γραμμών (*R*).

Το RC(M) μοντέλο δεν απαιτεί τη γνώση της διάταξης των κατηγοριών των μεταβλητών και η ιδιότητα που έχει στο να παραμένει αναλλοίωτο στις εναλλαγές των κατηγοριών των μεταβλητών σημαίνει ότι μεταχειρίζεται κατά κάποιο τρόπο τις μεταβλητές ταξινόμησης σαν ονομαστικές (Agresti, 2002). Ωστόσο μπορεί να χρησιμοποιηθεί και για την εύρεση της διάταξης των κατηγοριών.

4.2 Μοντέλα Συνάφειας για Τρισδιάστατους Πίνακες

4.2.1 Εισαγωγή

Όπως αναφέρθηκε νωρίτερα, όταν οι μεταβλητές είναι μόνο δύο τα λογαριθμογραμμικά μοντέλα που μπορούν να κατασκευαστούν είναι επίσης δύο και τα μοντέλα συνάφειας καθώς και οι ειδικές περιπτώσεις αυτών, δίνουν τη δυνατότητα μελέτης του όρου αλληλεπίδρασης των μεταβλητών αφού δεν είναι ποτέ κορεσμένα. Άμεση επέκταση των μοντέλων συνάφειας μπορεί να γίνει στην περίπτωση που έχουμε για μελέτη τρεις κατηγορικές μεταβλητές $\{A, B, C\}$ που ταξινομούνται σε έναν πίνακα συνάφειας $I \times J \times K$. Τώρα όμως μπορούμε να κατασκευάσουμε περισσότερα λογαριθμογραμμικά μοντέλα και να μελετήσουμε τις αλληλεπιδράσεις. Παρόλα αυτά, υπάρχουν ορισμένες περιπτώσεις που αυτά τα μοντέλα εμφανίζουν κάποιους περιορισμούς:

- Όταν ο αριθμός των κατηγοριών των μεταβλητών είναι μεγάλος. Τότε ο αριθμός των αλληλεπιδράσεων των παραγόντων που θα προκύψει θα είναι εκτενής και πολλές φορές η ερμηνεία τους είναι ιδιαίτερα δύσκολη ιδιαίτερα στις αλληλεπιδράσεις μεγαλύτερης τάξης αν συμβαίνει να είναι στατιστικά σημαντικές,
- Δε λαμβάνουν υπόψη τη διαταξιμότητα που μπορεί να έχουν κάποιες ή όλες οι μεταβλητές,
- Θεωρούν τις μεταβλητές ονομαστικές στην κλίμακα με αποτέλεσμα όταν η διάταξη των κατηγοριών τους αλλάξει, η προσαρμογή του μοντέλου να παραμένει ίδια,
- Δεν παρέχουν καμία πληροφορία για την απόσταση ανάμεσα στις κατηγορίες των κατηγορικών μεταβλητών,
- Δεν επιτρέπουν τη γραφική απεικόνιση των κατηγοριών των μεταβλητών για τον εντοπισμό συσχετισμών των κατηγοριών αλλά και πιθανών αλληλεπιδράσεων των μεταβλητών ούτως ώστε η ερμηνεία να απλοποιηθεί,
- Ειδικά εάν η αλληλεπίδραση των τριών παραγόντων πρέπει να μελετηθεί, τα μοντέλα αυτά γίνονται κορεσμένα και δεν μπορούν να χρησιμοποιηθούν.

4.2.2 Λογαριθμογραμμικά Μοντέλα

Για έναν $I \times J \times K$ πίνακα συνάφειας τριών μεταβλητών $\{A, B, C\}$ έστω n_{ijk} είναι η παρατηρούμενη συχνότητα του (i, j, k) κελιού, με $i = 1, \ldots, I$, $j = 1, \ldots J$, $k = 1, \ldots, K$ και με $\mathrm{E}(n_{ijk}) = m_{ijk}$ παριστάνουμε τις αναμενόμενες συχνότητες για δειγματοληψία *Poisson*. Συνήθως ο ερευνητής κάνει διάφορες υποθέσεις για τις σχέσεις μεταξύ των μεταβλητών και ενδιαφέρεται να μειώσει τις διαστάσεις του πίνακα για να μελετήσει σαφώς λιγότερο πολύπλοκα μοντέλα. Αυτό όμως δεν είναι πάντοτε εφικτό και αρκετές φορές οδηγείται σε λανθασμένα συμπεράσματα. Για να μπορεί ένας $I \times J \times K$ πίνακας να θεωρηθεί συρρικνωμένος (collapsible) θα πρέπει η συνάφεια στους μερικούς πίνακες να είναι ίση με τη συνάφεια του πίνακα περιθωρίου. Γενικά θα πρέπει να ισχύουν τα εξής: Α) Να μην υφίσταται η τριπλή αλληλεπίδραση των παραγόντων, Β) Ένας τουλάχιστον από τους όρους διπλής αλληλεπίδρασης είναι μηδέν. Οι συνθήκες αυτές είναι πολύ σημαντικές για τις περιπτώσεις εκείνες που θέλουμε να μελετήσουμε την τριπλή αλληλεπίδραση χρησιμοποιώντας διδιάστατους πίνακες. Όπως θα αναφέρουμε παρακάτω, ένας τρόπος ανάλυσης του $I \times J \times K$ πίνακα με τα μοντέλα συνάφειας είναι να τον τροποποιήσουμε κατάλληλα σε διδιάστατο.

Για να μπορέσουμε να το κάνουμε αυτό, ταξινομούμε τα λογαριθμογραμμικά μοντέλα σε 4 κλάσεις όπου ζητάμε να μελετήσουμε ορισμένες ή όλες τις αλληλεπιδράσεις τους. Εκτός από το μοντέλο της ανεξαρτησίας των τριών μεταβλητών (θα αναφερθούμε μόνο σε ιεραρχικά μοντέλα, δηλαδή όταν υπάρχει όρος αλληλεπίδρασης ανώτερης τάξης τότε θα υπάρχουν οι όροι αλληλεπίδρασης μικρότερης τάξης καθώς και οι κύριες επιδράσεις τους) κατασκευάζουμε τα εξής μοντέλα: Στην $1^{η}$ κλάση περιλαμβάνεται το «joint independence» μοντέλο που περιέχει ένα μόνο όρο διπλής αλληλεπίδρασης των παραγόντων που μπορούμε να μελετήσουμε και συμβολίζεται με (A, BC), (AB, C) ή (AC, B). Τότε ο τρισδιάστατος πίνακας μπορεί να συρρικνωθεί και στις 3 διαστάσεις αφού: *μερική συνάφεια* = *περιθώρια συνάφεια* = 0, για κάθε ζεύγος μεταβλητών ενώ δεν υπάρχει και η τριπλή αλληλεπίδραση τους. Η $2^{η}$ κλάση περιλαμβάνει το μοντέλο της «δεσμευμένης (conditional) ανεξαρτησίας» με δύο όρους διπλής αλληλεπίδρασης $(AB, BC), (AC, BC)$ ή (AB, AC). Τότε ο $I \times J \times K$ πίνακας μπορεί να συρρικνωθεί σε οποιεσδήποτε 2 από τις 3 διαστάσεις του. Για παράδειγμα για το (AB, AC) μοντέλο, οι B και C είναι ανεξάρτητες δοθείσης της A. Αυτό σημαίνει ότι ο πίνακας μπορεί να συρρικνωθεί είτε προς την B είτε προς την C, όχι όμως προς την A. Στην $3^{η}$ κλάση ανήκει το μοντέλο «ομοιογενούς συνάφειας ή χωρίς την αλληλεπίδραση τριών παραγόντων», (AB, BC, AC) συμβολικά. Το μοντέλο αυτό περιέχει όλες τις αλληλεπιδράσεις $2^{ης}$ τάξης ενώ κανένα ζεύγος μεταβλητών δεν είναι υπό συνθήκη ανεξάρτητο. Αποτέλεσμα είναι ο πίνακας να μην είναι δυνατόν να συρρικνωθεί προς καμία κατεύθυνση (διάσταση) αφού ισχύει ότι: *μερική συνάφεια* ≠ *περιθώρια συνάφεια*, για κάθε ζεύγος μεταβλητών. Η απουσία του όρου της τριπλής αλληλεπίδρασης προϋποθέτει την «ομοιογένεια στην συνάφεια». Αυτό σημαίνει ότι τα δεσμευμένα *odds ratios* για κάθε ζεύγος μεταβλητών είναι ίδια και δεν μεταβάλλονται στα επίπεδα της τρίτης μεταβλητής. Τέλος, στη $4^{η}$ κλάση βρίσκουμε το πλέον γενικό μοντέλο για τρισδιάστατους πίνακες που περιέχει όλες τις διπλές αλληλεπιδράσεις των μεταβλητών και την τριπλή το οποίο είναι το κορεσμένο μοντέλο και το συμβολίζουμε με (ABC).

Συνηθίζεται η φυσική ερμηνεία των παραμέτρων των λογαριθμογραμμικών μοντέλων να αποδίδεται μέσω των περιορισμών που έχουμε θέσει στα δεσμευμένα ή log-*odds ratios*. Στην περίπτωση του μοντέλου της $3^{η}$ κλάσης η ερμηνεία αυτή αφορά τις μερικές αλληλεπιδράσεις. Έτσι, για σταθερό επίπεδο k της μεταβλητής ταξινόμησης C, η υπό συνθήκη συνάφεια

ανάμεσα στην A και στην B ορίζεται μέσω των $(I-1)(J-1)$ *odds ratios* όπως τα τοπικά *odds ratios* (Agresti, 2002):

$$\theta_{ij(k)} = \frac{m_{ijk} m_{i+1,j+1,k}}{m_{i,j+1,k} m_{i+1,j,k}}, \quad 1 \le i \le I-1, \;\; 1 \le j \le J-1$$

Κατά ανάλογο τρόπο ορίζονται τα $(I-1)(K-1)$ *odds ratios* $\{\theta_{i(j)k}\}$ για την AC υπό συνθήκη ή μερική συνάφεια και τα $(J-1)(K-1)$ *odds ratios* $\{\theta_{(i)jk}\}$ για την BC υπό συνθήκη συνάφεια. Για το μοντέλο της κλάσης αυτής, το log-*odds ratio* για την AB υπό συνθήκη συνάφεια ισοδυναμεί με

$$\log \theta_{ij(k)} = \lambda_{ij}^{AB} + \lambda_{i+1,j+1}^{AB} - \lambda_{i,j+1}^{AB} - \lambda_{i+1,j}^{AB}.$$

Αντιστοίχως τα άλλα υπό συνθήκη log-*odds ratios* του μοντέλου είναι

$$\log \theta_{i(j)k} = \lambda_{ik}^{AC} + \lambda_{i+1,k+1}^{AC} - \lambda_{i,k+1}^{AC} - \lambda_{i+1,k}^{AC},$$

και

$$\log \theta_{(i)jk} = \lambda_{jk}^{BC} + \lambda_{j+1,k+1}^{BC} - \lambda_{j,k+1}^{BC} - \lambda_{j+1,k}^{BC}.$$

Τα δεσμευμένα log-*odds ratios* είναι ομοιόμορφα (uniform) σε όλα τα επίπεδα της τρίτης μεταβλητής πράγμα που αντανακλά την απουσία του όρου τριπλής αλληλεπίδρασης. Δηλαδή ισχύει η παρακάτω ισότητα για τα *odds ratios*

$$\theta_{ij(1)} = \theta_{ij(2)} = \ldots = \theta_{ij(K)}$$

για όλα τα i και j, ενώ αντίστοιχες ισότητες ισχύουν και για τα υπόλοιπα. Παρόμοια ερμηνεία δίνεται και για τον όρο τριπλής αλληλεπίδρασης στη 4^{η} κλάση όπου το log-*odds ratio* ορίζεται ως

$$\log \theta_{ijk} = \log \theta_{ij(k+1)} - \log \theta_{ij(k)} \tag{8}$$

δηλαδή ως η διαφορά ανάμεσα σε δύο διαδοχικά (τοπικά) log-*odds ratios*.

Με τον τρόπο αυτό αναλύονται τα λογαριθμογραμμικά μοντέλα και γενικά τα μοντέλα συνάφειας. Συνήθως η μεταβλητή απόκρισης είναι ο λογάριθμος των αναμενόμενων συχνοτήτων και όχι κάποια μεταβλητή του πίνακα. Έτσι οι εξαρτημένες και οι ανεξάρτητες μεταβλητές εμφανίζονται συμμετρικά. Το μοντέλο της $4^{\eta\varsigma}$ κλάσης έχει την εξής μορφή

$$\log m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC} + \lambda_{ijk}^{ABC} \tag{9}$$

όπου για την προσδιορισιμότητα του μπορούμε να χρησιμοποιήσουμε περιορισμούς τύπου ANOVA στις παραμέτρους (ή zero-sum constraints):

$$\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_k \lambda_k^C = 0$$
$$\sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = \sum_i \lambda_{ik}^{AC} = \sum_k \lambda_{ik}^{AC} = \sum_j \lambda_{jk}^{BC} = \sum_k \lambda_{jk}^{BC} = 0$$
$$\sum_i \lambda_{ijk}^{ABC} = \sum_j \lambda_{ijk}^{ABC} = \sum_k \lambda_{ijk}^{ABC} = 0$$

Εναλλακτικά μπορούμε να χρησιμοποιήσουμε ψευδομεταβλητές, με την πρώτη ή τελευταία κατηγορία να χρησιμοποιείται ως κελί αναφοράς.

Το μοντέλο (9) έχει 0 βαθμούς ελευθερίας και πολλές φορές θέλουμε να το μελετήσουμε ειδικά αν τα μοντέλα των προηγούμενων κλάσεων αποδειχτούν μη στατιστικά σημαντικά. Ακόμα και αυτό να μην συμβεί, θα υπάρχουν περιπτώσεις όπου τα λογαριθμογραμμικά μοντέλα θα χάνουν ιδιοτήτων όπως για παράδειγμα όταν ορισμένες ή όλες οι μεταβλητές έχουν μια φυσική διάταξη. Στις επόμενες παραγράφους θα αναλύσουμε το μοντέλο της $3^{ης}$ κλάσης και το μοντέλο (9) χρησιμοποιώντας μοντέλα συνάφειας για να διασπάσουμε τις αλληλεπιδράσεις σε πίνακες μικρότερης τάξης και να κερδίσουμε έτσι σε ερμηνεία όταν αυτό απαιτείται αλλά και σε οικονομία από άποψης παραμέτρων. Για τα μοντέλα αυτά χρειάζεται ο αριθμός των κατηγοριών των μεταβλητών να είναι τουλάχιστον τρία για να διασπαστούν οι αλληλεπιδράσεις σε συνιστώσες, αλλιώς δεν θα έχει ιδιαίτερη σημασία να προχωρήσουμε.

4.2.3 Αναλύοντας Μόνο τις Αλληλεπιδράσεις $2^{ης}$ Τάξης

Χρησιμοποιώντας τον συμβολισμό $\mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C$, γράφουμε το μοντέλο της $3^{η}$ κλάσης στη μορφή

$$\log m_{ijk} = \mu_{ijk} + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} \tag{10}$$

Τότε σύμφωνα με τον Becker (1989) η γενική λογαριθμογραμμική έκφραση για το αντίστοιχο μοντέλο συνάφειας τριών μεταβλητών χωρίς την αλληλεπίδραση $3^{ης}$ τάξης, του μοντέλου (10) είναι η εξής

$$\log m_{ijk} = \mu_{ijk} + \sum_{m=1}^{M_1} \phi_m^{AB} \nu_{1im}^A \nu_{1jm}^B + \sum_{m=1}^{M_2} \phi_m^{AC} \nu_{2im}^A \nu_{1km}^C + \sum_{m=1}^{M_3} \phi_m^{BC} \nu_{2jm}^B \nu_{2km}^C \tag{11}$$

όπου $0 \le M_1 \le \min(I-1, J-1), 0 \le M_2 \le \min(I-1, K-1)$ και $0 \le M_3 \le \min(J-1, K-1)$. Οι περιορισμοί που χρησιμοποιούνται για τον προσδιορισμό των παραμέτρων του μοντέλου είναι:

$$\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_k \lambda_k^C = 0$$
$$\sum_i \nu_{eim}^A = \sum_j \nu_{ejm}^B = \sum_k \nu_{ekm}^C = 0 \qquad e = 1, 2.$$
$$\sum_i \nu_{eim}^A \nu_{eim'}^A = \sum_j \nu_{ejm}^B \nu_{ejm'}^B = \sum_k \nu_{ekm}^C \nu_{ekm'}^C = \delta_{mm'}$$

όπου $\delta_{mm'}$ συμβολίζει το δέλτα του *Kronecker*, με $\delta_{mm'} = 1$ αν $m' = m$, αλλιώς $\delta_{mm'} = 0$ για $m' \neq m$. Οι βαθμοί ελευθερίας για τον έλεγχο καλής προσαρμογής του μοντέλου είναι:
$\mathrm{df} = IJK - (I + J + K) - M_1(I + J - M_1 - 2) - M_2(I + K - M_2 - 2) - M_3(J + K - M_3 - 2) + 2$.
Σημειώνουμε πως όταν τα M_1, M_2, M_3 είναι ίσα με το πάνω φράγμα τους τότε το μοντέλο (11) είναι ισοδύναμο (ισότιμο) με το αντίστοιχο λογαριθμογραμμικό μοντέλο (10) και προφανώς τότε θα έχει: $\mathrm{df} = (I-1)(J-1)(K-1)$. Επίσης όταν $M_1 = M_2 = M_3 = 0$ τότε το μοντέλο είναι ισότιμο με το μοντέλο ανεξαρτησίας με: $\mathrm{df} = IJK - (I + J + K) + 2$.

Επίσης μπορούμε να θέσουμε περιορισμούς στο μοντέλο (11) και να πάρουμε περιορισμένα (restricted) μοντέλα, όπως για παράδειγμα θέτοντας: (α) γνωστά σκορ, (β) ομοιογένεια στα σκορ (π.χ. $\nu_{1im}^A = \nu_{2im}^A = \nu_{im}^A$), (γ) συμμετρικά σκορ (π.χ. $\nu_{1im}^A = \nu_{1jm}^B$, όταν $i = j$, $\forall m \in M$). Για την μελέτη του μοντέλου μπορούμε να χρησιμοποιήσουμε τα υπό συνθήκη *odds ratios* $\{\theta_{i(j)k}\}, \{\theta_{(i)jk}\}$ και $\{\theta_{ij(k)}\}$. Επίσης οι Haberman (1981) και Becker (1989), αναφέρουν ότι η μορφή του μοντέλου δυσκολεύει και την συμπερασματολογία διότι δεν αιτιολογείται πλήρως η χρήση της X^2 κατανομής για τον έλεγχο καλής προσαρμογής. Για $M_1^* = M_2^* = M_3^* = 1$ συνιστώσες, το (11) συμβολίζεται με $RC(1) + RL(1) + CL(1)$ και γράφεται ως

$$\log m_{ijk} = \mu_{ijk} + \phi_1^{AB} \nu_{1i1}^A \nu_{1j1}^B + \phi_1^{AC} \nu_{2i1}^A \nu_{1k1}^C + \phi_1^{BC} \nu_{2j1}^B \nu_{2k1}^C \qquad (12)$$

το οποίο έχει: $\mathrm{df} = IJK - 3I - 3J - 3K + 11$ και είναι επίσης μη κορεσμένο.

Θέτοντας (α) γνωστά σκορ $\{\upsilon_i = i\}$ για τις γραμμές, $\{v_j = j\}$ για τις στήλες και $\{w_k = k\}$ για τα στρώματα, κατά αυτόν τον τρόπο γενικεύουμε το μοντέλο ομοιόμορφης συνάφειας (U) σε ένα μοντέλο μερικής συνάφειας για τρισδιάστατους πίνακες. Η μορφή του είναι η εξής

$$\log m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \beta^{AB} \upsilon_i v_j + \beta^{AC} \upsilon_i w_k + \beta^{BC} v_j w_k \qquad (13)$$

και ονομάζεται ομοιογενές ομοιόμορφο (homogenous uniform association) μοντέλο σύμφωνα με τον Agresti (1983). Τα $\beta^{AB}, \beta^{AC}, \beta^{BC}$ είναι παράμετροι που περιγράφουν τις μερικές

συνάφειες στον πίνακα $I \times J \times K$. Το μοντέλο αυτό δηλώνει ότι η συνάφεια σε κάθε πίνακα είναι ομοιόμορφη (uniform) για κάθε ζεύγος μεταβλητών και παράλληλα ομοιογενής (homogenous) σε όλα τα επίπεδα της τρίτης μεταβλητής. Οι βαθμοί ελευθερίας του μοντέλου (13) είναι $IJK - I - J - K - 1$, καθώς έχει τρεις παραμέτρους περισσότερες από το μοντέλο ανεξαρτησίας.

Αν θέσουμε (β) ομοιογένεια στα σκορ, τότε παίρνουμε το μοντέλο

$$\log m_{ijk} = \mu_{ijk} + \phi_1^{AB} \nu_{i1}^{A} \nu_{j1}^{B} + \phi_1^{AC} \nu_{i1}^{A} \nu_{k1}^{C} + \phi_1^{BC} \nu_{j1}^{B} \nu_{k1}^{C} \qquad (14)$$

με $\mathrm{df} = IJK - 2I - 2J - 2K + 5$. Στην περίπτωση αυτή τα σκορ για τις γραμμές, τις στήλες και τα στρώματα δεν διαφοροποιούνται για κάθε μερική συνάφεια των μεταβλητών οπότε δεν χρειάζεται να υπολογιστούν ξανά σε αντίθεση με το μοντέλο (12). Τότε τα log-odds ratios για την ερμηνεία του δίνονται από τις σχέσεις

$$\begin{aligned} \log\theta_{\mathrm{ij(k)}} &= \phi^{AB}\left(\nu_{i+1} - \nu_i\right)\left(\nu_{j+1} - \nu_j\right) \\ \log\theta_{\mathrm{i(j)k}} &= \phi^{AC}\left(\nu_{i+1} - \nu_i\right)\left(\nu_{k+1} - \nu_k\right), \quad \text{ενώ } \log\theta_{\mathrm{ijk}} = 0. \\ \log\theta_{\mathrm{(i)jk}} &= \phi^{BC}\left(\nu_{j+1} - \nu_j\right)\left(\nu_{k+1} - \nu_k\right) \end{aligned}$$

4.2.4 Αλληλεπίδραση Τριών Παραγόντων

Όταν υπάρχει ο όρος της τριπλής αλληλεπίδρασης το ιεραρχικό λογαριθμογραμμικό μοντέλο είναι το (9). Τότε χρησιμοποιώντας μοντέλα συνάφειας μπορούμε να μελετήσουμε το μοντέλο αυτό αφού είναι δυνατόν να προκύψουν οικονομικότερα μοντέλα. Υπάρχουν διάφορες δυνατότητες στη περίπτωση αυτή καθώς και διάφοροι τρόποι διάσπασης των αλληλεπιδράσεων σε έναν $I \times J \times K$ πίνακα. Στην παράγραφο αυτή θα αναφερθούμε στη γενικότερη περίπτωση όπου στο μοντέλο (9) θα διασπάσουμε την αλληλεπίδραση $3^{\eta\varsigma}$ τάξης και κάποιες από τις αλληλεπιδράσεις $2^{\eta\varsigma}$ τάξης σε συνδυασμό με εκείνης της $3^{\eta\varsigma}$ τάξης βρίσκοντας την προσέγγιση τους με το μοντέλο του Tucker-3 (3-mode principal components analysis model). Η παρουσίαση των μοντέλων αυτών βασίζεται στην πολύ γόνιμη εργασία της C.J.Anderson (1996) και στα αποτελέσματα της μετά από την εφαρμογή των μοντέλων.

Ωστόσο θα πρέπει να αναφέρουμε ότι διάφορες στρατηγικές κατά καιρούς έχουν προταθεί από ερευνητές για την ανάλυση τρισδιάστατων πινάκων. Ως πρώτη προσέγγιση ήταν η χρησιμοποίηση του RC(M) μοντέλου σε πίνακες τριών μεταβλητών. Έτσι κατασκεύασαν πολλαπλούς πίνακες συνδυάζοντας τις κατηγορίες 2 μεταβλητών σε μία μεταβλητή με βάση την 1^{η} κλάση μοντέλων την οποία ονόμασαν «joint» approach (Gilula and Haberman 1988,

Goodman 1986) για την μελέτη της από κοινού δεσμευμένης συνάφειας. Επίσης, θεωρώντας την τρίτη μεταβλητή του πίνακα σαν «control» μελέτησαν την υπό συνθήκη συνάφεια δύο μεταβλητών όταν μεταβάλλεται στα επίπεδα της τρίτης μεταβλητής (Goodman 1986, Becker and Clogg 1989, Becker 1989, Xie et al. 2000). Στην 1^{η} περίπτωση το μοντέλο περιγράφει σε πολλαπλασιαστικούς όρους τη διπλή αλληλεπίδραση ανάμεσα στη μεταβλητή A και στην joint (BC) μεταβλητή σε διδιάστατο πίνακα και μπορεί να χρησιμοποιηθεί για την ανάλυση όταν οι μεταβλητές θεωρούνται σαν απόκρισης και επεξηγηματικές αντίστοιχα, ενώ στην 2^{η} περίπτωση χρησιμοποιείται η υπό συνθήκη συνάφεια για την ανάλυση της τριπλής αλληλεπίδρασης («conditional» approach ή analysis of variation in association). Όμως για την ανάλυση τρισδιάστατου πίνακα με τις παραπάνω προσεγγίσεις οι μεταβλητές δεν χρησιμοποιούνται συμμετρικά και τα μοντέλα αυτά δεν είναι ιεραρχικά.

Επίσης για την ανάλυση τρισδιάστατων πινάκων διάφοροι ερευνητές χρησιμοποίησαν συνδυασμούς τριγραμμικών όρων (trilinear terms), επέκταση της μέχρι τότε χρησιμοποίησης διγραμμικών όρων (bilinear terms-RC(M) μοντέλο) για τη διάσπαση της τριπλής αλληλεπίδρασης και διγραμμικούς όρους για τη διάσπαση των όρων διπλής αλληλεπίδρασης. Τα μοντέλα αυτά είναι οικονομικότερα των μοντέλων στις δύο προηγούμενες περιπτώσεις και για την ανάλυση τους χρησιμοποιήθηκαν τεχνικές που εφαρμόζονταν στις κοινωνικές επιστήμες, το marketing ή την ψυχολογία και ιδιαίτερα στα ψυχομετρικά τεστ (psychometrics). Οι Choulakian (1988), Kroonenberg (1989), Siciliano (1990), Mooijaart (1992), Siciliano and Mooijaart (1997), ανέλυσαν τον πίνακα που περιέχει την τριπλή αλληλεπίδραση λ_{ijk}^{ABC} χρησιμοποιώντας για την προσέγγιση του μοντέλα όπως το CANDECOMP (CANonical DECOMposition), που αποτελεί τη γενίκευση του INDSCAL (INidividual Differences SCALing) για 3-mode, 3-διαστάσεις πίνακα (Carroll and Chang, 1970) και το οποίο είναι ισοδύναμο με το PARAFAC-1 (PARAllel profiles FACtor analysis) μοντέλο.

4.2.4.1 Διάφορα Log-trilinear Μοντέλα

Όπως αναφέραμε νωρίτερα, για τη διάσπαση των αλληλεπιδράσεων του μοντέλου (9) θα χρησιμοποιήσουμε το μοντέλο του Tucker-3. Το μοντέλο γράφεται στη γενική του μορφή

$$\lambda_{ijk}^{(2,3)} = \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{t=1}^{T}\phi_{rst}\mu_{ir}\nu_{js}\xi_{kt} \qquad (15)$$

όπου ο συμβολισμός $\lambda_{ijk}^{(2,3)}$ είναι γενικός και αναφέρεται στο σύνολο των αλληλεπιδράσεων 2 παραγόντων και εκείνης των 3, που θα διασπαστούν από το μοντέλο. Τα μ_{ir}, ν_{js} και ξ_{kt} είναι οι παράμετροι σκορ των μεταβλητών $\{A, B, C\}$ για τις τρεις συνιστώσες r, s και t, ενώ ϕ_{rst} είναι η παράμετρος εσωτερικής συνάφειας. Για τον προσδιορισμό των παραμέτρων του (15) κατάλληλοι περιορισμοί θα πρέπει να τεθούν όπου μπορούμε να χρησιμοποιήσουμε τις προτάσεις των Becker & Glogg (1989) σε γενικές γραμμές.

Επίσης αν αντί του μοντέλο (15) γράψουμε $\lambda_{ijk}^{(2,3)} = \sum_{r=1}^{R} \phi_r \mu_{ir} \nu_{jr} \xi_{kr}$, τότε η διάσπαση αυτή είναι γνωστή σαν το CANDECOMP/PARAFAC-1 μοντέλο που αναφέραμε προηγουμένως. Σύμφωνα πάντα με τους Siciliano and Mooijaart (1997), για το μοντέλο αυτό ο μέγιστος αριθμός των συνιστωσών που απαιτείται για την πλήρη ανάλυση του πίνακα είναι άγνωστος με αποτέλεσμα το R να είναι μεγαλύτερο από τα I, J και K. Ωστόσο από πρακτικής άποψης μόνο τις πρώτες κύριες συνιστώσες χρειαζόμαστε και έτσι κατά ένα τρόπο λύνεται το πρόβλημα. Παρακάτω, ο Πίνακας 1 περιέχει όλα τα πιθανά μοντέλα και τις αλληλεπιδράσεις που θα διασπαστούν με το μοντέλο του Tucker-3

Πίνακας 1: *Βασικά μοντέλα για τρισδιάστατο πίνακα.*

Μοντέλο	$\lambda_{ijk}^{(2,3)}$
(ABC)	λ_{ijk}^{ABC}
(AB, ABC)	$\lambda_{ij}^{AB} + \lambda_{ijk}^{ABC}$
(AC, ABC)	$\lambda_{ik}^{AC} + \lambda_{ijk}^{ABC}$
(BC, ABC)	$\lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$
(AB, AC, ABC)	$\lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{ijk}^{ABC}$
(AB, BC, ABC)	$\lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$
(AC, BC, ABC)	$\lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$
(AB, AC, BC, ABC)	$\lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$

Για το μοντέλο (ABC) του Πίνακα 1, διασπάται μόνο η λ_{ijk}^{ABC} αλληλεπίδραση των τριών παραγόντων ενώ για τις αλληλεπιδράσεις κατώτερης τάξης δεν χρησιμοποιείται καμία δομή. Το μοντέλο αυτό μπορεί να γραφτεί στη μορφή

$$\log m_{ijk} = \mu_{ijk} + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{t=1}^{T}\phi_{rst}\mu_{ir}\nu_{js}\xi_{kt} \tag{16}$$

όπου $R \le (I-1)$, $S \le (J-1)$ και $T \le (K-1)$. Όταν $R = (I-1)$, $S = (J-1)$ και $T = (K-1)$, το (16) είναι ισοδύναμο με το μοντέλο (9). Σύμφωνα με την Anderson (1996), για το μοντέλο αυτό δεν χρειάζεται ο αριθμός των συνιστωσών να είναι ίσος και για τις τρεις μεταβλητές και για τον λόγο αυτό πολλές φορές, ανάλογα όμως και με τη φύση των δεδομένων, το μοντέλο (16) είναι πιο οικονομικό από άλλα μοντέλα συνιστωσών που έχουν προταθεί. Για τον έλεγχο του μοντέλου οι βαθμοί ελευθερίας είναι:

$\mathrm{df} = (I-1)(J-1)(K-1) - R(I-R-1) - S(J-S-1) - T(K-T-1) - RST$.

Για τα μοντέλα της 2ης κατηγορίας του Πίνακα 1, πχ. το (BC, ABC) γράφεται

$$\log m_{ijk} = \mu_{ijk} + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{t=1}^{T}\phi_{rst}\mu_{ir}\nu_{js}\xi_{kt} \tag{17}$$

όπου $R \le I$, $S \le (J-1)$ και $T \le (K-1)$. Στο (17) αναλύονται η αλληλεπίδραση 2ης τάξης (BC) καθώς και η της 3ης τάξης (ABC). Το μοντέλο αυτό ομοιάζει με την «conditional» προσέγγιση του RC(M) μοντέλου για την ανάλυση τρισδιάστατων πινάκων. Αυτό συμβαίνει διότι βάσει της προσέγγισης αυτής, οι ερευνητές ξεκινάνε με τη διατύπωση του μοντέλου της «δεσμευμένης ανεξαρτησίας» (AB, AC) και για τους όρους αλληλεπίδρασης λ_{jk}^{BC} και λ_{ijk}^{ABC} που ενδιαφέρονται να αναλύσουν προσπαθούν να τους προσδιορίσουν χρησιμοποιώντας πολλαπλασιαστικούς «layer-effect» όρους. Όπως και πριν όταν $R = I$, $S = (J-1)$ και $T = (K-1)$ το μοντέλο είναι κορεσμένο. Οι βαθμοί ελευθερίας του μοντέλου είναι: $\mathrm{df} = I(J-1)(K-1) - R(I-R) - S(J-S-1) - T(K-T-1) - RST$. Παρόμοια προκύπτουν και οι βαθμοί ελευθερίας για τα υπόλοιπα μοντέλα της κατηγορίας αυτής.

Για τα μοντέλα της 3ης κατηγορίας του Πίνακα 1, πχ. το (AC, BC, ABC) γράφεται στη μορφή

$$\log m_{ijk} = \mu_{ijk} + \lambda_{ij}^{AB} + \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{t=1}^{T}\phi_{rst}\mu_{ir}\nu_{js}\xi_{kt} \tag{18}$$

όπου $R \le I, S \le J$ και $T \le (K-1)$. Στο μοντέλο αυτό δεν αναλύεται μόνο η αλληλεπίδραση λ_{ij}^{AB} ενώ δεν μπορεί να είναι ισοδύναμο του (9) (βλ. Anderson, 1996 - 481:482). Το μοντέλο αυτό έχει την ίδια μορφή με το RC(M) μοντέλο της «joint» προσέγγισης αφού γράφεται και ως: $\sum_m \phi_m^{ABC} \nu_{im}^{A} \nu_{jkm}^{BC}$. Όμως με τον τρόπο αυτό δεν δύναται να υπολογιστούν τα σκορ για κάθε μία από τις μεταβλητές ξεχωριστά. Οι βαθμοί ελευθερίας για τον έλεγχο όπως και πριν είναι: $\mathrm{df} = (IJ-1)(K-1) - R(I-R) - S(J-S) - T(K-T-1) - RST$ και με αντίστοιχο τρόπο προκύπτουν για τα υπόλοιπα μοντέλα της κατηγορίας οι βαθμοί ελευθερίας.

Τέλος, το μοντέλο της 4ης κατηγορίας του Πίνακα 1, (AB, AC, BC, ABC), γράφεται

$$\log m_{ijk} = \mu_{ijk} + \sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{t=1}^{T} \phi_{rst} \mu_{ir} \nu_{js} \xi_{kt} \qquad (19)$$

Στο μοντέλο (19) όλες οι αλληλεπιδράσεις αναλύονται ταυτόχρονα μέσω της Tucker-3 διάσπασης. Θεωρείται ότι είναι αντίστοιχο του CANDECOMP/PARAFAC-$RCL(R)$ όπου χρησιμοποιούνται 3 διγραμμικοί όροι και 1 τριγραμμικός όρος για τη διάσπαση των αλληλεπιδράσεων. Όμως, υπάρχει πάλι το πρόβλημα του μέγιστου αριθμού των συνιστωσών.

Κλείνοντας μπορούμε να πούμε ότι είναι δυνατόν για τα παραπάνω μοντέλα να θέσουμε επιπλέον περιορισμούς ομοιογένειας στα σκορ έτσι ώστε σε κάθε μεταβλητή να μην υπολογίζουμε διαφορετικά σκορ για κάθε αλληλεπίδραση. Επίσης για όλα τα μοντέλα του Πίνακα 1 το μέγεθος της τριπλής αλληλεπίδρασης δίνεται από τα τοπικά *odds ratios* και είναι ίσο με το γινόμενο των αποστάσεων μεταξύ διαδοχικών κατηγοριών επί την παράμετρο εσωτερικής συνάφειας.

4.3 Εφαρμογή

Ο πίνακας που ακολουθεί είναι ένα μέρος από τα δεδομένα που συλλέχθηκαν το 1978 από την ACS (American Couples Survey). Η πλήρης περιγραφή των ερωτήσεων και η μέθοδος της δειγματοληψίας που χρησιμοποιήθηκε μπορεί να βρεθεί στους Blumstein and Schwartz (1983). Διάφοροι ειδικοί επιστήμονες έχουν ασχοληθεί κατά καιρούς με θέματα που αφορούν τις οικογένειες μέσα από τις έρευνες που έχουν διεξαχθεί. Κάποιες συγκεκριμένες μελέτες αφορούσαν την κατανομή των εργασιών στο σπίτι με βάση το φύλο. Για το σκοπό αυτό έχουν γίνει διάφορες υποθέσεις από τους ερευνητές (England and Farkas, 1986, Kamo, 1988, Ishii-Kuntz and Coltrane, 1992, Cubbins and Vannoy, 2004).

Μια πολύ συχνή υπόθεση που γίνεται αφορά την ιδεολογία του άντρα και της γυναίκας στον τρόπο που επηρεάζει τα διαφορετικά ζευγάρια να μοιράζονται τις δουλειές «ρουτίνας» του σπιτιού. Με την προσέγγιση αυτή, οι ερευνητές αναφέρουν ότι οι άνθρωποι είναι συνήθως ελεύθεροι να υιοθετούν αξίες και πιστεύω σχετικές με το τι είναι «κατάλληλο» και τι δεν είναι για τους άντρες και τις γυναίκες και ότι αυτές παρακινούν ή εμποδίζουν τη συμμετοχή τους στις διάφορες δουλειές του σπιτιού. Επιπλέον, ορισμένοι αναφέρουν ότι η ιδεολογία των αντρών είναι περισσότερο ισχυρή στην κατανομή των εργασιών ενώ άλλοι πιστεύουν ότι η ιδεολογία των γυναικών είναι η πιο σημαντική.

Ο Πίνακας 2 είναι ένας $3\times3\times3$ πίνακας συνάφειας που προέκυψε από την διασταύρωση τριών κατηγορικών μεταβλητών: Ιδεολογία του Άντρα (Husband Ideology), Ιδεολογία της Γυναίκας (Wife Ideology) και τη Συνεισφορά στις Εργασίες του Σπιτιού (Sharing of Housework). Η μεταβλητή Ιδεολογία είναι κοινή για Άντρες και Γυναίκες και μετριέται από το βαθμό της Συμφωνίας/Διαφωνίας στην ερώτηση: «Είναι καλύτερα για όλους, εάν ο Άντρας είναι κυρίως υπεύθυνος για τα χρήματα μέσα στην οικογένεια και η Γυναίκα είναι αποκλειστικά υπεύθυνη για το νοικοκυριό και τα παιδιά». Η μεταβλητή Ιδεολογία του Άντρα και της Γυναίκας αποδίδεται σε τριτοβάθμια κλίμακα με τις εξής κατηγορίες:

1. Traditional (Αυτοί που συμφωνούν με την παραπάνω ερώτηση)
2. Moderate (Αυτοί που είναι ουδέτεροι, δηλαδή κάπου στη «μέση»)
3. Liberal (Αυτοί που διαφωνούν με την παραπάνω άποψη)

Η τρίτη μεταβλητή αφορά το βαθμό που ο Άντρας και η Γυναίκα, δηλαδή το ζευγάρι μοιράζονται τις δουλειές του σπιτιού, όπως το μαγείρεμα, τα ψώνια, το πλύσιμο και το σιδέρωμα, το καθάρισμα, δηλαδή ότι μπορεί να αφορά δουλειές «ρουτίνας». Οι κατηγορίες της μεταβλητής αυτής δίνονται επίσης με την εξής τριτοβάθμια κλίμακα:

1. Rarely (Σπάνια μοιράζομαι τις εργασίες στο σπίτι)
2. Sometimes (Μερικές φορές μοιράζομαι τις εργασίες στο σπίτι)
3. Often (Συχνά μοιράζομαι τις εργασίες στο σπίτι)

Το μέγεθος του δείγματος είναι $N = 1399$ ζευγάρια. Για να μοντελοποιήσουμε τις συχνότητες των κελιών του Πίνακα 2 αρχικά θα κατασκευάσουμε διάφορα ιεραρχικά λογαριθμογραμμικά μοντέλα κάνοντας τη παραδοχή ότι οι μεταβλητές είναι σε ονομαστική κλίμακα και στη συνέχεια, με βάση το μοντέλο εκείνο που θα είναι καλύτερο για τα δεδομένα, θα βρούμε μοντέλα συνάφειας θεωρώντας τις κατηγορικές μεταβλητές διατάξιμες όπου η διάταξη τους είναι άγνωστη και θα την εκτιμήσουμε, ενώ θα ελέγξουμε και την υπόθεση για την χρησιμοποίηση ή όχι ακεραίων σκορ για τις κατηγορίες τους.

Πίνακας 2: *Παρατηρούμενες συχνότητες - Ιδεολογία του Άντρα και της Γυναίκας και Συνεισφορά στις Εργασίες του Σπιτιού.*

		Share of Housework			
Husband Ideology	Wife Ideology	Rarely	Sometimes	Often	Total
Traditional	Traditional	73	137	43	253
	Moderate	38	89	32	159
	Liberal	13	59	35	107
Moderate	Traditional	21	69	28	118
	Moderate	36	121	49	206
	Liberal	17	103	72	192
Liberal	Traditional	8	21	6	35
	Moderate	8	56	47	111
	Liberal	17	95	106	218
Total		231	750	418	1399

Πηγή: Ishii-Kuntz (1994).

[**Σημ.:** Από εδώ και στο εξής θα συμβολίζουμε τις μεταβλητές με [H] για την Ιδεολογία του Άντρα, [W] για την Ιδεολογία της Γυναίκα και με [S] τη Συνεισφορά στις Εργασίες του Σπιτιού.]

Για τη διατύπωση και την προσαρμογή των μοντέλων θα αναλύσουμε τον παραπάνω πίνακα συμμετρικά, άρα θα θεωρήσουμε τις τρεις μεταβλητές σαν μεταβλητές απόκρισης και θα μεταχειριστούμε τις $N = IJK$ συχνότητες των κελιών του $I \times J \times K$ πίνακα, σαν ανεξάρτητες παρατηρήσεις που προέρχονται από δειγματοληψία *Poisson* με αναμενόμενες

συχνότητες m_{ijk} . Οι εκτιμήσεις των παραμέτρων των μοντέλων θα είναι εκτιμήσεις μέγιστης πιθανοφάνειας. Ωστόσο ίδιες εκτιμήσεις θα πάρουμε και αν θεωρήσουμε ότι έχουμε ανεξάρτητες πολυωνυμικές, απλώς θα πρέπει να θέσουμε διαφορετικούς περιορισμούς. Αν το μοντέλο της ανεξαρτησίας των μεταβλητών απορριφθεί, τότε μπορούμε να κατασκευάσουμε επιπλέον 7 λογαριθμογραμμικά μοντέλα χωρίς το κορεσμένο. Προφανώς όλα τα μοντέλα αυτά δεν είναι χρήσιμα για την ανάλυση και για τον λόγο αυτό ο Πίνακας 3 περιέχει μοντέλα με τουλάχιστον δύο από τις τρεις αλληλεπιδράσεις των παραγόντων.

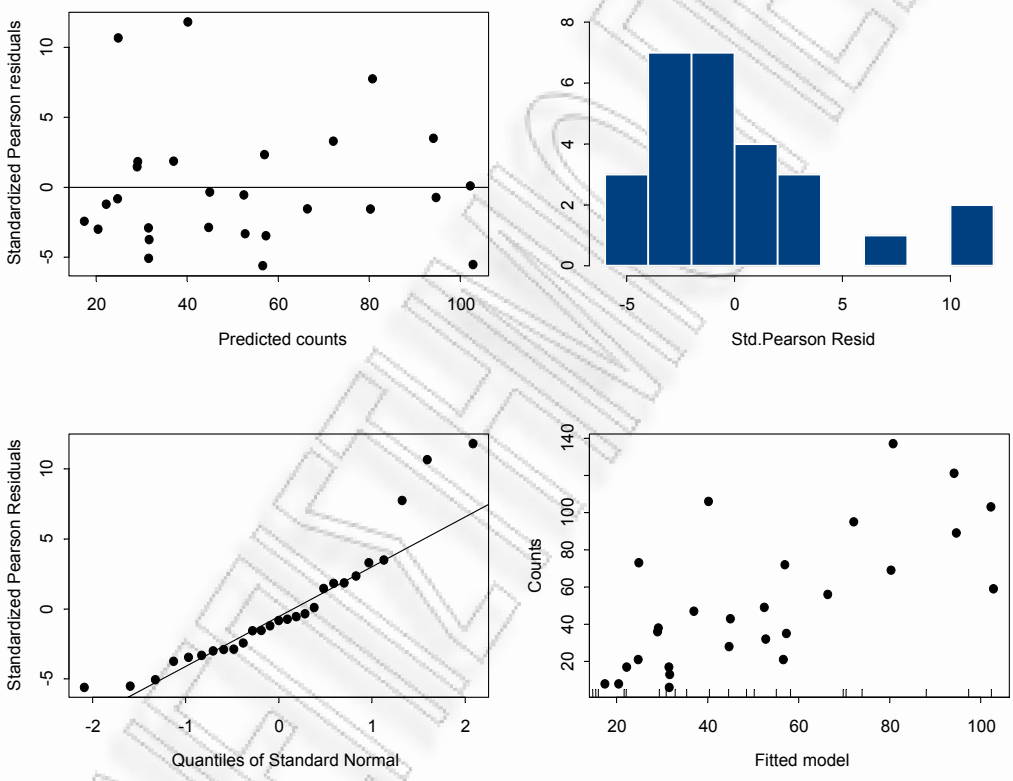
Πίνακας 3: *Διατύπωση και προσαρμογή ιεραρχικών λογαριθμογραμμικών μοντέλων στον Πίνακα 2.*

Μοντέλο	β.ε.	G^2	X^2
(H,W,S)	20	345.306	387.005
(SW,HW)	12	41.264	40.962
(SW,HS)	12	196.354	195.844
(SH,HW)	12	49.764	49.685
(WS,HW,HS)	8	9.698	9.184
(HWS)	0	0.000	0.000

Από τον Πίνακα 3 βλέπουμε ότι το μοντέλο της ανεξαρτησίας των μεταβλητών (H,W,S) δεν έχει καθόλου ικανοποιητική προσαρμογή αφού η στατιστική συνάρτηση του γενικευμένου λόγου πιθανοφανειών είναι $G^2 = 345.306$ με 20 βαθμούς ελευθερίας και η τιμή του στατιστικού του Pearson είναι $X^2 = 387.005$. Η διαπίστωση αυτή μπορεί να ελεγχθεί και γραφικά εξετάζοντας τις εκτιμημένες τιμές που δίνει το μοντέλο και τα κατάλοιπα του Pearson. Για παράδειγμα, το $1^{ο}$ γράφημα της Εικόνα 4.1 είναι το διάγραμμα διασποράς (scatter plot) των καταλοίπων ως προς τις προσαρμοσμένες συχνότητες από το οποίο διαφαίνεται ότι υπάρχει συστηματικό λάθος αφού έχουμε αρνητικά κατάλοιπα σε χαμηλές τιμές και θετικά στις υψηλότερες ενώ ξεχωρίζουν τρεις ακραίες παρατηρήσεις (όμως κάτι τέτοιο είναι αναμενόμενο για τέτοιου είδους δεδομένα). Από το ιστόγραμμα των καταλοίπων του $2^{ου}$ γραφήματος, δεν έχουμε ξεκάθαρη εικόνα για το αν είναι κανονικά κατανεμημένα. Πάντως για να εξετάσουμε καλύτερα αν όντως τα κατάλοιπα ακολουθούν προσεγγιστικά την

κανονική κατανομή, από το QQ-plot του 3ου γραφήματος φαίνεται πως εκτός από τρεις θετικές παρατηρήσεις και μία αρνητική τα κατάλοιπα του Pearson είναι κοντά στην ευθεία γραμμή άρα μπορούμε να θεωρήσουμε ότι προσεγγιστικά την ακολουθούν. Τέλος, στο 4ο γράφημα των παρατηρούμενων τιμών ως προς τις προσαρμοσμένες τιμές που δίνει το μοντέλο μπορούμε να δούμε ότι δεν προσαρμόζεται καθόλου καλά στα δεδομένα αφού υπάρχει μεγάλη διασπορά των τιμών.

Εικόνα 4.1: *Διαγνωστικά γραφήματα για το μοντέλο της ανεξαρτησίας του Πίνακα 3.*



Αφού το μοντέλο της ανεξαρτησίας απορρίπτεται δεχόμαστε ότι οι μεταβλητές συσχετίζονται και ότι υπάρχει τουλάχιστον ένας όρος αλληλεπίδρασης που είναι σημαντικός. Παρόλα αυτά, από τον Πίνακα 3 οι τιμές του G^2 και του X^2 δείχνουν ότι το μόνο μοντέλο που είναι ικανοποιητικό για τα δεδομένα είναι το (WS, HW, HS), δηλαδή το μοντέλο της «ομοιογενούς συνάφειας ή χωρίς την αλληλεπίδραση τριών παραγόντων» και ότι κανένα από τα άλλα μοντέλα της «δεσμευμένης ανεξαρτησίας» με τις δύο από τις τρεις αλληλεπιδράσεις είναι βάσιμα. Λαμβάνοντας υπόψη τη διαταξιμότητα που έχουν οι μεταβλητές του πίνακα και

βλέποντας ότι μόνο το μοντέλο (WS, HW, HS) είναι ικανοποιητικό, θα βρούμε μοντέλα συνάφειας για να αναλύσουμε τις μερικές αλληλεπιδράσεις αλλά και επιπλέον για να μελετήσουμε την τριπλή αλληλεπίδραση των παραγόντων χωρίς ωστόσο να είναι κορεσμένα. Το θέμα αυτό γενικά μπορούμε να το προσεγγίσουμε με δύο τρόπους: με τη χρήση ενός συστήματος ανάθεσης ακεραίων σκορ στις κατηγορίες των μεταβλητών ή με εκτίμηση των σκορ των κατηγοριών από τα δεδομένα για να καθοριστεί η διάταξη και ερμηνεία μέσω της απόστασης των κατηγοριών με βάση το *odds ratio*. Θα αναφερθούμε εδώ στην 2η περίπτωση που όπως είπαμε θα μπορέσουμε να ελέγξουμε και αν ευσταθή η χρήση των equal-interval σκορ. Για τα μοντέλα συνάφειας που μελετήθηκαν και που παρουσιάζονται στον Πίνακα 4, χρησιμοποιήθηκε μια συνιστώσα, αφού είναι $M_1 = M_2 = M_3 = 2$ και να μην καταλήξουμε στα αντίστοιχα τους λογαριθμογραμμικά μοντέλα σε κάθε περίπτωση. Έτσι θα είναι $M_1^* = M_2^* = M_3^* = 1$, απ' όπου θα εκτιμήσουμε μονοδιάστατα σκορ για τις κατηγορίες των μεταβλητών γραμμής, στήλης και στρώματος και το ίδιο ισχύει για την παράμετρο συνάφειας ϕ.

Πίνακας 4: *Μοντέλα συνάφειας για τον Πίνακα 2.*

Μοντέλο	β.ε.	G^2	p-value
Unrestricted $RC(1)+RL(1)+CL(1)$	11	14.281	0.218
Restricted $RC(1)+RL(1)+CL(1)$	14	17.341	0.238
$RCL(1)$ - (HW, HWS)	7	12.101	0.097
$RCL(1)$ - (HWS)	4	1.443	0.836
(HW, HS, WS, HWS)	13	17.816	0.165

Τα 2 πρώτα μοντέλα Πίνακα 4 είναι τα μοντέλα μερικής συνάφειας (partial association) και τα οποία αναλύουν όλες τις διπλές αλληλεπιδράσεις των παραγόντων χρησιμοποιώντας πολλαπλασιαστικούς όρους, όχι όμως την τριπλή. Τα μοντέλα αυτά είναι αντίστοιχα του μόνου αποδεκτού μοντέλου στον Πίνακα 3, (WS, HW, HS). Η απουσία του όρου της τριπλής αλληλεπίδρασης προϋποθέτει την ομοιογένεια στη συνάφεια και αυτό σημαίνει ότι τα odds

είναι σταθερά (uniform) σε όλα τα επίπεδα της τρίτης μεταβλητής. Το Unrestricted $RC(1)+RL(1)+CL(1)$ υπολογίζει σκορ για τις κατηγορίες των μεταβλητών χωρίς να έχουμε θέσει κάποιους περιορισμούς σε αυτά και συγκεκριμένα για τον Πίνακα 2 βρίσκει συνολικά 6 διανύσματα από σκορ (για τις γραμμές, στήλες και τα στρώματα). Αντιθέτως, στο Restricted $RC(1)+RL(1)+CL(1)$ έχουν τεθεί περιορισμοί ομοιογένειας στα σκορ των μεταβλητών και υπολογίζει 3 διανύσματα από σκορ. Στη περίπτωση που θέλουμε να μελετήσουμε πώς η ιδεολογία των αντρών και η ιδεολογία των γυναικών μπορεί να αλλάξει κατά κάποιο τρόπο με τη συνεισφορά στις εργασίες του σπιτιού, τότε πρέπει να μοντελοποιήσουμε την τριπλή αλληλεπίδραση. Αυτό όπως είδαμε δεν μπορεί να γίνει για τα λογαριθμογραμμικά μοντέλα αφού το (HWS) είναι κορεσμένο. Όμως με τα μοντέλα συνάφειας του Πίνακα 4 είναι εφικτό. Για τον λόγο αυτό, το μοντέλο της υπό συνθήκης συνάφειας PARAFAC/CANDECOMP (HW, HWS), αναλύει την ετερογενή συνάφεια [HW] και [S] εισάγοντας έναν τριγραμμικό όρο και έναν διγραμμικό και υπολογίζει τα σκορ των κατηγοριών για κάθε μεταβλητή ξεχωριστά. Επίσης είναι δυνατόν να μελετήσουμε μόνο τον όρο της τριπλής αλληλεπίδρασης χωρίς το μοντέλο να είναι κορεσμένο χρησιμοποιώντας μόνο τριγραμμικούς όρους για τη διάσπαση της αλληλεπίδρασης. Αυτό γίνεται με το μοντέλο (HWS) που επίσης η συνάφεια στους μερικούς πίνακες δεν είναι ομοιόμορφη. Τέλος υπάρχει μη κορεσμένο μοντέλο για τρισδιάστατους πίνακες συνάφειας όπου όλες οι αλληλεπιδράσεις μεγαλύτερης και μικρότερης τάξης αναλύονται και όπως φαίνεται από τον Πίνακα 4, είναι το μοντέλο $RCL(1)$ - (HW, HS, WS, HWS). Με το μοντέλο αυτό μας ενδιαφέρει να εξηγήσουμε για παράδειγμα, πώς η τάση που έχουν οι Liberal γυναίκες να συνεισφέρουν πιο Συχνά στις εργασίες απ' ότι οι non-Liberal είναι μεγαλύτερη ή μικρότερη σε σχέση όταν οι άντρες είναι λιγότερο Liberal;

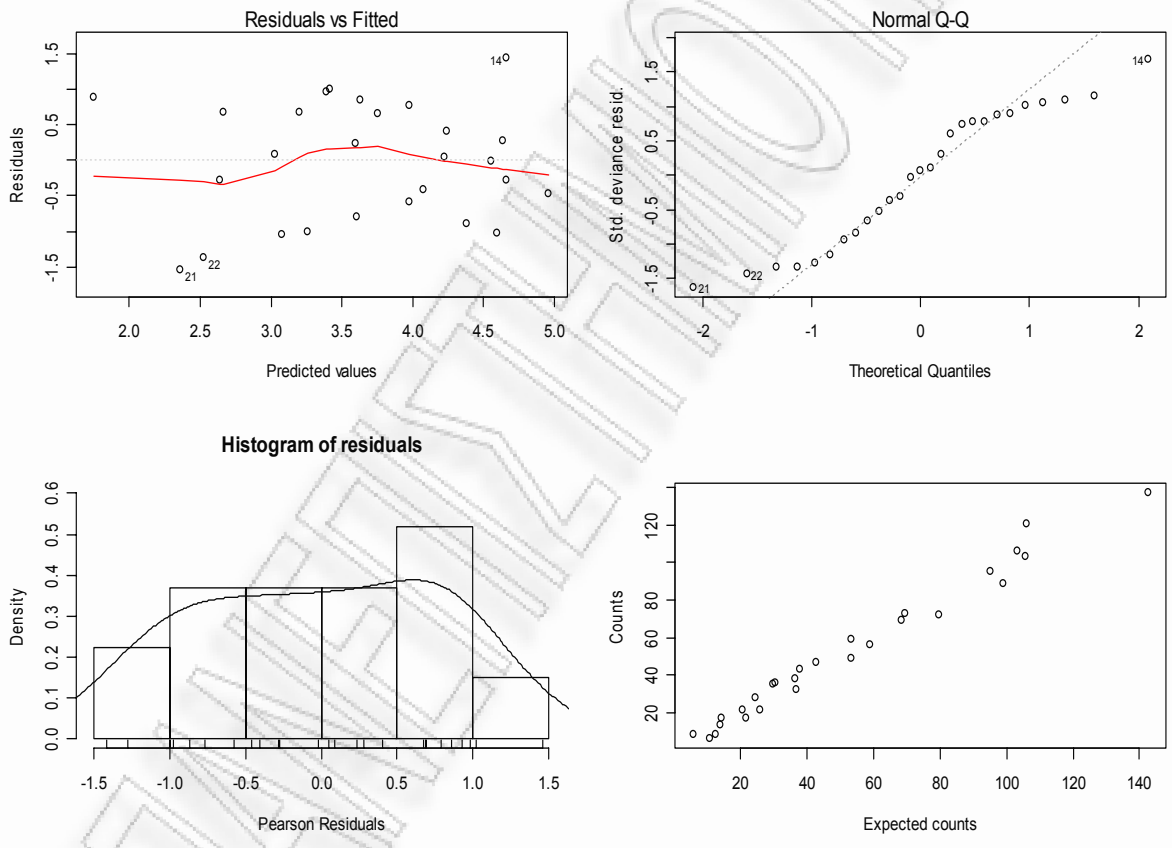
Από τα μοντέλα συνάφειας του Πίνακα 4 θα δώσουμε την ερμηνεία για το Restricted $RC(1)+RL(1)+CL(1)$ αφού από τον βασικό Πίνακα 3 της ανάλυσης για την εύρεση καλύτερου μοντέλου που να προσαρμόζεται ικανοποιητικά στα δεδομένα, μόνο το μοντέλο «χωρίς την αλληλεπίδραση τριών παραγόντων» (WS, HW, HS), είναι βάσιμο. Η προσαρμογή του μοντέλου είναι ικανοποιητική $\left(G^2 = 17.341 \text{ , p-value} = 0.238, 14 \text{ df}\right)$. Οι εκτιμήσεις μέγιστης πιθανοφάνειας για τις παραμέτρους του μοντέλου μας δίνουν ότι τα σκορ για τις

γραμμές είναι $\hat{\mu}_{i1} = \{1.177,-0.241,-1.337\}$, για τις στήλες $\hat{\nu}_{j1} = \{-1.289,-0.174,1.172\}$ και για τα στρώματα $\hat{\xi}_{k1} = \{-1.600,-0.257,1.346\}$, όπου μ_{11} = Traditional Husband … μ_{13} = Liberal Husband Ideology, ν_{11} = Traditional Wife … ν_{13} = Liberal Wife, ξ_{11} = Rarely … ξ_{13} = Often. Η εκτίμηση της συνάφεια για κάθε ζεύγος μεταβλητών είναι $\hat{\phi}^{HW} = 0.299$, $\hat{\phi}^{HS} = 0.237$ και $\hat{\phi}^{WS} = 0.178$, αντίστοιχα.

Από τις εκτιμήσεις των σκορ μπορούμε να βρούμε τις αποστάσεις μεταξύ των κατηγοριών και ύστερα να υπολογίσουμε το σχετικό πηλίκο των αποστάσεων $r_i = d_i / d_{i+1}$ και να βγάλουμε επιπλέον συμπεράσματα. Θα πρέπει εδώ να σημειώσουμε ότι για το μοντέλο που συζητάμε έχει την ιδιότητα να παραμένει αναλλοίωτο στις εναλλαγές των κατηγοριών των μεταβλητών. Δηλαδή η προσαρμογή του δεν αλλάζει, ούτε οι εκτιμήσεις των $\hat{\phi}_i$, ούτε το πηλίκο των αποστάσεων r_i, αν αλλάξουμε τη διάταξη των κατηγοριών. Έτσι βάσει των εκτιμήσεων που βρήκαμε για τις κατηγορίες όλων των μεταβλητών, για την μεταβλητή Ιδεολογία του άντρα, η απόσταση ανάμεσα στην πρώτη και τη δεύτερη κατηγορία είναι -1.417 και για την δεύτερη και τρίτη κατηγορία είναι -1.096. Αυτό σημαίνει ότι η πρώτη και τρίτη κατηγορία της μεταβλητής αυτής θα πρέπει να αντιστραφούν, για να υπάρχει σωστή διάταξη, αν αυτό είναι εφικτό να γίνει από τη κωδικοποίηση των κατηγοριών της μεταβλητής. Το σχετικό πηλίκο των αποστάσεων τους είναι ίσο με 1.294, άρα η απόσταση μεταξύ των κατηγοριών, Traditional και Moderate, είναι περίπου 1,3 φορές την απόσταση μεταξύ των κατηγοριών Moderate και Liberal. Για την μεταβλητή Ιδεολογία της γυναίκας, οι μεταξύ αποστάσεις των διαδοχικών κατηγοριών είναι 1.115 και 1.346 αντίστοιχα, ενώ το πηλίκο των αποστάσεων αυτών είναι 0.828. Συνεπώς η απόσταση μεταξύ των κατηγοριών Traditional και Moderate, είναι 0,8 φορές την απόσταση μεταξύ των κατηγοριών Moderate και Liberal. Τέλος, για την μεταβλητή Συνεισφορά, η απόσταση μεταξύ της πρώτης και δεύτερης κατηγορίας είναι 1.342 και ανάμεσα στη δεύτερη και τρίτη κατηγορία είναι 1.604. Το πηλίκο των αποστάσεων αυτών είναι 0.837 συνεπώς η απόσταση μεταξύ των κατηγοριών, Rarely και Sometimes είναι επίσης περίπου 0.83 φορές την απόσταση μεταξύ των κατηγοριών, Sometimes και Often. Άρα μπορούμε να πούμε ότι σε πρώτο επίπεδο η χρήση equal-interval σκορ για τις κατηγορίες των μεταβλητών Συνεισφορά και Ιδεολογία της Γυναίκας είναι ορθή και αναμένουμε να πάρουμε παρόμοια αποτελέσματα από την προσαρμογή ενός τέτοιου restricted μοντέλου (μοντέλο 13) αφού τα εκτιμημένα σκορ παρουσιάζουν μονοτονικότητα, αρκεί τα *a priori* σκορ που θα

θέσουμε να έχουν παρόμοιες αποστάσεις με εκείνες που υπολογίσαμε. Επιπλέον για τη μεταβλητή Ιδεολογία του Άντρα, ίσως η διάταξη της να πρέπει να αλλάξει σε $\mu_{i1}^{*} = \left\{ \mu_{i3}^{*}, \mu_{i2}^{*}, \mu_{i1}^{*} \right\}$. Παρόλα αυτά με την υπάρχουσα μονότονη διάταξη μπορούμε να χρησιμοποιήσουμε equal-interval σκορ για τις κατηγορίες της αλλά ίσως με μια διαφορετική απόσταση από τις άλλες δύο μεταβλητές. Εν κατακλείδι, και στις τρεις μεταβλητές του Πίνακα 2 μπορούμε να κάνουμε χρήση ιδίων σκορ (πχ. ακεραίων τιμών) με ίσες αποστάσεις μόνο αν οι κατηγορίες των γραμμών αλλάξουν διάταξη.

Εικόνα 4.2 : *Διαγνωστικά γραφήματα για το Restricted* $RC(1)+RL(1)+CL(1)$ *μοντέλο του Πίνακα 4.*



Από τα γραφήματα της Εικόνας 4.2 ελέγχουμε το μοντέλο και βλέπουμε ότι τα κατάλοιπα ακολουθούν προσεγγιστικά την κανονική κατανομή. Επίσης από το $4^{ο}$ γράφημα των παρατηρούμενων και προσαρμοσμένων τιμών παρατηρούμε ότι είναι αρκετά συμπαγές, απόδειξη της καλής προσαρμογής του μοντέλου. Δηλαδή οι προσαρμοσμένες τιμές που δίνει το μοντέλο είναι πολύ κοντά στις πραγματικές.

Η ερμηνεία του μοντέλου γίνεται για τις μερικές συνάφειες και αποδίδεται μέσω των log-odds και των odds ratios. Έτσι από τις εκτιμημένες παραμέτρους του μπορούμε να πούμε για την υπό συνθήκη συνάφεια [WS], πως αφού είναι $\hat{\phi}^{WS} = 0.178$ και τα εκτιμημένα σκορ είναι μονότονα με αύξουσα διάταξη, η συνάφεια αυτή είναι παντού θετική. Αυτό σημαίνει ότι τα log-odds και αντίστοιχα τα odds ratios σε κάθε πίνακα είναι παντού θετικά. Για την μεταβλητή [H] τα εκτιμημένα σκορ γραμμών είναι σε φθίνουσα διάταξη. Αυτό έχει σαν αποτέλεσμα οι μερικές συνάφειες [HS] και [HW] να είναι παντού αρνητικές για διαδοχικές κατηγορίες. Άρα τα log-odds είναι αρνητικά και τα odds ratios μικρότερα της μονάδας. Αν αλλαχθεί η διάταξη της [H], τότε οι σχέσεις αυτές θα γίνουν παντού θετικές. Έτσι, λαμβάνοντας υπόψη τη συνάφεια των μεταβλητών [W] και [S], στην περίπτωση αυτή, οι εκτιμημένες συχνότητες του πίνακα θα είναι ισοτροπικές (Goodman, 1981b).

Για παράδειγμα, για την ερμηνεία της [HS] συνάφειας όπου τα $\{\hat{\mu}_i\}$ είναι σε φθίνουσα διάταξη και τα $\{\hat{\xi}_k\}$ σε αύξουσα μπορούμε να υποθέσουμε ότι η ιδεολογία των αντρών μετατοπίζεται κατά κάποιο τρόπο προς την πιο Liberal θέση όταν Σπάνια οι άντρες συνεισφέρουν στις δουλειές του σπιτιού και αντίστροφα. Έτσι τα *odds* των αντρών που έχουν Traditional ιδεολογία και συνεισφέρουν στις δουλειές του σπιτιού πιο Συχνά είναι περίπου 5.8 φορές μεγαλύτερα από τα αντίστοιχα *odds* των αντρών που έχουν Liberal ιδεολογία. Επιπλέον ο βαθμός συνάφειας είναι ομοιογενής *εγκαρσίως* (across) στα επίπεδα της μεταβλητής [W].

ΚΕΦΑΛΑΙΟ 5

Ανάλυση Αντιστοιχιών και Μοντέλα Συσχέτισης

5.1 Γενικά

Η ανάλυση αντιστοιχιών είναι μια μέθοδος κατάλληλη για το χειρισμό κατηγορικών δεδομένων που δίνονται με τη μορφή πίνακα συνάφειας. Η μέθοδος αυτή έχει κυρίως περιγραφικό και διερευνητικό χαρακτήρα και είναι πολύ ευσταθής (robust). Χαρακτηρίζεται ως «model-free» μέθοδος αφού βασίζεται σε λίγες υποθέσεις και σε αντίθεση με τις «model-based» προσεγγίσεις δεν θέτει εξαρχής κάποιο μοντέλο και στη συνέχεια προσπαθεί να εκτιμήσει τις παραμέτρους του. Για την εύρεση της δομής των δεδομένων του πίνακα συνάφειας, προσπαθούμε να αναπαραστήσουμε τις γραμμές και τις στήλες με σημεία στο χώρο λίγων διαστάσεων. Έτσι, από τη γεωμετρική απεικόνιση των μεταβλητών του πίνακα αποκτάμε περισσότερη πληροφορία για τις σχέσεις που συνδέουν τις κατηγορίες των διακριτών μεταβλητών.

Σε γενικές γραμμές με την ανάλυση αντιστοιχιών τα αποτελέσματα που αναμένουμε είναι τα εξής: Συσχετισμός μεταξύ γραμμών και στηλών. Σημεία πάνω στο γράφημα που είναι γειτονικά μεταξύ τους υποδηλώνουν εκτός των άλλων και συσχετισμό ανάμεσα στις αντίστοιχες γραμμές/στήλες. Επίσης περιμένουμε κάποιο είδος διάταξης των κατηγοριών. Επιπλέον μπορούμε να ελέγξουμε το κατά πόσο υπάρχουν διαφορές μεταξύ γραμμών και στηλών. Αυτό συνδέεται με το θέμα της ανεξαρτησίας δύο μεταβλητών. Για τον σκοπό αυτό, η μέθοδος αναπαριστά το χ^2 τεστ ανεξαρτησίας γραφικά διασπώντας το σε M όρους ή αδράνειες και χρησιμοποιώντας την χ^2 απόσταση ανάμεσα στις κατηγορίες των μεταβλητών ελέγχει με αυτόν τον τρόπο την ανεξαρτησία των γραμμών και των στηλών του πίνακα συνάφειας. Για παράδειγμα, όταν μια από τις κατηγορίες της μεταβλητής γραμμής που απεικονίζεται γραφικά με το σημείο $R1$, είναι κοντά σε μια από τις κατηγορίες της μεταβλητής στήλης που απεικονίζεται με το σημείο $C1$, τότε ο συνδυασμός αυτός έχει μεγαλύτερη συχνότητα εμφάνισης από ότι στην περίπτωση της ανεξαρτησίας. Παράλληλα, οι

άξονες που αντιπροσωπεύουν τις κύριες αδράνειες μπορούν να θεωρηθούν σαν νέες μεταβλητές οι οποίες έχουν μια φυσική ερμηνεία και μπορούν να χρησιμοποιηθούν για περαιτέρω στατιστικές αναλύσεις.

Για την επιλογή μεταξύ μεθόδων «model-free» όπως είναι η ανάλυση αντιστοιχιών και «model-based» όπως είναι τα λογαριθμογραμμικά μοντέλα, προτείνεται η επιλογή της ανάλυσης αντιστοιχιών όταν έχουμε μεγάλους πίνακες αφού η δομή των κατηγοριών των διαφορετικών μεταβλητών είναι δύσκολη να βρεθεί με μια απλή παρατήρηση ή με απλές στατιστικές μεθόδους και των λογαριθμογραμμικών μοντέλων όταν οι μεταβλητές ταξινόμησης του πολυδιάστατου πίνακα συνάφειας έχουν λίγες κατηγορίες. Επίσης θα πρέπει να αναφέρουμε ότι μια σημαντική διαφορά μεταξύ των δύο μεθόδων ανάλυσης είναι ότι τα λογαριθμογραμμικά μοντέλα κυρίως εξετάζουν τις συσχετίσεις των μεταβλητών του πίνακα ενώ η ανάλυση αντιστοιχιών εξετάζει τις σχέσεις μεταξύ των κατηγοριών των μεταβλητών.

Η ανάλυση αντιστοιχιών προκύπτει ως το αποτέλεσμα της διάσπαση ιδιόμορφων τιμών (Singular Value Decomposition SVD) του πίνακα καταλοίπων και τα βασικά αποτελέσματα της τεχνικής αυτής είναι οι ιδιόμορφες τιμές (singular values) και τα ιδιόμορφα διανύσματα. Παρόμοια αποτελέσματα με την ανάλυση αντιστοιχιών επιτυγχάνονται και με τα μοντέλα συσχέτισης τα οποία συμπληρώνουν και επεκτείνουν την μέθοδο.

5.2 Ανάλυση Αντιστοιχιών για Πίνακες Συνάφειας 2 Διαστάσεων

5.2.1 Εισαγωγή

Στο τμήμα αυτό του κεφαλαίου η ανάλυση αντιστοιχιών θα παρουσιαστεί σαν μια τεχνική που μπορεί να χρησιμοποιηθεί για την μελέτη των συσχετισμών μεταξύ γραμμών και στηλών σε έναν $R \times C$ πίνακα συνάφειας δύο μεταβλητών. Όπως είδαμε και στο Κεφάλαιο 4, τα κελιά ενός πίνακα συνάφειας περιέχουν συχνότητες n_{ij}, $i = 1,\ldots,R$, $j = 1,\ldots,C$, για όλους τους συνδυασμούς των γραμμών και των στηλών. Στη θεωρία της ανάλυσης αντιστοιχιών ο πίνακας αυτός μετατρέπεται σε ένα πίνακα που τα κελιά του περιέχουν τις παρατηρούμενες σχετικές συχνότητες p_{ij}, όπου $p_{ij} = n_{ij} / n$ και n είναι το ολικό μέγεθος του δείγματος.

5.2.2 Μάζες και Προφίλ

Ο πίνακας που μόλις περιγράψαμε ονομάζεται πίνακας αντιστοιχιών (*correspondence matrix*) και συμβολίζεται με **P**

$$\mathbf{P} = \left(p_{ij} \right) = \left(n_{ij} / n \right).$$

Στον πίνακα **P** οι περιθώριες κατανομές των γραμμών και των στηλών δίνονται από τις σχέσεις $p_{i\bullet} = n_{i\bullet}/n$ και $p_{\bullet j} = n_{\bullet j}/n$ αντίστοιχα, όπου $n_{i\bullet}$ και $n_{\bullet j}$ είναι οι περιθώριες συχνότητες. Το $(r \times 1)$ διάνυσμα της περιθώριας κατανομής των γραμμών $p_{i\bullet}$, συμβολίζεται με **r** και επίσης το $(c \times 1)$ διάνυσμα της περιθώριας κατανομής των στηλών $p_{\bullet j}$ συμβολίζεται με **c**. Τότε τα διανύσματα **r** και **c** ονομάζονται μάζες (*masses*) των γραμμών και των στηλών αντίστοιχα. Θα πρέπει το άθροισμα όλων των στοιχείων του διανύσματος **r** να ισούται με τη μονάδα και αντίστοιχα για τα στοιχεία του **c**. Μπορούμε να κατασκευάσουμε διαγώνιους πίνακες με τα στοιχεία των μαζών και να τους συμβολίσουμε $\mathbf{D}_r (r \times r)$ και $\mathbf{D}_c (c \times c)$, οι οποίοι έχουν ιδιαίτερη σημασία για το αντικείμενο της ανάλυσης αντιστοιχιών όπου με μορφή πινάκων παριστάνονται ως εξής

$$\mathbf{D}_\mathbf{r} = \text{diag}(\mathbf{r}) = \begin{pmatrix} p_{1\bullet} & & 0 \\ & \ddots & \\ 0 & & p_{r\bullet} \end{pmatrix} \text{ και } \mathbf{D}_\mathbf{c} = \text{diag}(\mathbf{c}) = \begin{pmatrix} p_{\bullet 1} & & 0 \\ & \ddots & \\ 0 & & p_{\bullet c} \end{pmatrix} \quad (1)$$

Στη συνέχεια μετατρέπουμε τις γραμμές και τις στήλες του πίνακα **P** σε προφίλ. Χρησιμοποιώντας την δεσμευμένη κατανομή για τις γραμμές $n(j/i) = n_{ij}/n_{i\bullet}$ ή αφού μιλάμε για τον πίνακα **P** την $p(j/i) = p_{ij}/p_{i\bullet}$, συμβολίζουμε με $\mathbf{r}_i$, $i = 1, \ldots, R$ και $\forall j \in C$, το προφίλ της γραμμής i. Αντίστοιχα το διάνυσμα με τις δεσμευμένες κατανομές των στηλών $p(i/j) = p_{ij}/p_{\bullet j}$ το οποίο συμβολίζεται με $\mathbf{c}_j$, $j = 1, \ldots, C$ και $\forall i \in R$, ονομάζεται προφίλ της στήλης j. Αφού τα $p(j/i)$ και $p(i/j)$ είναι οι δεσμευμένες σχετικές συχνότητες πρέπει το άθροισμα των στοιχείων κάθε γραμμής και το άθροισμα των στοιχείων κάθε στήλης να είναι ίσο με την μονάδα. Κάθε προφίλ γραμμής και στήλης είναι δυνατόν να αναπαρασταθεί με ένα σημείο στο χώρο αφού θεωρούνται διανύσματα και για την αναπαράσταση τους πρέπει να βρούμε τις συντεταγμένες τους.

Μπορούμε να κατασκευάσουμε ολόκληρο τον πίνακα $\mathbf{R}(r \times c)$ με τα προφίλ των γραμμών του **P** μέσω της σχέσης (1) γράφοντας $\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P}$. Επίσης μπορούμε να κατασκευάσουμε

ολόκληρο τον πίνακα $\mathbf{C}(c \times r)$ με τα προφίλ των γραμμών πάλι μέσω της σχέσης (1) από τις ποσότητες: $\mathbf{C} = \mathbf{D}_c^{-1}\mathbf{P}^{\mathbf{T}}$ ή $\mathbf{C} = \mathbf{P}\mathbf{D}_c^{-1}$.

5.2.3 Μέσο Προφίλ και Απόκλιση από την Ανεξαρτησία

Ο σταθμισμένος μέσος των προφίλ γραμμής με σταθμά το σύνολο των παρατηρήσεων κάθε γραμμής λέγεται μέσο προφίλ γραμμής ή κεντροειδές (centroid) γραμμών και δίνεται από την ποσότητα $\sum_{i=1}^{r} \mathbf{r}_i n_{i\cdot} / n = \mathbf{c}$, το οποίο είναι το διάνυσμα των μαζών των στηλών. Παρόμοια, ο σταθμισμένος μέσος των προφίλ των στηλών με σταθμά το σύνολο των παρατηρήσεων κάθε στήλης λέγεται μέσο προφίλ στήλης ή κεντροειδές στηλών και δίνεται από την ποσότητα $\sum_{j=1}^{c} \mathbf{c}_j n_{\cdot j} / n = \mathbf{r}$ και το οποίο είναι το διάνυσμα των μαζών των γραμμών. Το κεντροειδές μετέπειτα τοποθετείται στην αρχή (origin) των κυρίων αξόνων (principal axes) για τη γεωμετρική απεικόνιση των προφίλ των μεταβλητών. Εάν κάποιο προφίλ είναι πολύ διαφορετικό από το μέσο προφίλ, το σημείο αυτό στο χάρτη θα απέχει αρκετά από την αρχή των αξόνων, με άλλα λόγια θα είναι απομακρυσμένο, ενώ για τα προφίλ που είναι κοντά στο μέσο προφίλ θα αντιπροσωπεύονται με σημεία στο χάρτη κοντά στο κεντροειδές. Αν πάλι όλες οι κατηγορίες έχουν ίδια προφίλ τότε όλα τα σημεία θα πέσουν πάνω στο κεντροειδές.

Για να αποφανθούμε την ανεξαρτησία των μεταβλητών θεωρητικά τουλάχιστον, θα πρέπει να ισχύει ότι $\mathbf{r}\mathbf{c}^{'} = \mathbf{P}$. Άρα ο πίνακας $\left(\mathbf{P} - \mathbf{r}\mathbf{c}^{'}\right)$ είναι ένα μέτρο της απόκλισης από την ανεξαρτησία. Παρόμοια, μπορεί κανείς να συγκρίνει τα προφίλ των γραμμών και των στηλών με τις μάζες των στηλών και των γραμμών αντίστοιχα, για να κρίνει την απόκλιση από την ανεξαρτησία. Δηλαδή να συγκρίνει το $\mathbf{r}_i$ με το $\mathbf{c}$ και το $\mathbf{c}_j$ με το $\mathbf{r}$.

5.2.4 Μέτρα Απόστασης - Σχέση με το χ^2 τεστ Ανεξαρτησίας του Pearson

Ένα σημαντικό θέμα σχετικά με τα προφίλ είναι το πόσο διαφορετικά αυτά είναι μεταξύ τους (εδώ έχει έννοια η διαφοροποίηση μόνο των προφίλ του ίδιου σετ μεταβλητών) όσο και με το μέσο προφίλ. Για το λόγο αυτό χρησιμοποιώντας κάποιο μέτρο απόστασης μεταξύ των προφίλ γραμμών και μεταξύ των προφίλ στηλών, μπορεί κανείς να ποσοτικοποιήσει τον βαθμό διαφοροποίησης μεταξύ δύο γραμμών ή μιας γραμμής και του μέσου, δηλαδή να μετρήσει την διαφορά τους.

Όπως είναι γνωστό, για να μετρήσουμε την απόσταση μεταξύ δύο παρατηρήσεων μπορούμε να χρησιμοποιήσουμε την Ευκλείδεια απόσταση. Το μέτρο αυτό αν και είναι το πιο συνηθισμένο έχει ένα σοβαρό μειονέκτημα. Το μειονέκτημα του είναι ότι δε λαμβάνει υπόψη τον αριθμό των παρατηρήσεων σε κάθε κελί του πίνακα συνάφειας με αποτέλεσμα διαφορές σε κελιά με μικρές συχνότητες να έχουν την ίδια βαρύτητα με διαφορές σε κελιά με μεγάλες συχνότητες στον τελικό υπολογισμό της απόστασης.

Μια λύση στο πρόβλημά μας είναι να χρησιμοποιήσουμε την σταθμισμένη (*weighted*) Ευκλείδεια απόσταση με σταθμά το αντίστροφο των αντίστοιχων μαζών ή κεντροειδές στον υπολογισμό της απόστασης. Με τον τρόπο αυτό, οι κατηγορίες που έχουν λίγες παρατηρήσεις συνεισφέρουν σχετικά περισσότερο στον υπολογισμό των αποστάσεων από ότι οι κατηγορίες που έχουν πολλές παρατηρήσεις. Αυτό το μέτρο απόστασης ονομάζεται χ^2 απόσταση. Επίσης μπορούμε να δούμε ότι η απόσταση μεταξύ μιας γραμμής από τον μέσο ομοιάζει με τον χ^2 έλεγχο ανεξαρτησίας, αφού κάθε όρος είναι το τετράγωνο της διαφοράς της παρατηρούμενης σχετικής συχνότητας από την αναμενόμενη, εάν θεωρήσουμε ως αναμενόμενη τιμή τον μέσο όρο, προς την αναμενόμενη σχετική συχνότητα. Αυτό μπορεί να δειχθεί ως εξής:

Το στατιστικό του *Pearson* για τον έλεγχο ανεξαρτησίας δύο κατηγορικών μεταβλητών όπως γνωρίζουμε δίνεται από την σχέση

$$X^2 = \sum_i \sum_j \frac{\left(n_{ij} - n_{i\bullet} n_{\bullet j}/n\right)^2}{n_{i\bullet} n_{\bullet j}/n} \tag{2}$$

και το οποίο ασυμπτωτικά ακολουθεί την χ^2 κατανομή με $(r-1)(c-1)$ βαθμούς ελευθερίας. Την σχέση (2) μπορούμε να την γράψουμε εναλλακτικά ως εξής

$$X^2 = \sum_i n_{i\bullet} \left[\sum_j \left(\frac{n_{ij}}{n_{i\bullet}} - n_{\bullet j}/n \right)^2 / n_{\bullet j}/n \right] \tag{3}$$

$$X^2 = \sum_j n_{\bullet j} \left[\sum_i \left(\frac{n_{ij}}{n_{\bullet j}} - n_{i\bullet}/n \right)^2 / n_{i\bullet}/n \right] \tag{4}$$

Στην (3) σε κάθε γραμμή i, το τετράγωνο των αποκλίσεων κάθε προφίλ γραμμής για την στήλη j διαιρείται με την περιθώρια κατανομή της στήλης j και το αποτέλεσμα αθροίζεται για όλες τις στήλες. Έτσι σε κάθε γραμμή i, το άθροισμα μας δίνει ένα σταθμισμένο μέσο για τετράγωνα των αποκλίσεων των προφίλ πάνω στις στήλες. Αφού τα σταθμά των στηλών είναι

το αντίστροφο των μαζών τους, μεγάλες αποκλίσεις που λαμβάνουν χώρα στις στήλες με μικρές μάζες έχουν μεγαλύτερο βάρος στον υπολογισμό του μέσου. Τέλος, τον σταθμισμένο μέσο τον πολλαπλασιάζουμε με την μάζα των γραμμών και παίρνουμε μία συνολική τετραγωνική απόκλιση για την γραμμή. Η τιμή αυτή αθροίζεται για όλες τις γραμμές και αποτελεί το χ^2 τεστ ανεξαρτησίας του *Pearson*. Παρόμοια ερμηνεία μπορεί να δοθεί και για την σχέση (4) όσον αφορά τις στήλες. Θα πρέπει εδώ να σημειωθεί ότι η απόσταση μπορεί να υπολογιστεί μόνο ανάμεσα στις κατηγορίες της ίδιας μεταβλητής και όχι ανάμεσα στις κατηγορίες διαφορετικών μεταβλητών.

5.2.5 Ολική Αδράνεια και χ^2 απόσταση

Σε γενικές γραμμές, για την γραφική απεικόνιση των σημείων, προφίλ με μικρή τιμή στην απόσταση θα πρέπει να αντιπροσωπεύονται με σημεία που είναι κοντινά μεταξύ τους και προφίλ με μεγάλη τιμή για το μέτρο της απόστασης θα πρέπει να αντιπροσωπεύονται με σημεία που είναι απομακρυσμένα. Για να μετρήσουμε τις διαφορές σε κάθε ζεύγος σημείων ορίζουμε ένα συνολικό μέτρο ομοιογένειας ή ετερογένειας των προφίλ ανάλογο της διακύμανσης που ονομάζουμε αδράνεια (inertia). Με άλλα λόγια η αδράνεια μετράει την μεταβλητότητα των σημείων που απεικονίζονται πάνω στον χάρτη. Όσο πιο συγκεντρωμένα είναι μεταξύ τους τα σημεία τόσο μικρότερη είναι και η αδράνεια.

Η σχέση της αδράνειας και του χ^2 τεστ είναι η εξής: από την (2) έχουμε ότι

$$n^{-1}X^2 = n^{-1}\sum_i \sum_j \frac{\left(n_{ij} - n_{i\bullet}n_{\bullet j}/n\right)^2}{n_{i\bullet}n_{\bullet j}/n}$$

και με τη βοήθεια της (3) γράφουμε

$$\begin{aligned} n^{-1}X^2 &= \sum_i n_{i\bullet}\left[\sum_j \left(\frac{n_{ij}}{n_{i\bullet}} - n_{\bullet j}/n\right)^2 / n_{\bullet j}\right] \\ &= \sum_i n_{i\bullet}\left(\mathbf{r}_i - \mathbf{c}\right)'\mathbf{D}_c^{-1}\left(\mathbf{r}_i - \mathbf{c}\right) \end{aligned} \quad (5)$$

Παρόμοια, από τις σχέσεις (2) και (4) παίρνουμε αντίστοιχη σχέση για τις στήλες

$$n^{-1}X^2 = \sum_j n_{\bullet j}\left(\mathbf{c}_j - \mathbf{r}\right)'\mathbf{D}_r^{-1}\left(\mathbf{c}_j - \mathbf{r}\right) \quad (6)$$

όπου τα $\mathbf{r}, \mathbf{c}, \mathbf{r}_i, \mathbf{c}_j, \mathbf{D}_r$ και $\mathbf{D}_c$ ορίστηκαν προηγουμένως. Τότε γενικά, το στατιστικό X^2/n ονομάζεται ολική αδράνεια.

Η χ^2 απόσταση ανάμεσα σε δύο προφίλ γραμμών $\mathbf{r}_i$ και $\mathbf{r}_j$ δίνεται από τη σχέση

$$d_{ij}^2 = (\mathbf{r}_i - \mathbf{r}_j)' \mathbf{D}_c^{-1} (\mathbf{r}_i - \mathbf{r}_j)$$

και ανάλογα υπολογίζεται η απόσταση μεταξύ δύο προφίλ των στηλών. Εάν δύο σημεία γραμμών ή στηλών είναι πολύ κοντά, τα προφίλ των γραμμών (στηλών) είναι παρόμοια. Τότε μπορούμε να ενοποιήσουμε τις 2 κατηγορίες σε μία και να βελτιώσουμε την προσέγγιση της τιμής του χ^2. Θα πρέπει και πάλι να τονίσουμε ότι οι αποστάσεις αυτές έχουν νόημα για τις κατηγορίες της ίδιας μεταβλητής και όχι για τις κατηγορίες διαφορετικών μεταβλητών.

Από τις (5) και (6) μπορούμε να δούμε ότι η ολική αδράνεια μπορεί να οριστεί ξεχωριστά για τις γραμμές και τις στήλες ενώ λόγω της συμμετρίας του χ^2 στατιστικού αυτές είναι ίσες. Επιπλέον ορίζεται ως το σταθμισμένο άθροισμα των χ^2 αποστάσεων των σημείων των προφίλ γραμμών και στηλών από τον μέσο:

$$I_r = \sum_i r_i (\mathbf{r}_i - \mathbf{c})' \mathbf{D}_c^{-1} (\mathbf{r}_i - \mathbf{c}), \text{ η ολική αδράνεια των γραμμών}$$

$$I_c = \sum_j c_j (\mathbf{c}_j - \mathbf{r})' \mathbf{D}_r^{-1} (\mathbf{c}_j - \mathbf{r}), \text{ η ολική αδράνεια των στηλών,}$$

όπου r_i είναι η μάζα της γραμμής i και c_j είναι η μάζα της στήλης j. Άρα η αδράνεια μπορεί να θεωρηθεί σαν ένα μέτρο διασποράς των σημείων των προφίλ από το κεντροειδές. Ακόμα, όσο μεγαλύτερη είναι η τιμή της, τόσο μεγαλύτερη είναι η μεταβλητότητα των σημείων από τον μέσο.

Η ανάλυση αντιστοιχιών χρησιμοποιεί το χ^2 μέτρο για τη δημιουργία του γεωμετρικού χώρου και τον υπολογισμό των αποστάσεων ενώ παράλληλα μπορεί να θεωρηθεί σαν μια τεχνική διάσπασης του χ^2 στατιστικού. Η διάσπαση επιτυγχάνεται μέσω της Γενικευμένης Singular Value Decomposition (SVD) του πίνακα των αποκλίσεων ή του πίνακα $\mathbf{P}$ με την οποία η ολική αδράνεια διασπάται σε έναν αριθμό ιδιοτιμών και ο αριθμός αυτός είναι ίσος με τον αριθμό των διαστάσεων.

5.2.6 Σύνδεση της SVD με την Ολική Αδράνεια - Συντεταγμένες των Προφίλ

Κάθε γραμμή και κάθε στήλη του αρχικού πίνακα συνάφειας αποτελεί ουσιαστικά ένα σημείο σε ένα πολυδιάστατο χώρο. Προσπαθούμε να αναπαραστήσουμε τα σημεία αυτά (προφίλ) σε ένα χώρο λιγότερων διαστάσεων με βέλτιστο τρόπο έτσι ώστε η αναπαράστασή τους να προσφέρει ουσιαστικότερη πληροφόρηση από την αναπαράσταση αυτών στον αρχικό

χώρο των περισσότερων διαστάσεων. Για να γίνει αυτό θα πρέπει να επιλέξουμε τον αριθμό των διαστάσεων που θα κρατήσουμε, να βρούμε τις συντεταγμένες των σημείων και μετά μέσα από ένα γράφημα σε χώρο λίγων διαστάσεων (2 ή 3 το πολύ) να μελετήσουμε τις σχέσεις μεταξύ των κατηγοριών των μεταβλητών, αλλά και τις συσχετίσεις μεταξύ των μεταβλητών εάν αυτές προκύψουν.

Για να αποκτήσουμε τις συντεταγμένες για τα προφίλ γραμμών και τα προφίλ στηλών σε σχέση με τους κύριους άξονες, χρησιμοποιούμε την γενικευμένη SVD για να προσεγγίσουμε τον πίνακα $\left(\mathbf{P}-\mathbf{rc}'\right)$ των αποκλίσεων. Για να το επιτύχουμε αυτό αρχικά, παραγοντοποιούμε τον πίνακα των τυποποιημένων καταλοίπων $\mathbf{S}=\mathbf{D}_r^{-1/2}(\mathbf{P}-\mathbf{rc}')\mathbf{D}_c^{-1/2}$, όπου έχει για στοιχεία του τα $\left(1/\sqrt{n}\right)\left(p_{ij}-r_ic_j\Big/\sqrt{r_ic_j}\right)$. Οπότε η SVD του $\mathbf{S}$ θα είναι:

$$\mathbf{S}=\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}' \tag{7}$$

Από την (7) παίρνουμε τις ιδιόμορφες τιμές $\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_k}$ του $\mathbf{S}$, τα αριστερά ιδιόμορφα διανύσματα και τα δεξιά ιδιόμορφα διανύσματα. Για την προσέγγιση του πίνακα $\mathbf{P}-\mathbf{rc}'$, η σχέση μεταξύ της γενικευμένης SVD και της SVD θα μας δώσει την διάσπαση του πίνακα και την σύνδεση που υπάρχει με την ολική αδράνεια. Από την (7) έχουμε ότι

$$\begin{aligned}\mathbf{D}_r^{-1/2}(\mathbf{P}-\mathbf{rc}')\mathbf{D}_c^{-1/2}=\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}' \Rightarrow\\ \mathbf{P}-\mathbf{rc}'=\mathbf{D}_r^{1/2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'\mathbf{D}_c^{1/2}\\ =\mathbf{A}\boldsymbol{\Lambda}\mathbf{B}'=\sum_{i=1}^{k}\sqrt{\lambda_i}\mathbf{a}_i\mathbf{b}_i'\end{aligned} \tag{8}$$

με τους περιορισμούς $\mathbf{A}'\mathbf{D}_r^{-1}\mathbf{A}=\mathbf{B}'\mathbf{D}_c^{-1}\mathbf{B}=\mathbf{I}$, αφού $\mathbf{U}'\mathbf{U}=\mathbf{V}'\mathbf{V}=\mathbf{I}$ στη σχέση (7) και από την (8), $\mathbf{A}=\mathbf{D}_r^{1/2}\mathbf{U}$ και $\mathbf{B}=\mathbf{D}_c^{1/2}\mathbf{V}$.

Τα $\sqrt{\lambda_i}$, $i=1,\ldots,k$, είναι τα στοιχεία του πίνακα $\boldsymbol{\Lambda}=\operatorname{diag}\left(\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_k}\right)$ και τα διανύσματα $\mathbf{a}_i,\mathbf{b}_i$, είναι οι στήλες των πινάκων $\mathbf{A}$, $\mathbf{B}$ αντίστοιχα. Η σχέση (8) μας λέει ότι για την πλήρη διάσπαση του πίνακα ο μέγιστος αριθμός των διαστάσεων (components) που μπορούμε να χρησιμοποιήσουμε είναι $k=\min\left(r-1,c-1\right)$ και είναι ίσος με την τάξη του πίνακα $\mathbf{P}-\mathbf{rc}'$. Στην πράξη όμως θα πάρουμε λιγότερες διαστάσεις με αποτέλεσμα να χρησιμοποιήσουμε τις πρώτες k^* ιδιόμορφες τιμές άρα η προσέγγιση του πίνακα $\mathbf{P}-\mathbf{rc}'$ θα γίνει από τον πίνακα

$$\left(\mathbf{P}-\mathbf{rc}^{'}\right)_{k^*}=\sum_{i=1}^{k^*}\sqrt{\lambda_i}\mathbf{a}_i\mathbf{b}_i^{'} \tag{9}$$

Στην πραγματικότητα η προσέγγιση από τον $\left(\mathbf{P}-\mathbf{rc}^{'}\right)_{k^*}$, με $k^*<k$, στην (9) ελαχιστοποιεί το

$$\mathrm{tr}\left[\mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)\mathbf{D}_c^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)^{'}\right] \tag{10}$$

Τότε από την (2) και την (10) το χ^2 στατιστικό μπορεί να πάρει την μορφή

$$X^2=n\mathrm{tr}\left[\mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)\mathbf{D}_c^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)^{'}\right]=n\sum_{i=1}^{k}\lambda_i \tag{11}$$

όπου $\lambda_1,\lambda_2,\ldots,\lambda_k$ είναι οι μη-μηδενικές ιδιοτιμές του πίνακα $\mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)\mathbf{D}_c^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)^{'}$ και $k=\mathrm{rank}[\mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)\mathbf{D}_c^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)^{'}]=\mathrm{rank}\left(\mathbf{P}-\mathbf{rc}^{'}\right)$. Με τον τρόπο που ορίστηκε παραπάνω η ολική αδράνεια βλέπουμε ότι ισούται με το άθροισμα των ιδιοτιμών:

$$I=\frac{X^2}{n}=\sum_{i=1}^{k}\lambda_i \tag{12}$$

Έτσι με την ανάλυση αντιστοιχιών επιτυγχάνεται η διαμέριση του χ^2 στατιστικό σε k συνιστώσες ή διαστάσεις που καθεμία εξηγεί και ένα μέρος του μέσω της σχέσης που υπάρχει μεταξύ του χ^2 και της αδράνειας.

Θα δείξουμε τώρα πως μπορούμε να γράψουμε εναλλακτικά την αδράνεια και να καταλήξουμε και πάλι στη σχέση (12). Από την (5), χρησιμοποιώντας ορισμένες από τις ιδιότητες των πινάκων και ότι $\mathbf{A}^{'}\mathbf{D}_r^{-1}\mathbf{A}=\mathbf{B}^{'}\mathbf{D}_c^{-1}\mathbf{B}=\mathbf{I}$, έχουμε ότι η αδράνεια I ισούται με

$$\begin{aligned}
I&=\mathrm{tr}\left[\mathbf{D}_r\left\{\mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)\right\}\mathbf{D}_c^{-1}\left\{\mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{rc}^{'}\right)^{'}\right\}\right]\\
&=\mathrm{tr}\left[\mathbf{D}_r\left\{\mathbf{D}_r^{-1}\left(\mathbf{A}\Lambda\mathbf{B}^{'}\right)\right\}\mathbf{D}_c^{-1}\left\{\mathbf{D}_r^{-1}\left(\mathbf{A}\Lambda\mathbf{B}^{'}\right)^{'}\right\}\right]\\
&=\mathrm{tr}\left[\mathbf{D}_r\left\{\mathbf{D}_r^{-1}\left(\mathbf{A}\Lambda\mathbf{B}^{'}\right)\right\}\mathbf{D}_c^{-1}\left\{\left(\mathbf{B}\Lambda\mathbf{A}^{'}\right)\left(\mathbf{D}_r^{'}\right)^{-1}\right\}\right]\\
&=\mathrm{tr}\left[\left(\mathbf{A}\Lambda\mathbf{B}^{'}\right)\mathbf{D}_c^{-1}\left(\mathbf{B}\Lambda\mathbf{A}^{'}\right)\mathbf{D}_r^{-1}\right]\\
&=\mathrm{tr}\left[\mathbf{A}\Lambda^2\mathbf{A}^{'}\mathbf{D}_r^{-1}\right]=\mathrm{tr}\left(\Lambda^2\underbrace{\mathbf{A}^{'}\mathbf{D}_r^{-1}\mathbf{A}}_{\mathbf{I}}\right)\Rightarrow\\
I&=\mathrm{tr}\left(\Lambda^2\right)
\end{aligned}$$

Οι απαιτούμενες διαστάσεις για να απεικονίσουμε ικανοποιητικά τα προφίλ των γραμμών και των στηλών μπορούν να αποφασιστούν εάν χρησιμοποιούμε την συνεισφορά στην ολική

αδράνεια των k^* διαστάσεων. Οι παραπάνω ιδιοτιμές μας δίνουν το σχετικό μέγεθος. Ως στόχο έχουμε να ερμηνεύσουμε τα δεδομένα σε λιγότερες διαστάσεις με όσο το δυνατόν μικρότερο χάσιμο πληροφορίας. Έτσι η ποσότητα $\sum_{i=1}^{k^*}\lambda_i \Big/ \sum_{i=1}^{k}\lambda_i$ μας δίνει το ποσοστό της αδράνειας που εξηγείται για τα δεδομένα για τον αριθμό των διαστάσεων που θα επιλέξουμε για την λύση.

Μπορεί να δειχτεί ότι οι τυποποιημένες κύριες συντεταγμένες για τη θέση των σημείων των προφίλ γραμμών και στηλών δίνονται από τις σχέσεις

$$\begin{aligned}\mathbf{r}_k &= \mathbf{D}_r^{-1}\mathbf{A}\boldsymbol{\Lambda} = \mathbf{D}_r^{-1}\left(\mathbf{P}-\mathbf{r}\mathbf{c}^{'}\right)\mathbf{D}_c^{-1}\mathbf{B}\\ \mathbf{s}_k &= \mathbf{D}_c^{-1}\mathbf{B}\boldsymbol{\Lambda} = \mathbf{D}_c^{-1}\left(\mathbf{P}-\mathbf{r}\mathbf{c}^{'}\right)^{'}\mathbf{D}_r^{-1}\mathbf{A}\end{aligned} \tag{13}$$

Για τη γραφική απεικόνιση των προφίλ στον διδιάστατο χώρο αρκεί να πάρουμε τις δύο πρώτες στήλες του πίνακα $\mathbf{r}_k$ και $\mathbf{s}_k$, αντίστοιχα. Επίσης είναι δυνατόν τα σημεία αυτά να συνδυαστούν σε ένα μόνο γράφημα αφού οι πίνακες $\mathbf{A}$ και $\mathbf{B}$ στην (8) μοιράζονται τις ίδιες ιδιόμορφες τιμές (*singular values*) $\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_k}$ του πίνακα $\boldsymbol{\Lambda}$ και έτσι μπορούμε να έχουμε τη δυνατότητα εντοπισμού συσχετισμό των γραμμών και των στηλών (**Σημ**. Για τον μετασχηματισμό αυτό βλ., Cox et al. 2001:190). Τέλος είναι πολύ σημαντικό να παρατηρήσουμε ότι οι Ευκλείδειες αποστάσεις μεταξύ των γραμμών του πίνακα $\mathbf{r}_k$ και των στηλών του πίνακα $\mathbf{s}_k$ είναι ίσες με τις χ^2 αποστάσεις μεταξύ των προφίλ του πίνακα $\mathbf{R}$ και των προφίλ του πίνακα $\mathbf{C}$.

5.2.7 Απόλυτες Συνεισφορές και Σχετικές Συνεισφορές

Η συνεισφορά των σημείων στις διαστάσεις (απόλυτες συνεισφορές), μεταφράζεται σαν το ποσοστό που συνεισφέρει στην αδράνεια για την διάσταση που μελετάμε ένα συγκεκριμένο σημείο και μαζί με το μέγεθος και το πρόσημο των συντεταγμένων του σημείου συμβάλλουν ιδιαίτερα στην φυσική ερμηνεία της συγκεκριμένης διάστασης. Τα σημεία εκείνα με την μεγαλύτερη συνεισφορά είναι περισσότερο σημαντικά για τον άξονα που μελετάμε. Η σχέση:

$$C_a(i,r_k) = \frac{\mathbf{r}r_{ki}^2}{\lambda_k}$$

για $i=1,2,\ldots,r,\ \ k=1,2,\ldots,K$, ονομάζεται απόλυτη συνεισφορά του $i-$σημείου στον r_k άξονα. Αντίστοιχα η σχέση:

$$C_a(j,s_k)=\frac{\mathbf{c}s_{kj}^2}{\lambda_k}$$

για $j=1,2,\ldots,c,\ \ k=1,2,\ldots,K$, ονομάζεται απόλυτη συνεισφορά του $j-$σημείου στον s_k άξονα. Θα πρέπει να σημειωθεί ότι το άθροισμα των απόλυτων συνεισφορών σε κάθε διάσταση ισούται με την μονάδα.

Το επόμενο βήμα για την ερμηνεία των αποτελεσμάτων της ανάλυσης αντιστοιχιών είναι να αποφασίσουμε για το πόσο καλά κάθε σημείο περιγράφεται από τους άξονες. Αυτό εκφράζεται από την συνεισφορά των αξόνων στα σημεία (σχετική συνεισφορά) και μας παρέχει την πληροφορία για το ποσό της αδράνειας (μεταβλητότητας) ενός συγκεκριμένου σημείου που ερμηνεύεται από τους άξονες. Το τετραγωνικό συνημίτονο ενός σημείου και ενός άξονα είναι το τετράγωνο του συνημίτονου της γωνίας που σχηματίζεται ανάμεσα στον άξονα και της ευθείας που ενώνει το κέντρο των αξόνων με το σημείο που αντιστοιχεί σε ένα προφίλ γραμμής ή στήλης. Αν συμβολίσουμε την γωνία αυτή με θ_{ik} τότε έχουμε ότι

$$\cos^2\theta_{ik}=r_{ki}^2/d_{ki}^2$$

όπου d_{ki} είναι η χ^2 απόσταση του σημείου i από το κεντροειδές. Τιμές κοντά στη μονάδα υπονοούν ότι το σημείο i είναι πολύ κοντά στον συγκεκριμένο άξονα αφού η γωνία θα είναι σχεδόν 0 και το $\cos^2 0^\circ=1$, άρα υπάρχει μεγαλύτερη συσχέτιση με την διάσταση αυτή, με αποτέλεσμα το συγκεκριμένο σημείο να είναι περισσότερο σημαντικό για τον άξονα αυτό. Σύμφωνα με τον Greenacre (1984), τα σημεία με μεγάλες απόλυτες συνεισφορές έχουν και μεγάλες σχετικές συνεισφορές όμως το αντίστροφο δεν ισχύει.

Η «Ποιότητα» της παρουσίασης για κάθε σημείο είναι το άθροισμα των σχετικών συνεισφορών και στις k-διαστάσεις που χρησιμοποιούνται στην ανάλυση. Εάν οι άξονες αντιπροσωπεύουν πλήρως τα δεδομένα η «Ποιότητα» τότε είναι 1. Σε οποιαδήποτε άλλη περίπτωση μόνο ένα ποσοστό αυτής εξηγείται από τους άξονες. Το στατιστικό αυτό είναι παρόμοιο της εταιρικότητας (*communality*) στην παραγοντική ανάλυση.

5.2.8 Σχέση με την Γενικευμένη SVD του Πίνακα Αντιστοιχιών P

Η παρουσίαση της μεθόδου στηρίχθηκε στην γενικευμένη διάσπαση ιδιόμορφων τιμών του πίνακα των αποκλίσεων $\left(\mathbf{P}-\mathbf{rc}'\right)$. Είναι όμως δυνατόν τα αποτελέσματα της ανάλυσης αντιστοιχιών να προκύψουν από την γενικευμένη SVD του πίνακα αντιστοιχιών **P** (Greenacre 1984, 1994; Cox, 2001; Beh, 2003, 2008). Στην περίπτωση αυτή η γενικευμένη

SVD του $\mathbf{P}$ δίνεται από την σχέση: $\mathbf{P}=\mathbf{A}^{*}\mathbf{\Lambda}^{*}\mathbf{B}^{*'}$, όπου $\mathbf{A}^{*'}\mathbf{D}_r^{-1}\mathbf{A}^{*}=\mathbf{B}^{*'}\mathbf{D}_c^{-1}\mathbf{B}^{*}=\mathbf{I}$ και $\mathbf{A}^{*}=[\mathbf{r},\mathbf{A}]$, $\mathbf{B}^{*}=[\mathbf{c},\mathbf{B}]$, $\mathbf{\Lambda}^{*}=\begin{bmatrix}1 & 0\\ 0 & \mathbf{\Lambda}\end{bmatrix}$. Μπορεί κανείς να δει ότι οι ιδιόμορφες τιμές του πίνακα $\mathbf{\Lambda}^{*}$ είναι οι τετραγωνικές ρίζες των μη-μηδενικών ιδιοτιμών του πίνακα $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1/2}$ ή του πίνακα $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}\mathbf{P}'$ (Cox et.al. 2001:188). Έτσι βλέπουμε ότι με τη γενικευμένη SVD του $\mathbf{P}$ ουσιαστικά προστίθεται μία ακόμη διάσταση που συνήθως ονομάζεται κοινότοπη (*trivial*) χωρίς όμως να αλλάζει τις υπόλοιπες ιδιόμορφες τιμές ή τις στήλες του πίνακα $\mathbf{A}$ και του $\mathbf{B}$, αντίστοιχα. Τελικά οι πίνακες των προφίλ που χρησιμοποιούνται στην ανάλυση αντιστοιχιών είναι οι: $\mathbf{D}_r^{-1}\mathbf{P}-\mathbf{1}\mathbf{c}^{\mathbf{T}}$ και $\mathbf{D}_c^{-1}\mathbf{P}^{\mathbf{T}}-\mathbf{1}\mathbf{r}^{\mathbf{T}}$, όπου τα διανύσματα $\mathbf{r}$ και $\mathbf{c}$ είναι οι μάζες. Επιπλέον οι συντεταγμένες των σημείων των προφίλ των γραμμών και των στηλών δίνονται από τις σχέσεις: $\mathbf{r}_k^{*}=\mathbf{D}_r^{-1}\mathbf{A}^{*}\mathbf{\Lambda}^{*}$ και $\mathbf{s}_k^{*}=\mathbf{D}_c^{-1}\mathbf{B}^{*}\mathbf{\Lambda}^{*}$.

5.3 Πολλαπλή Ανάλυση Αντιστοιχιών

Η ανάλυση αντιστοιχιών που περιγράφηκε στις προηγούμενες παραγράφους ανάλυε πίνακες συνάφειας δύο διαστάσεων. Για να αναλύσουμε πίνακες συνάφειας τρισδιάστατους ή μεγαλύτερης διάστασης χρησιμοποιούμε την μέθοδο της πολλαπλής ανάλυσης αντιστοιχιών.

Η μέθοδος αυτή είναι κατάλληλη για την αναπαράσταση των κατηγοριών των μεταβλητών από μεγάλα σετ δεδομένων σε ένα γράφημα λίγων διαστάσεων (συνήθως 2 ή 3 είναι αρκετές) για τον εντοπισμό και τη μελέτη των διαφοροποιήσεων ή συσχετισμών ανάμεσα στις κατηγορίες. Επίσης μπορεί να αποδώσει τη διάταξη των κατηγοριών με βέλτιστο τρόπο.

Είναι σημαντικό να σημειώσουμε ότι η πολλαπλή ανάλυση αντιστοιχιών αν και γενικεύει την ιδέα της ανάλυσης αντιστοιχιών στην ουσία οι δύο μέθοδοι δεν είναι ισοδύναμες. Όμως τα εργαλεία που χρησιμοποιούνται για τη μείωση των διαστάσεων είναι ανάλογα με εκείνα της ανάλυσης αντιστοιχιών που περιγράψαμε προηγουμένως. Θα πρέπει επίσης να τονίσουμε ότι η πολλαπλή ανάλυση αντιστοιχιών δουλεύει μόνο με τον πίνακα των αρχικών παρατηρήσεων $\mathbf{Y}(n\times p)$, έπειτα από κατάλληλα μετατροπή των κατηγορικών μεταβλητών. Άρα στον χάρτη που θα απεικονιστούν τα σημεία των στηλών μπορούμε να απεικονίσουμε και τις παρατηρήσεις κάτι που δεν μπορούσαμε να κάνουμε έως τώρα αφού δουλεύαμε με πίνακες συνάφειας όπου οι γραμμές και οι στήλες είναι οι κατηγορίες των μεταβλητών. Ο τρόπος που δουλεύει η μέθοδος περιγράφεται στα παρακάτω εδάφια.

5.3.1 Πίνακας Δείκτης και Πίνακας Burt

Στην τυπική μορφή ενός πίνακα πολυδιάστατων δεδομένων οι γραμμές αντιστοιχούν στις παρατηρήσεις και οι στήλες στις μεταβλητές. Η πολλαπλή ανάλυση αντιστοιχιών βασίζεται στην αναπαράσταση ενός τέτοιου πίνακα, με τη δημιουργία ενός πίνακα δείκτη (*indicator matrix*) $\mathbf{Z}$. Ο πίνακας $\mathbf{Z}$ αποτελείται από ψευδομεταβλητές για όλες τις κατηγορίες των μεταβλητών του πίνακα δεδομένων $\mathbf{Y}$. Κάθε ψευδομεταβλητή παίρνει την τιμή 1 αν η παρατήρηση ανήκει στην συγκεκριμένη κατηγορία της μεταβλητής ή 0 αν δεν ανήκει. Επιπλέον στον πίνακα $\mathbf{Z}$, υπάρχουν τόσες γραμμές όσες είναι και οι αρχικές παρατηρήσεις. Έτσι μια τυπική γραμμή του πίνακα δεδομένων με έξι μεταβλητές θα έχει την εξής μορφή για τον πίνακα $\mathbf{Z}$:

$$100:01000:10:010:0001:01000$$

Γενικά για έναν πίνακα p διαστάσεων όταν η j κατηγορική μεταβλητή αποτελείται από c_j κατηγορίες, μπορούμε να πάρουμε συνολικά $I = \sum_{j=1}^{p} c_j$ ψευδομεταβλητές για αυτή. Για κάθε παρατήρηση από τις n συνολικά μπορούμε να σχηματίσουμε νέο πίνακα με διαστάσεις $n \times I$, όπου οι ψευδομεταβλητές I να αντιστοιχούν στις στήλες. Τότε κάθε γραμμή του νέου πίνακα δείκτη θα αποτελείται από p στοιχεία με τιμή 1 και $I - p$ στοιχεία με τιμή 0. Έτσι ο αριθμός των 1 σε κάθε γραμμή θα είναι ίσος με τον αριθμό των μεταβλητών. Ο πίνακας δείκτης μπορεί να γραφτεί ως $\mathbf{Z} = \left[\mathbf{Z}_1, \ldots, \mathbf{Z}_p\right]$ με διαστάσεις $n \times c$ και μπορεί να αναλυθεί με την μέθοδο της ανάλυσης αντιστοιχιών όπως και κάθε άλλος πίνακας δύο διαστάσεων.

Κάποιες ιδιότητες του πίνακα είναι οι ακόλουθες (Καρλής, 2005):

- Η μάζα κάθε γραμμής είναι p/n,
- Η μάζα κάθε στήλης είναι το αντίστοιχο ποσοστό των περιθώριων συχνοτήτων,
- Οι χ^2 αποστάσεις μεταξύ των γραμμών είναι μια διαφοροποίηση του συντελεστή «matching coefficient». Ο «matching coefficient» μετρά τον αριθμό των διαφορετικών αποκρίσεων μεταξύ δύο παρατηρήσεων (γραμμών).

Από τον πίνακα $\mathbf{Z}$ μπορούμε να κατασκευάσουμε τον πίνακα $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$ ο οποίος ονομάζεται πίνακας Burt και να δουλέψουμε με αυτόν τον πίνακα εναλλακτικά.

Ο πίνακας $\mathbf{B}$ είναι ένας συμμετρικός πίνακας διάστασης $c \times c$, όπου c είναι το πλήθος όλων των διαφορετικών κατηγοριών των μεταβλητών και ο οποίος μπορεί να διασπαστεί σε p^2 υποπίνακες, με p ο αριθμός των κατηγορικών μεταβλητών. Στη διαγώνιο του βρίσκονται

οι υποπίνακες που είναι οι πίνακες συνάφειας κάθε μεταβλητής με τον εαυτό της. Εκτός της διαγωνίου είναι οι πίνακες συνάφειας δύο μεταβλητών. Τα στοιχεία του είναι οι συχνότητες για κάθε κελί του πίνακα συνάφειας. Το άθροισμα των διαγώνιων στοιχείων για κάθε υποπίνακα ξεχωριστά ισούται με το πλήθος των παρατηρήσεων. Η μορφή του πίνακα $\mathbf{B}$ είναι η εξής

$$\mathbf{B} = \begin{bmatrix} \mathbf{Z}_1'\mathbf{Z}_1 & \mathbf{Z}_1'\mathbf{Z}_2 & \dots & \mathbf{Z}_1'\mathbf{Z}_p \\ \mathbf{Z}_2'\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{Z}_2 & \dots & \mathbf{Z}_2'\mathbf{Z}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}_p'\mathbf{Z}_1 & \mathbf{Z}_p'\mathbf{Z}_2 & \dots & \mathbf{Z}_p'\mathbf{Z}_p \end{bmatrix}.$$

Λόγω της συμμετρίας του πίνακα $\mathbf{B}$, η ανάλυση καταλήγει σε μια απλή ανάλυση στηλών. Η SVD για τις γραμμές και τις στήλες του πίνακα $\mathbf{B}$ δίνει τα ίδια αποτελέσματα με την SVD για τις στήλες του πίνακα $\mathbf{Z}$. Η γεωμετρική ερμηνεία όμως που παίρνουμε όταν χρησιμοποιούμε τον πίνακα $\mathbf{B}$ είναι καλύτερη. Έτσι η πολλαπλή ανάλυση αντιστοιχιών ορίζεται ως η ανάλυση αντιστοιχιών του πίνακα Burt λόγω της σχέσης που υπάρχει μεταξύ της SVD του πίνακα δείκτη και του πίνακα Burt.

5.3.2 Προσαρμοσμένες Βασικές Αδράνειες

Καθώς η πολλαπλή ανάλυση αντιστοιχιών ισοδυναμεί με την ανάλυση του πίνακα Burt η αδράνεια που προκύπτει με τον τρόπο αυτό είναι υπερκτιμημένη καθώς οι διαγώνιοι υποπίνακες έχουν μέγιστη αδράνεια. Ενώ σε σχέση με τον πίνακα δείκτη οι συντεταγμένες των σημείων θα είναι ίδιες δεν συμβαίνει το ίδιο με τις ιδιοτιμές. Συγκεκριμένα οι ιδιοτιμές που προκύπτουν από τον πίνακα Burt είναι το τετράγωνο εκείνων που προκύπτουν από την ανάλυση του πίνακα δείκτη. Αυτό έχει σαν αποτέλεσμα της υπερκτίμησης της αδράνειας αφού οι ιδιοτιμές συνδέονται με την αδράνεια.

Για τον λόγο αυτό προσαρμόζουμε τις βασικές αδράνειες χρησιμοποιώντας την σχέση

$$\lambda_k^* = \left(\frac{p}{p-1}\right)\left(\lambda_k - \frac{1}{p}\right)^2$$

Οι λ_k^*, $k = 1, 2, \dots$ καλούνται προσαρμοσμένες αδράνειες και οι λ_k είναι οι *standard* ιδιοτιμές που προκύπτουν από την SVD του πίνακα Burt. Ενώ όμως ο αριθμός των διαστάσεων είναι ίσος με $\sum_{j=1}^{p}(c_j - 1)$, πρακτικά δεν έχουν όλες το ίδιο ενδιαφέρον. Σαν πρακτικός κανόνας

χρησιμοποιούνται εκείνες για τις οποίες ισχύει $\lambda_k > 1/p$ σύμφωνα με τον Greenacre και οι λ_k^* προσαρμόζονται σε αυτή την περίπτωση.

Εναλλακτικά μπορούμε να χρησιμοποιήσουμε την από κοινού ανάλυση αντιστοιχιών (*joint correspondence analysis*) στον πίνακα Burt η οποία προσαρμόζει μόνο τα μη-διαγώνια στοιχεία του.

5.4 Μοντέλα Συσχέτισης

5.4.1 Εισαγωγή

Όλα τα μοντέλα που θα παρουσιάσουμε στην ενότητα αυτή μπορούν να χρησιμοποιηθούν για τη μελέτη $I \times J$ πινάκων συνάφειας. Τα μοντέλα αυτά θα τα αναφέρουμε με τη γενική ονομασία «μοντέλα συσχέτισης» λόγω της αντιστοιχίας που υπάρχει μεταξύ των μοντέλων κανονικής συσχέτισης (canonical correlation models) και των μοντέλων ανάλυσης αντιστοιχιών (correspondence analysis models). Για τα μοντέλα συσχέτισης μπορούμε να πούμε πως απλοποιούν και συμπληρώνουν τα αποτελέσματα που παίρνουμε από την ανάλυση αντιστοιχιών και την canonical correlation ανάλυση.

Τα μοντέλα αυτά μπορούν να εφαρμοστούν σε πίνακες ταξινόμησης όταν οι κατηγορίες της μεταβλητής γραμμής και οι κατηγορίες της μεταβλητής στήλης έχουν καθορισμένη διάταξη, όπως επίσης και όταν δεν υπάρχει καμία διάταξη στις κατηγορίες αυτές (μεταβλητές σε ονομαστική κλίμακα) ή όταν η διάταξη των κατηγοριών δεν είναι γνωστή *a priori*. Επιπλέον μπορούν να χρησιμοποιηθούν όταν η μία μεταβλητή είναι ονομαστική και η άλλη είναι διατάξιμη καθώς επίσης και στις περιπτώσεις που η απόσταση (spacing) μεταξύ των κατηγοριών των μεταβλητών θεωρείται γνωστή, ή ακόμα και όταν θέσουμε περιορισμούς ομοιογένειας στα σκορ, αλλά και όταν χρησιμοποιούμε προκαθορισμένα σκορ. Έτσι λοιπόν φαίνεται ότι τα μοντέλα συσχέτισης έχουν πολλά κοινά στοιχεία με τα μοντέλα συνάφειας για διδιάστατους πίνακες που παρουσιάσαμε στο 4^{ο} Κεφάλαιο, ενώ υπάρχουν περιπτώσεις που τα αποτελέσματα που δίνει η μία ανάλυση να μοιάζουν με τα αποτελέσματα της άλλης (βλ. Goodman, 1981, 1991).

Στο υπόλοιπο τμήμα του κεφαλαίου θα αναφερθούμε στην γενικότερη περίπτωση όπου τα σκορ και η απόσταση μεταξύ των κατηγοριών των μεταβλητών ταξινόμησης του πίνακα συνάφειας είναι άγνωστα και πρέπει να εκτιμηθούν. Αναλυτικότερη και εκτενέστερη συζήτηση πάνω στα θέματα αυτά μπορεί να βρεθεί στους Gilula and Haberman (1986, 1988),

Goodman (1985, 1986, 1991), Greenacre (1984, 1994), van der Heijden, de Falguerolles and de Leeuw (1985, 1989), van der Heijden and Worsley (1988), Wasserman and Faust (1989, 1993), Gilula and Ritov (1990) και Beh (1997, 1998, 1999, 2008).

5.4.2 Μοντέλα Κανονικής Συσχέτισης και Μοντέλα Ανάλυσης Αντιστοιχιών

5.4.2.1 Canonical correlation ανάλυση

Πριν προχωρήσουμε στη διατύπωση των μοντέλων θα κάνουμε μια σύντομη αναφορά στη μέθοδο της canonical correlation analysis. Η ανάλυση αντιστοιχιών μπορεί να θεωρηθεί σαν μια ειδική μέθοδος της canonical correlation analysis και τα αποτελέσματα που παίρνουμε από την SVD είναι ίδια. Εν συντομία αναφέρουμε ότι, γενικά με την canonical correlation ανάλυση ο σκοπός είναι να βρούμε ταυτόχρονες γραμμικές σχέσεις ανάμεσα σε 2 σετ μεταβλητών $\mathbf{X}(n\times q)$ και $\mathbf{Y}(n\times s)$ που μπορούν να δειχθούν ως $\mathbf{w}=\mathbf{Xb}$ και $\mathbf{z}=\mathbf{Ya}$ έτσι ώστε ο συντελεστής κανονικής συσχέτισης ρ_{zw} ανάμεσα στο $\mathbf{z}$ και $\mathbf{w}$ να μεγιστοποιείται. Χωρίς να αναφερθούμε σε λεπτομέρειες, λέμε πως η διαδικασία είναι ένα πρόβλημα εύρεσης των ιδιοδιανυσμάτων $\mathbf{a}_j$, $\mathbf{b}_j$ από όπου προκύπτει ότι η συσχέτιση ανάμεσα στις κανονικές μεταβλητές $\mathbf{b}_j'\mathbf{x}$ και $\mathbf{a}_j'\mathbf{y}$ είναι ίση με την τετραγωνική ρίζα των ιδιοτιμών λ_j δηλαδή $\sqrt{\lambda_j}$, για $j=1,2,\ldots,t$ και $t=\min(s,q)$ είναι η τάξη των δύο σετ μεταβλητών. Όταν έχουμε πίνακα συνάφειας 2 διαστάσεων τότε οι κανονικές συσχετίσεις του πίνακα $\mathbf{Z}(n\times(r+c))=(\mathbf{Z}_1,\mathbf{Z}_2)$ που αποτελείται από ψευδομεταβλητές για τις κατηγορίες του, δίνονται από τις ιδιόμορφες τιμές της διάσπασης του. Επίσης ισχύει και τώρα ότι η μεγαλύτερη κανονική συσχέτιση δίνεται από την πρώτη ιδιόμορφη τιμή $\sqrt{\lambda_1}$ της διάσπασης, ενώ με τη μέθοδο αυτή μπορούμε επίσης να υπολογίσουμε τα σταθμισμένα σκορ των κατηγοριών των γραμμών και των στηλών από τις ποσότητες $\sqrt{\lambda}\mathbf{c}$ και $\sqrt{\lambda}\mathbf{d}$ αντίστοιχα. Οι κανονικές μεταβλητές τώρα δίνονται από τις $\mathbf{Z}_2\mathbf{d}$ και $\mathbf{Z}_1\mathbf{c}$. Τα ιδιοδιανύσματα $\mathbf{c},\mathbf{d}$ είναι οι τυποποιημένοι κανονικοί συντελεστές ή τα κανονικά βάρη και αντιστοιχούν στις κύριες συντεταγμένες (principal coordinates) για τις γραμμές και τις στήλες.

5.4.2.2 Μοντέλα Συσχέτισης

Προχωράμε τώρα με τη διατύπωση των μοντέλων όπου στο τέλος θα καταλήξουμε να έχουμε ένα σετ με τα σκορ γραμμής $\{x_{im}\}$, ένα σετ με τα σκορ στήλης $\{y_{jm}\}$ και ένα σετ με τις κύριες αδράνειες $\{\lambda_m\}$ ή τα τετράγωνα των ιδιόμορφων τιμών $\{\sqrt{\lambda_m}\}$.

Για έναν $I \times J$ πίνακα συνάφειας έστω P_{ij} είναι η πιθανότητα μια παρατήρηση να ανήκει στο (i, j) κελί. Η ανεξαρτησία μεταξύ των γραμμών και των στηλών μπορεί να δειχθεί ως

$$P_{ij} = P_{i\bullet}P_{\bullet j} \tag{14}$$

όπου $P_{i\bullet} = \sum_j P_{ij}$ και $P_{\bullet j} = \sum_i P_{ij}$ είναι οι κατανομές περιθωρίου. Ορίζουμε ως ένα συνολικό μέτρο για την μη-ανεξαρτησία το:

$$\Delta = \left[\sum_i \sum_j c_{ij}^2 P_{i\bullet}P_{\bullet j}\right]^{1/2} = \sum_i \sum_j P_{ij} - P_{i\bullet}P_{\bullet j} \Big/ \left(P_{i\bullet}P_{\bullet j}\right)^{1/2} \tag{15}$$

όπου στην (15) είναι: $c_{ij} = \left(P_{ij} - P_{i\bullet}P_{\bullet j}\right) \Big/ P_{i\bullet}P_{\bullet j}$ (16)

Από την (16) προκύπτει η σχέση

$$D_{ij} \propto c_{ij} + 1 = P_{ij} \Big/ P_{i\bullet}P_{\bullet j} \tag{17}$$

όπου το D_{ij} ονομάζεται Pearson's ratio σύμφωνα με τον Goodman (1996, 2002) και ορίζει ένα μέτρο για την απόκλιση από την ανεξαρτησία για την i γραμμή και την j στήλη. Με βάση τα όσα έχουν αναφερθεί σε προηγούμενες παραγράφους, η (17) μπορεί να γραφτεί στη πινακική μορφή

$$\mathbf{D} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}$$

και για να αποφανθούμε για την ύπαρξη συνάφειας μεταξύ των γραμμών και των στηλών χρησιμοποιείται η SVD του πίνακα του Pearson's ratio έτσι ώστε $\mathbf{D} - \mathbf{U} = \mathbf{A}\mathbf{\Lambda}\mathbf{B}'$, όπου $\mathbf{U}$ είναι ο πίνακας που έχει για στοιχεία του μονάδες. Εάν συμβαίνει $\mathbf{D} = \mathbf{U}$ τότε οδηγούμαστε στην ανεξαρτησία γραμμών και στηλών. Όλα αυτά που αναφέραμε συνοπτικά μας οδηγούν στα μοντέλα συσχέτισης.

Το μοντέλο (14) μπορεί να διατυπωθεί στην πιο γενική του μορφή ως εξής:

$$P_{ij} = P_{i\bullet}P_{\bullet j}\left(1 + \sum_{m=1}^{M} \sqrt{\lambda_m} x_{im} y_{jm}\right) \tag{18}$$

όπου $M = \min(I-1, J-1)$. Οι κανονικές μεταβλητές (canonical variables) $\mathbf{x}_m$ και $\mathbf{y}_m$ ικανοποιούν παρόμοιες συνθήκες (περιορισμούς) με εκείνες των μοντέλων συνάφειας:

$$\sum_{i=1}^{I} x_{im} P_{i\bullet} = 0, \qquad \sum_{j=1}^{J} y_{jm} P_{\bullet j} = 0$$

$$\sum_{i=1}^{I} x_{im}^2 P_{i\bullet} = 1, \qquad \sum_{j=1}^{J} y_{jm}^2 P_{\bullet j} = 1$$

$$\sum_{i=1}^{I} x_{im} x_{im'} P_{i\bullet} = 0, \sum_{j=1}^{J} y_{jm} y_{jm'} P_{\bullet j} = 0 \qquad (19)$$

για $m \neq m'$. Με την παραπάνω κλιμακοποίηση τα $\mathbf{x}$ και $\mathbf{y}$ ονομάζονται κανονικές συντεταγμένες (standard coordinates) κατά τον Greenacre στα πλαίσια της ανάλυσης αντιστοιχιών. Επίσης από τον τρίτο περιορισμό στην (19) παρατηρούμε ότι για $m \neq m'$, τα x_{im} και $x_{im'}$ είναι ασυσχέτιστα όπως επίσης είναι και τα y_{jm} και $y_{jm'}$. Οι κανονικές μεταβλητές $\mathbf{x}_m$ και $\mathbf{y}_m$ είναι τα τυποποιημένα σκορ των κατηγοριών που θα εκτιμηθούν από τα δεδομένα για την m συνιστώσα ή διάσταση.

Για την m συνιστώσα, $m = 1, \ldots, M$, η συσχέτιση μεταξύ των x_{im} και y_{jm} ισούται με συντελεστή κανονικής συσχέτισης ρ_{XY}, δηλαδή το $\sqrt{\lambda_m}$. Η παράμετρος $\sqrt{\lambda_m}$ μετράει την συσχέτιση των γραμμών και των στηλών αφού ισχύει ότι

$$\sum_{i,j} x_{im} y_{jm} P_{ij} = \sqrt{\lambda_m} \qquad (20)$$

Συνεπώς οι παράμετροι $\mathbf{x}_m$ και $\mathbf{y}_m$ είναι τα σκορ των γραμμών και των στηλών οι οποίοι μεγιστοποιούν τη συσχέτιση $\sqrt{\lambda_m}$ στη σχέση (20) βάσει των περιορισμών που τίθενται σε αυτά στην (19): οι παράμετροι x_{i1} και y_{j1} είναι τα σκορ των κατηγοριών των μεταβλητών τα οποία μεγιστοποιούν την κανονική συσχέτιση $\sqrt{\lambda_1}$, παρόμοια οι παράμετροι x_{i2} και y_{j2} είναι τα σκορ που μεγιστοποιούν την κανονική συσχέτιση $\sqrt{\lambda_2}$, με τον περιορισμό ότι τα x_{i1} και x_{i2} όπως και τα y_{j2} και y_{j1} είναι ασυσχέτιστα μεταξύ τους κ.ο.κ.. Η σχέση (20) προκύπτει από το γεγονός ότι: $\Delta^2 = \sum_i \sum_j c_{ij}^2 P_{i\bullet} P_{\bullet j} = \sum_i \sum_j \left(P_{ij} - P_{i\bullet} P_{\bullet j}\right)^2 \Big/ \left(P_{i\bullet} P_{\bullet j}\right) = \sum_m \rho_m^2$, όπου $\rho_m = \sqrt{\lambda_m}$ ο γνωστός συντελεστής συσχέτισης και μέσω των δύο πρώτων περιορισμών της σχέσης (19).

Το μοντέλο στη σχέση (18) ονομάζεται κορεσμένο RC μοντέλο κανονικής συσχέτισης συμφώνα με τον Goodman (1986) και για τον έλεγχο προσαρμογής έχει 0 βαθμούς ελευθερίας αφού περιγράφει πλήρως τα δεδομένα. Συνεπώς για να μπορέσουμε να μελετήσουμε τα δεδομένα χρειαζόμαστε οικονομικότερα μοντέλα και στην περίπτωση αυτή αναφερόμαστε στον αριθμό των συνιστωσών που θα επιλέξουμε για να διασπάσουμε το $\mathbf{D}-\mathbf{U}$. Έτσι, εάν στο μοντέλο αυτό αντικαταστήσουμε το M με το $M^* < \min(I-1, J-1)$ παίρνουμε μη-κορεσμένα μοντέλα που μπορούμε να μελετήσουμε. Τότε ο έλεγχος καλής προσαρμογής για τα μοντέλα αυτά μπορεί να γίνει αφού θα έχουμε $(I-M^*-1)(J-M^*-1)$ βαθμούς ελευθερίας. Δηλαδή θα πάρουμε την προσέγγιση

$$D_{ij}-1=\sum_{m=1}^{M^*}\sqrt{\lambda_m}x_{im}y_{jm} \tag{21}$$

μέσω της SVD, που όπως είπαμε νωρίτερα μας δείχνει το μέγεθος της απόκλισης από την ανεξαρτησία.

Επίσης πριν είδαμε ότι

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(P_{ij}-P_{i\cdot}P_{\cdot j}\right)^2 \Big/ P_{i\cdot}P_{\cdot j}=\sum_{i=1}^{M}\lambda_m \tag{22}$$

Δηλαδή το χ^2 τεστ του Pearson διαμερίζεται σε M όρους λ_m, ο καθένας από τους οποίους εκφράζει το ποσοστό του τεστ που ερμηνεύεται από την αντίστοιχη διάσταση. Το αποτέλεσμα αυτό είναι ίδιο με εκείνο που είχαμε βρει στην ανάλυση αντιστοιχιών στη σχέση (11) όταν διασπάσαμε την αδράνεια σε k ιδιοτιμές.

Οι Goodman και Agresti σημειώνουν ότι όταν η τιμή του $\sqrt{\lambda}$ είναι κοντά στο μηδέν, η τιμή για το ϕ και οι τιμές των σκορ των κατηγοριών των μεταβλητών που παίρνουμε από το RC μοντέλο συνάφειας είναι περίπου ίδιες με εκείνες που παίρνουμε από το μοντέλο κανονικής συσχέτισης. Όμως τα αποτελέσματα μπορεί να διαφέρουν κατά πολύ όταν η τιμή του $\sqrt{\lambda}$ δεν είναι κοντά στο μηδέν.

Η αντιστοιχία ανάμεσα στα μοντέλα κανονικής συσχέτισης και στα μοντέλα ανάλυσης αντιστοιχιών βασίζεται στο γεγονός ότι για το μοντέλο ανάλυσης αντιστοιχιών έχουμε κλιμακοποιήσει με διαφορετικό τρόπο τα σκορ. Ο Goodman (1986) διατυπώνει το μοντέλο αυτό ως RC μοντέλο ανάλυσης αντιστοιχιών και η γενική του (κορεσμένη) μορφή είναι

$$P_{ij}=P_{i\cdot}P_{\cdot j}\left(1+\sum_{m=1}^{M}x'_{im}y'_{jm}\Big/\sqrt{\lambda_m}\right) \tag{23}$$

όπου $x_{im}^{'} = \sqrt{\lambda_m} x_{im}$ και $y_{jm}^{'} = \sqrt{\lambda_m} y_{jm}$. (24)

Τα $x_{im}^{'}$ και $y_{jm}^{'}$ στην σχέση (24) είναι τα σκορ των κατηγοριών των μεταβλητών που χρησιμοποιούνται από το μοντέλο της ανάλυσης αντιστοιχιών και μας δείχνουν την αντιστοιχία που υπάρχει ανάμεσα στα δύο μοντέλα.

Από τις σχέσεις (19) και (24) βρίσκουμε ότι

$$\sum_{i=1}^{I} x_{im}^{'} P_{i\bullet} = 0, \quad \sum_{j=1}^{J} y_{jm}^{'} P_{\bullet j} = 0 , \quad \sum_{i=1}^{I} x_{im}^{'2} P_{i\bullet} = \lambda_m, \quad \sum_{j=1}^{J} y_{jm}^{'2} P_{\bullet j} = \lambda_m \tag{25}$$

Με τον τρόπο αυτό κλιμακοποίησης των σκορ στην σχέση (25), τα σκορ ονομάζονται κύριες συντεταγμένες (principal coordinates) κατά τον Greenacre, ενώ όπως είπαμε παραπάνω από την σχέση (24) μπορούμε να δούμε την αντιστοιχία που υπάρχει μεταξύ των κύριων συντεταγμένων και των κανονικών συντεταγμένων. Με τον τρόπο αυτό κλιμακοποίησης των συντεταγμένων ή σκορ, το σταθμισμένο άθροισμα των τετραγώνων των συντεταγμένων είναι ίσο με την κύρια αδράνεια ή ιδιοτιμή για την m διάσταση, ενώ για τις κανονικές συντεταγμένες αυτό είναι 1 (Greenacre, 1994). Επιπλέον από τις (20) και (24) βλέπουμε ότι η παράμετρος $\sqrt{\lambda_m}$ για το μοντέλο (23) είναι ένα μέτρο συσχέτισης μεταξύ των σκορ $x_{im}^{'}$ και $y_{jm}^{'}$.

Για τα x_{im}, y_{jm}, $x_{im}^{'}$ και $y_{jm}^{'}$ από τις σχέσεις (18), (19) και (24) μπορούμε πολύ εύκολα να δούμε ότι τα $y_{jm}^{'}$ είναι ο σταθμισμένος μέσος των x_{im} χρησιμοποιώντας την δεσμευμένη κατανομή $P_{i/j}$ σαν σταθμά καθώς και ότι τα $x_{im}^{'}$ είναι ο σταθμισμένος μέσος των y_{jm} χρησιμοποιώντας την δεσμευμένη κατανομή $P_{j/i}$ σαν σταθμά:

$$\sum_{i=1}^{I} x_{im} P_{ij} \big/ P_{\bullet j} = \sqrt{\lambda_m} y_{im} = y_{im}^{'} \quad \text{και} \quad \sum_{j=1}^{J} y_{im} P_{ij} \big/ P_{i\bullet} = \sqrt{\lambda_m} x_{im} = x_{im}^{'} \tag{26}$$

Έτσι από την (26) αποκτούμε μεγαλύτερη κατανόηση για την ερμηνεία των σκορ. Για τη μελέτη του μοντέλου (23) χρησιμοποιούμε έναν μικρότερο αριθμό κανονικών συσχετίσεων $\sqrt{\lambda_m}$ έτσι πάμε σε οικονομικότερα μοντέλα.

5.5 Εφαρμογή 1η

Στο 4ο Κεφάλαιο χρησιμοποιήσαμε για την ανάλυση του Πίνακα 2 τριπλής εισόδου λογαριθμογραμμικά μοντέλα και μοντέλα συνάφειας, κάναμε ελέγχους για τη σημαντικότητα των αλληλεπιδράσεων και ερμηνεύσαμε τις παραμέτρους του $RC(1)+RL(1)+CL(1)$ μοντέλου. Στην εφαρμογή αυτή αναλύουμε ξανά τον Πίνακα 2 χρησιμοποιώντας την ανάλυση αντιστοιχιών και τα μοντέλα συσχέτισης αφού όμως πρώτα γίνουν ορισμένες απαραίτητες διευκρινήσεις.

Τα λογαριθμογραμμικά μοντέλα και η ανάλυση αντιστοιχιών μας παρέχουν δύο διαφορετικούς τρόπους για τη μελέτη πινάκων συνάφειας. Ιδιαίτερα, η ανάλυση αντιστοιχιών είναι μια τεχνική κατάλληλη για διδιάστατους πίνακες συνάφειας κυρίως λόγω της έμφασης της στη γεωμετρική απεικόνιση, ενώ τα λογαριθμογραμμικά μοντέλα χρησιμοποιούνται κατά κόρον για την ανάλυση πινάκων συνάφειας μεγαλύτερης διάστασης. Για να αναλύσουμε έναν τρισδιάστατο πίνακα συνάφειας, τότε, κατασκευάζουμε έναν νέο πίνακα συνάφειας διπλής εισόδου ενώνοντας δύο από τις μεταβλητές σε μια νέα μεταβλητή που έχει για κατηγορίες κάθε συνδυασμό των κατηγοριών τους και φτιάχνουμε με τον τρόπο αυτό μια μεταβλητή αλληλεπίδρασης (*interactive variable*). Στην Γαλλική σχολή οι πίνακες αυτοί ονομάζονται πολλαπλοί (*multiple*) πίνακες και χρησιμοποιούνται συχνά για την ανάλυση πινάκων συνάφειας μεγάλης διάστασης. Όπως οι van der Heijden and de Leeuw (1985) αναφέρουν, η ανάλυση αντιστοιχιών στους πίνακες αυτούς θεωρείται σαν μια μέθοδο για τη διάσπαση της διαφοράς του χ^2 στατιστικού μεταξύ δύο λογαριθμογραμμικών μοντέλων. Επιπλέον είναι σημαντικό να αποφασίσουμε ποιες από τις μεταβλητές θα ενώσουμε καθώς η αλληλεπίδραση τους δεν θα επηρεάσει το αποτέλεσμα της ανάλυσης. Έτσι ενδέχεται να υπάρξει απώλεια πληροφορίας και για τον λόγο αυτόν είθισται να ενώνουμε τις μεταβλητές των οποίων οι αλληλεπιδράσεις έχουν το λιγότερο ενδιαφέρον για την ανάλυση ή δεν είναι στατιστικά σημαντικές, δηλαδή όταν απορρίπτονται από τον έλεγχο. Στην περίπτωση μελέτης πίνακα συνάφειας μεγαλύτερης διάστασης του τρία, τότε υπάρχει μια ποικιλία μοντέλων μερικής συνάφειας από τα οποία επιλέγουμε διάφορους συνδυασμούς μεταξύ των μεταβλητών ώστε να σχηματιστεί ο τελικός πίνακα δύο διαστάσεων.

Από τον Πίνακα 2 μπορούν να προκύψουν τρεις πολλαπλοί πίνακες: οι $(H\times W)\times S$, $(H\times S)\times W$ και $(S\times W)\times H$ με την μεταβλητή αλληλεπίδρασης να εμφανίζεται κάθε φορά

μέσα στις παρενθέσεις. Στην περίπτωση μας έχει περισσότερο νόημα να μεταχειριστούμε την μεταβλητή Συνεισφορά στις Εργασίες του Σπιτιού [S] σαν εξαρτημένη και να μελετήσουμε τη σχέση της μεταβλητής αυτής με την Ιδεολογία των αντρών [H] και την Ιδεολογία των γυναικών [W]. Αυτό σημαίνει τη δημιουργία του 9×3 πολλαπλού πίνακα $(H\times W)\times S$, όπου οι [H] και [W] μαζί δημιουργούν την *interactive* μεταβλητή. Θα πρέπει να σημειώσουμε ότι τώρα η αλληλεπίδραση [HW] δεν θεωρείται σημαντική και δεν επηρεάζει τα αποτελέσματα ενώ η ανάλυση αντιστοιχιών στον πίνακα αυτόν διασπά τη διαφορά του χ^2 στατιστικού μεταξύ των λογαριθμογραμμικών μοντέλων [HWS] (κορεσμένου) και [S][HW] (με ένα όρο αλληλεπίδρασης $2^{ης}$ τάξης).

Πίνακας 5.1 : *Ο πολλαπλός πίνακας* $(H\times W)\times S$ *της ανάλυσης αντιστοιχιών.*

	Share of Housework		
Husband Ideology & Wife Ideology	Rarely	Sometimes	Often
Traditional Husband / Traditional Wife	73	137	43
Traditional Husband / Moderate Wife	38	89	32
Traditional Husband / Liberal Wife	13	59	35
Moderate Husband / Traditional Wife	21	69	28
Moderate Husband / Moderate Wife	36	121	49
Moderate Husband / Liberal Wife	17	103	72
Liberal Husband / Traditional Wife	8	21	6
Liberal Husband / Moderate Wife	8	56	47
Liberal Husband / Liberal Wife	17	95	106

Στον Πίνακα 5.1 το χ^2 τεστ μας δείχνει ότι υπάρχει ισχυρή ένδειξη συσχέτισης μεταξύ της *interactive* μεταβλητής $(H\times W)$ και της [S] ($X^2 = 120.398$, df = 16, p-value<$2.2e$-16). Η ανάλυση αντιστοιχιών διασπά την τιμή του χ^2 τεστ σε k ή M διαστάσεις/συνιστώσες. Για τον Πίνακα 5.1 ο μέγιστος αριθμός των διαστάσεων που μπορούμε να χρησιμοποιήσουμε είναι: $\min(9-1,3-1)=2$. Όπως φαίνεται από τον Πίνακα 5.2, δύο διαστάσεις εξηγούν ολόκληρο το ποσοστό της αδράνειας ή της τιμής του χ^2 στατιστικού.

Πίνακας 5.2 : *Ιδιοτιμές και ποσοστό αδράνειας που εξηγείται.*

	Ιδιοτιμές	Ποσοστό αδράνειας	Αθροιστικό ποσοστό αδράνειας
Dim1	0.080936	94.045	94.045
Dim2	0.005124	5.955	100.000
Άθροισμα: 0.08606, $\chi^2 = (0.08606)\times1399 = 120.398$ με df = 16			

Βλέπουμε ότι η 1^{η} διάσταση εξηγεί το 94.05% της αδράνειας και είναι δεσπόζουσα, ενώ η 2^{η} διάσταση εξηγεί το υπόλοιπο 5.95%. Οι ιδιοτιμές είναι 0.080936 και 0.005124, ενώ οι ιδιόμορφες τιμές είναι ίσες με 0.2845 και 0.0715 αντίστοιχα και μπορούν να ερμηνευτούν σαν μέγιστες κανονικές συσχετίσεις μεταξύ της *interactive* μεταβλητής $(H \times W)$ και της εξαρτημένης μεταβλητής Συνεισφορά στις Εργασίες του Σπιτιού, [S]. Στον παραπάνω πίνακα αναγράφεται και η τιμή του χ^2 τεστ ανεξαρτησίας που υπολογίσαμε στην αρχή και η οποία βρίσκεται πολλαπλασιάζοντας την ολική αδράνεια με το μέγεθος του δείγματος.

Στον Πίνακα 5.3 παραθέτουμε τα περιγραφικά στατιστικά των γραμμών και των στηλών μετά την ανάλυση αντιστοιχιών. Σε γενικές γραμμές τα σημεία των γραμμών και των στηλών που έχουν μεγάλες μάζες και υψηλές συντεταγμένες για την k – διάσταση θα συνεισφέρουν περισσότερο στην αδράνεια που ερμηνεύεται από την διάσταση αυτή. Επίσης θα πρέπει να ελέγχεται η συνεισφορά των σημείων στην αδράνεια ξεχωριστά για τις γραμμές και για τις στήλες στην κάθε διάσταση, όπως και οι σχετικές συνεισφορές που δείχνουν το πόσο καλά κάθε σημείο περιγράφεται από την κάθε διάσταση.

Πίνακας 5.3 : *Στατιστικά στοιχεία της ανάλυσης για τις γραμμές και τις στήλες του Πίνακα 5.1.*

		Συντεταγμένες		Συνεισφορά στην αδράνεια		Σχετικές Συνεισφορές		Ποιότητα
Κατηγορία	Μάζα	Dim1	Dim2	Dim1	Dim2	Dim1	Dim2	Dim1+Dim2
Rarely (**V1**)	0.165	0.46	0.11	44.9	38.6	0.95	0.05	1.00
Sometimes (**V2**)	0.536	0.06	-0.06	2.82	43.5	0.51	0.49	1.00
Often (**V3**)	0.298	-0.37	0.05	52.3	17.8	0.98	0.02	1.00
Trad.H/Trad.W	0.181	0.37	0.08	31.4	25.1	0.95	0.05	1.00
Trad.H/Mod.W	0.113	0.25	0.02	9.22	0.59	0.99	0.01	1.00
Trad.H/Lib.W	0.076	-0.10	-0.05	1.06	5.13	0.77	0.23	1.00
Mod.H/Trad.W	0.084	0.11	-0.07	1.34	8.46	0.71	0.29	1.00
Mod.H/Mod.W	0.147	0.10	-0.08	2.13	17.7	0.65	0.35	1.00
Mod.H/Lib.W	0.137	-0.22	-0.06	8.74	9.18	0.94	0.06	1.00
Lib.H/Trad.W	0.025	0.28	-0.06	2.56	1.69	0.96	0.04	1.00
Lib.H/Mod.W	0.079	-0.32	-0.01	10.4	0.47	0.99	0.01	1.00
Lib.H/Lib.W	0.156	-0.41	0.10	33.1	31.6	0.94	0.06	1.00

Ξεκινώντας την ερμηνεία για τις στήλες, για την 1^{η} διάσταση παρατηρούμε ότι οι κατηγορίες «Rarely» και «Often» εξηγούν περίπου ίδιο ποσοστό της αδράνειας. Αυτές οι δύο κατηγορίες έχουν μεγάλες συντεταγμένες με αντίθετα πρόσημα και σαν αποτέλεσμα θα έχουν αντίθετη κατεύθυνση στο χάρτη. Η κατηγορία «Sometimes» έχει καταρχάς, την μεγαλύτερη μάζα άρα το προφίλ της είναι ιδιαίτερα διαφορετικό από τα άλλα 2 προφίλ των στηλών και

επιπλέον συνεισφέρει περισσότερο στη $2^{η}$ διάσταση. Οπότε μπορούμε να υποθέσουμε ότι η Συνεισφορά στις Εργασίες του Σπιτιού περνάει από το «Often» στο «Rarely» ή αντίστροφα μέσω του «Sometimes» στην $1^{η}$ διάσταση, ενώ η $2^{η}$ διάσταση ξεχωρίζει εκείνους που έχουν μια πιο ενδιάμεση αντιμετώπιση από εκείνους που είναι στο ένα άκρο ή στο άλλο. Για τα σημεία των γραμμών της *interactive* μεταβλητής, στη $1^{η}$ διάσταση παρατηρούμε ότι οι κατηγορίες «Traditional Husband / Traditional Wife» και «Liberal Husband / Liberal Wife» εξηγούν το ίδιο ποσοστό της αδράνειας και έχουν συντεταγμένες με αντίθετα πρόσημα. Για τη $2^{η}$ διάσταση βλέπουμε πως και η κατηγορία «Moderate Husband / Moderate Wife» μαζί με τις δύο προαναφερθείσες κατηγορίες συνεισφέρει περισσότερο στην αδράνεια και έχει συντεταγμένη με αντίθετο πρόσημο. Άρα μπορούμε να υποθέσουμε ότι τα ζευγάρια με κοινή Ιδεολογία Συμφωνίας ή Διαφωνίας, τοποθετούνται στα 2 άκρα και εκείνα με Ουδέτερη άποψη στη μέση. Οπότε μπορούμε να ονομάσουμε τη $1^{η}$ διάσταση ως «Κατεύθυνση Συμπεριφοράς». Οι υπόλοιπες κατηγορίες δεν συνεισφέρουν ιδιαίτερα σημαντικά στην αδράνεια. Για παράδειγμα, η κατηγορία «Liberal Husband / Traditional Wife» έχει ελάχιστη συνεισφορά στις 2 διαστάσεις και την μικρότερη μάζα συνεπώς το προφίλ της δεν σημαντικό.

Η «Ποιότητα» των αποτελεσμάτων του Πίνακα 5.3 δείχνεται από το άθροισμα των σχετικών συνεισφορών για τις 2 διαστάσεις. Καθώς η $1^{η}$ διάσταση είναι δεσπόζουσα κάθε σημείο των γραμμών και στηλών περιγράφεται πολύ καλά από αυτή. Το άθροισμα των τιμών αυτών και για τις δύο διαστάσεις ισούται με το 1.0 προφανώς.

Με το Γράφημα 5.1 αναπαριστούμε τις κατηγορίες των μεταβλητών με σημεία από τα οποία είναι ευκολότερο να εντοπίσουμε ενδιαφέρουσες σχέσεις μεταξύ των κατηγοριών αλλά και μεταξύ των μεταβλητών του Πίνακα 5.1. Αφού ο $1^{ος}$ άξονας δεσπόζει, σχεδόν όλη η πληροφορία του πίνακα συνάφειας συνοψίζεται από αυτόν. Από τον χάρτη βλέπουμε ότι τα σημεία των γραμμών 4 και 5 σχεδόν ταυτίζονται. Αυτό σημαίνει ότι οι κατηγορίες «Mod.H/Trad.W» και «Mod.H/Mod.W» της *interactive* μεταβλητής έχουν παρόμοια προφίλ. Άρα η χ^2 απόσταση μεταξύ τους είναι πολύ μικρή. Έτσι βάσει μιας ιδιότητας που έχει η χ^2 απόσταση, μπορούμε να ενοποιήσουμε τις 2 γραμμές σε 1 γραμμή χωρίς να μεταβληθεί η χ^2 απόσταση μεταξύ των στηλών. Επίσης παρατηρούμε ότι δεν υπάρχουν άλλα σημεία για τις γραμμές που να είναι κοντά μεταξύ τους άρα βγάζουμε το συμπέρασμα ότι τα προφίλ των υπολοίπων κατηγοριών είναι περίπου ανόμοια.

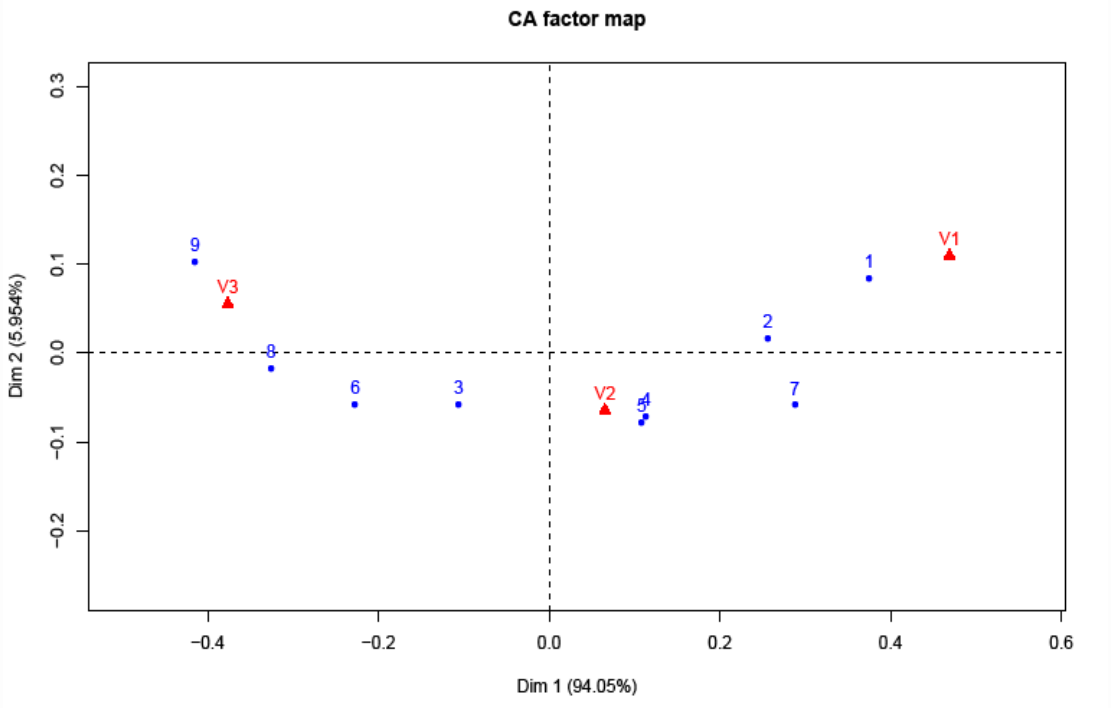
Ακόμα μπορούμε να δούμε ότι τα σημεία 4 και 5 είναι κοντά με το σημείο **V2** της *interactive* μεταβλητής. Άρα υπάρχει μια έντονη συσχέτιση μεταξύ τους. Παρόμοιο

συμπέρασμα βγάζουμε για το σημείο 1 με το σημείο **V1** αλλά και για τα σημεία 8 και 9 με το σημείο **V3**. Αυτοί οι συσχετισμοί μπορούν να ειδωθούν σαν τις αλληλεπιδράσεις που είναι στατιστικά σημαντικά στα επίπεδα της εξαρτημένης μεταβλητής.

Επίσης βλέπουμε ότι σχεδόν όλα τα «σημεία» των γραμμών στο χάρτη διαφέρουν από το μέσο προφίλ τους ή κεντροειδές των γραμμών αφού κανένα δεν είναι κοντά στην αρχή των αξόνων (origin). Τα «σημεία» που διαφέρουν περισσότερο από αυτό είναι το 1 και 9 καθώς είναι τα πιο απομακρυσμένα στο χάρτη και συνεπώς τα προφίλ τους είναι κατά πολύ διαφορετικά από το μέσο προφίλ.

Τελειώνοντας με το biplot λέμε ότι η μεταβλητότητα των σημείων γύρω από το κέντρο είναι αρκετά μεγάλη. Αυτό έχει ως αποτέλεσμα η αδράνεια να είναι και αυτή μεγάλη. Για τον Πίνακα 5.1 η αδράνεια είναι ίση με $120.39/1399 = 0.086$. Επίσης από τον ορισμό της αδράνειας προκύπτει πως οι γραμμές (στήλες) με μικρή μάζα την επηρεάζουν μόνο αν βρίσκονται μακριά από το κέντρο. Έτσι το σημείο 7 των γραμμών που αντιστοιχεί στην κατηγορία «Lib.Husb/Trad.Wife», ενώ όπως είδαμε πριν έχει μικρή μάζα παρόλα αυτά επηρεάζει την αδράνεια αρκετά αφού έχει σχετικά μεγάλη απόσταση από το κέντρο.

Γράφημα 5.1 : *Biplot (χάρτης) του Πίνακα 5.1.*



Τον Πίνακα 5.1 τον αναλύουμε ξανά χρησιμοποιώντας μοντέλα συσχέτισης για τη σύγκριση των αποτελεσμάτων με την ανάλυση αντιστοιχιών και μοντέλα συνάφειας για τη σύγκριση με τα μοντέλα συσχέτισης. Ο Πίνακας 5.4 που ακολουθεί μας δίνει τις εκτιμήσεις των μονοδιάστατων παραμετρικών σκορ που παίρνουμε από το RC(1) μοντέλο συνάφειας, RC(1) μοντέλο κανονικής συσχέτισης και RC(1) μοντέλο ανάλυσης αντιστοιχιών, με $M^* = 1$ αφού μία διάσταση είναι επαρκής για τα δεδομένα, ενώ για $M = 2$ τα μοντέλα περιγράφουν πλήρως τα δεδομένα. Αφού λοιπόν θα πάρουμε μονοδιάστατα σκορ η γραφική τους παράσταση θα είναι μιας διάστασης, ενώ για τα σκορ δύο διαστάσεων παραπέμπουμε στο Γράφημα 5.1 για την απεικόνιση τους.

Πίνακας 5.4 : *Εκτιμήσεις των παραμέτρων του RC(1) μοντέλου συνάφειας και των αντίστοιχων μοντέλων συσχέτισης για τα δεδομένα του Πίνακα 5.1.*

RC Model	Association model	Canonical model	CorrespAnal.model
	Association ϕ_1	Correlation $\sqrt{\lambda_1}$	
	0.306	0.2844	
Column score	μ_{i1}	x_{i1}	x'_{i1}
1	1.705	1.648	0.468
2	0.195	0.229	0.065
3	-1.292	-1.323	-0.376
Row score	ν_{j1}	y_{j1}	y'_{j1}
1	1.311	1.318	0.375
2	0.898	0.901	0.256
3	-0.351	-0.372	- 0.105
4	0.404	0.399	0.113
5	0.385	0.380	0.108
6	-0.784	-0.797	-0.226
7	0.996	1.011	0.287
8	-1.145	-1.144	-0.325
9	-1.474	-1.457	-0.414

Από τον πίνακα βλέπουμε ότι για τα RC(1) μοντέλα συσχέτισης είναι $\hat{\lambda}_1 \cong 0.080884$ η πρώτη ιδιοτιμή και $\sqrt{\hat{\lambda}_1} = 0.2844$ η κανονική συσχέτιση. Η τιμή αυτή είναι ίδια με την πρώτη ιδιόμορφη από την ανάλυση αντιστοιχιών. Επίσης τα σκορ για τα μοντέλα συσχέτισης δείχνουν μια συνέπεια. Για παράδειγμα, τα σκορ από το μοντέλο ανάλυσης αντιστοιχιών μπορούν να συγκριθούν με εκείνα από το μοντέλο κανονικής συσχέτισης μέσω των σχέσεων $\hat{x}'_i = \sqrt{\hat{\lambda}}\hat{x}_i$ και $\hat{y}'_i = \sqrt{\hat{\lambda}}\hat{y}_j$. Ακόμα βλέπουμε ότι τα σκορ για τις στήλες δίνουν μία διάταξη στη μεταβλητή με αρνητική τιμή για την κατηγορία «Often», καθώς και αρνητικά σκορ για τις

κατηγορίες «Lib.H/Mod.W», «Lib.Husb/Lib.Wife». Ωστόσο αν τα σκορ που υπολογίσαμε τα πολλαπλασιάσουμε με -1 η διάταξη αντιστρέφεται αλλά τα αποτελέσματα είναι παρόμοια. Επιπλέον δεν τίθεται θέμα ίσων αποστάσεων των κατηγοριών της *interactive* μεταβλητής αλλά είναι αρκετά πιθανό να γίνει για την [S].

Επίσης αφού είναι $\sqrt{\lambda_1} > 0$ και μετράει τη συσχέτιση μεταξύ των μεταβλητών του πίνακα, μπορούμε να πούμε ότι «Συχνή» Συνεισφορά στις εργασίες του σπιτιού σχετίζεται περισσότερο με την κατηγορία Liberal άντρες και Liberal γυναίκες της *interactive* μεταβλητής. Έτσι, μια πιθανή διάταξη των κατηγοριών των μεταβλητών [S] και *interactive* είναι η $\{O, S, R\}$ και $\{9,8,6,3,5,4,2,7,1\}$.

Κλείνοντας την εφαρμογή αυτή λέμε ότι το RC(1) μοντέλο συνάφειας όπως βλέπουμε από τον Πίνακα 5.4, δίνει παρόμοια σκορ με το RC(1) μοντέλο κανονικής συσχέτισης. Το γεγονός αυτό οφείλεται κυρίως στην μικρή τιμή για την παράμετρο συσχέτισης $\sqrt{\lambda}$. Επίσης, η θετική παράμετρος ϕ_1 είναι ένδειξη της συνάφειας των μεταβλητών προς την ίδια κατεύθυνση. Από τις τιμές των σκορ φαίνεται ότι οι κατηγορίες $\{4,5\}$ είναι αδιαχώριστες, συνεπώς μπορούμε να τις ενοποιήσουμε σε μία κατηγορία.

5.6 Εφαρμογή $2^{η}$

Για να εφαρμόσουμε την πολλαπλή ανάλυση αντιστοιχιών στις μεταβλητές του ερωτηματολογίου της τράπεζας θα πρέπει πρώτα να υπενθυμίσουμε ορισμένα αποτελέσματα που βρήκαμε σε προηγούμενες αναλύσεις. Στο ερωτηματολόγιο με τη παραγοντική ανάλυση βρήκαμε 4 παράγοντες όπου στον κάθε παράγοντα οι μεταβλητές είχαν υψηλά φορτία και αντίστοιχα η ανάλυση κατά συστάδες έφτιαξε 4 ομάδες μεταβλητών όπου οι μεταβλητές σε κάθε ομάδα (συστάδα) είχαν μικρή απόσταση μεταξύ τους, ενώ στο αντίστοιχο κεφάλαιο είπαμε και πως ορίζεται η απόσταση μεταξύ των μεταβλητών. Οι δύο μέθοδοι μας έδωσαν τα ίδια αποτελέσματα και διαπιστώθηκε μια σχετικά καλή συμφωνία με την αρχική δομή του ερωτηματολογίου. Επιπλέον για τις 4 ομάδες ή group ερωτήσεων δώσαμε μια γενική περιγραφή του καθενός group (δηλαδή φυσική ερμηνεία) την οποία επαναλαμβάνουμε:
Για την $1^{η}$ ομάδα ή παράγοντα δώσαμε την ονομασία «*Συνάντηση αξιολόγησης και ικανότητα των αξιολογητών-προϊσταμένων*», για την $2^{η}$ κατηγορία ή παράγοντα δώσαμε την ονομασία «*Γενική εικόνα των προϊσταμένων*», για την $3^{η}$ κατηγορία δώσαμε την ονομασία «*Στήριξη από την επιχείρηση και ποιότητα των κριτηρίων αξιολόγησης*», ενώ την $4^{η}$ ομάδα ή παράγοντα την χαρακτηρίσαμε ως «*Χρησιμότητα και αξιοπιστία του συστήματος αξιολόγησης*».

Για να κάνουμε την ανάλυση πιο κατανοητή θα στηριχτούμε σε ένα μικρότερο σύνολο μεταβλητών (ερωτήσεων) ενώ παράλληλα θα χρησιμοποιήσουμε ορισμένες από τις δημογραφικές μεταβλητές του ερωτηματολογίου οι οποίες δεν χρησιμοποιήθηκαν στην αρχική ανάλυση των Κεφαλαίων 2 και 3. Ο λόγος είναι ότι αφού τα άτομα είναι ανώνυμα στην έρευνα, οι μεταβλητές αυτές θα μας δώσουν επιπρόσθετες πληροφορίες και θα μας βοηθήσουν στην ερμηνεία των σχέσεων ανάμεσα στις διάφορες κατηγορικές μεταβλητές.

Πριν προχωρήσουμε με τη μέθοδο ελέγχουμε πρώτα τα συνολικά ποσοστά όλων των κατηγοριών των μεταβλητών που χρησιμοποιήθηκαν από τις προηγούμενες αναλύσεις καθώς και το ποσοστό εκείνων που δεν απάντησαν. Υπενθυμίζουμε ότι οι μεταβλητές δίνονται σε πενταβάθμια *Likert* κλίμακα με κατηγορίες: 1=Διαφωνώ απολύτως,….,5=Συμφωνώ απολύτως. Έτσι, το συνολικό ποσοστό για την $1^{η}$ κατηγορία είναι 12.3% και το οποίο σημαίνει ότι κατά μέσο όρο οι ερωτηθέντες απάντησαν ότι Διαφωνούν απολύτως σε όλες τις ερωτήσεις με ένα ποσοστό της τάξης περίπου του 12%. Αντίστοιχα το ποσοστό για την $2^{η}$ κατηγορία στο σύνολο των ερωτήσεων είναι 24.3%, για την $3^{η}$ κατηγορία ο μέσος όρος είναι 28% που είναι και ο μεγαλύτερος, παρόμοια για την $4^{η}$ κατηγορία το ποσοστό της είναι 27% και τέλος για την $5^{η}$ κατηγορία είναι μόλις το 8.4% του συνόλου. Άρα από τα νούμερα αυτά

βλέπουμε ότι το 79.3% των απαντήσεων δίνεται στην $2^{η}$, $3^{η}$ και $4^{η}$ κατηγορία ενώ μόλις το 20.7% στις 2 ακραίες κατηγορίες. Συνεπώς αναμένουμε ότι με την πολλαπλή ανάλυση αντιστοιχιών στο χάρτη όπου θα απεικονίσουμε τις κατηγορίες των μεταβλητών, οι αποστάσεις μεταξύ των σημείων να είναι τέτοιες έτσι ώστε ο διαφορές ανάμεσα στις κατηγορίες των μεταβλητών να είναι εμφανείς. Δηλαδή για τη $2^{η}$, $3^{η}$ και $4^{η}$ κατηγορία των μεταβλητών, κατά μέσο όρο, θα είναι μαζεμένες κοντά, ενώ η $1^{η}$ και η $5^{η}$ κατηγορία των μεταβλητών θα είναι απομακρυσμένες. Όσον αφορά τις χαμένες απαντήσεις (missing values) το ποσοστό τους είναι στο 0.25% και είναι πολύ μικρό για να επηρεάσει το αποτέλεσμα.

Το επόμενο που πρέπει να αποφασίσουμε είναι η επιλογή των μεταβλητών που θα χρησιμοποιήσουμε για την ανάλυση. Οι μεταβλητές που θα επιλέξουμε θα πρέπει κατά τη γνώμη μας να αντικατοπτρίζουν όσο το δυνατό περισσότερο τον στόχο της έρευνας που δεν είναι άλλος από τη διερεύνηση της αποτελεσματικότητας του συστήματος αξιολόγησης του προσωπικού της τράπεζας. Επιπλέον, οι μεταβλητές θα προέρχονται και από τις 4 διαφορετικές ομάδες που δημιουργήθηκαν από προηγούμενη ανάλυση για να προσεγγίσουμε το θέμα από όλες τις πλευρές. Αυτό είναι μια δύσκολη επιλογή. Ωστόσο δεν υπάρχει κάποιος περιορισμός στον αριθμό των μεταβλητών που θα επιλέξουμε αλλά για να έχει πιο «ευδιάκριτη» εικόνα η παρουσίαση θα βασιστούμε σε ένα μειωμένο αριθμό.

Μια εναλλακτική λύση στην επιλογή μεταβλητών είναι να πάρουμε μόνο της «ηγούσες» μεταβλητές (surrogate variables) των παραγόντων αλλά η επιλογή αυτή δεν βοηθάει ιδιαίτερα τον σκοπό που αναφέραμε στην προηγούμενη παράγραφο, καθώς οι «ηγούσες» απλώς είναι μεταβλητές που έχουν υψηλά φορτία σε κάθε παράγοντα. Αυτό δεν τις καθιστά απαραίτητα και τις πιο «κατάλληλες» για το στόχο της έρευνας. Παρόλο αυτά είναι μια εναλλακτική λύση αρκετά συχνή ειδικά στις περιπτώσεις που οι «ηγούσες» παίρνουν το ρόλο των εξαρτημένων (response) μεταβλητών στην κατασκευή μοντέλων.

Η επιλογή των μεταβλητών αυτών δεν είναι μοναδική. Με μια προσεχτική μελέτη του ερωτηματολογίου ο ενδιαφερόμενος αναγνώστης πιθανώς να εντοπίσει άλλες για να εξετάσει το ζήτημα αυτό. Πάντως μπορούμε πούμε ότι σε γενικές γραμμές τα συμπεράσματα που θα βγάλει θα συμφωνούν σε μεγάλο βαθμό με αυτά που βγάλαμε και εμείς, γιατί υπάρχει μια ομοιογένεια (uniformity) στις απαντήσεις, δεν παρατηρούνται δύσκολες ερωτήσεις που να μπερδεύουν τους ερωτώμενους και οι ερωτήσεις μεταξύ τους είναι σχετικές.

Συνεπώς, θεωρούμε ότι η αποτελεσματικότητα του συστήματος αξιολόγησης που διατηρεί η τράπεζα, σε γενικές γραμμές μπορεί να κριθεί από τα εξής 4 σημεία:

- → Τους αξιολογητές-προϊστάμενους
- → Τα κριτήρια αξιολόγησης
- → Την «ποιότητα» κατά την διάρκεια της συνάντησης αξιολόγησης
- → Την ικανότητα για τον εντοπισμό συγκεκριμένων ιδιοτήτων και θέσεων

Τα παραπάνω σημεία είναι ποιοτικά και ποσοτικά χαρακτηριστικά που κατά τη γνώμη μας μπορούν να καθορίσουν και παράλληλα να εντοπίσουν σε μεγάλο βαθμό ένα καλό σύστημα αξιολόγησης στελεχών και προσωπικού μιας επιχείρησης. Θέτοντας η επιχείρηση τις κατάλληλες ερωτήσεις και χωρίς κατά ανάγκη να «πλατιάζει», είναι πιθανό μέσω κατάλληλων ποσοτικών μεθόδων να εκτιμηθεί οποιοσδήποτε βαθμός αποτελεσματικότητας και καταλληλότητας αυτού. Όσον αφορά την πολλαπλή ανάλυση αντιστοιχιών, με τη μέθοδο αυτή θα προσπαθήσουμε να εντοπίσουμε συσχετισμούς των κατηγοριών των μεταβλητών και αποστάσεις μεταξύ αυτών που θα αντικατοπτρίζουν τις διαφοροποιήσεις τους. Επιπροσθέτως, η χρησιμοποίηση συμπληρωματικών μεταβλητών να μας βοηθήσει να δούμε ποιοι δίνουν τις σχετικές απαντήσεις και πιθανώς αυτό να φανεί πιο χρήσιμο για την ερμηνεία των αποτελεσμάτων.

Οι ερωτήσεις/μεταβλητές που θα χρησιμοποιήσουμε από το ερωτηματολόγιο είναι οι εξής:

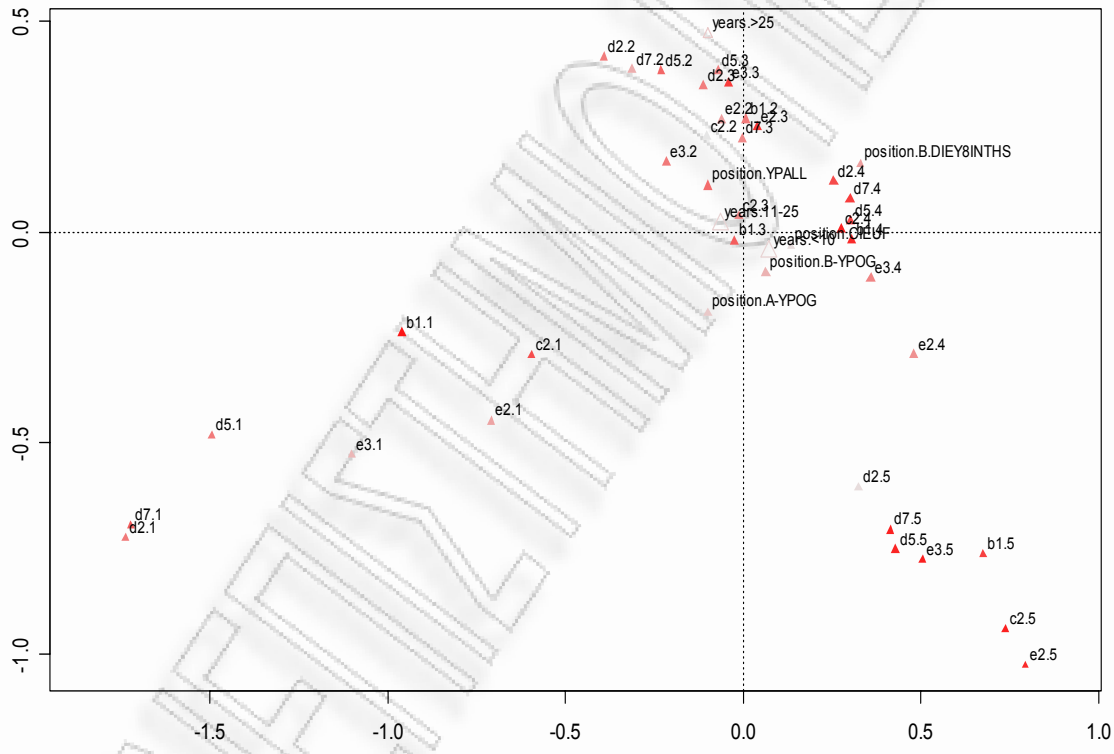
- → Η ερώτηση: «Κατά τη διάρκεια της χρονιάς υπάρχουν τακτικές συναντήσεις για βοήθεια και κατεύθυνση», με την κωδικοποίηση «b1»,
- → Η ερώτηση: «Οι εργαζόμενοι αξιολογούνται με ποσοτικά κριτήρια (στόχοι)», με την κωδικοποίηση «e2»,
- → Η ερώτηση: «Οι εργαζόμενοι αξιολογούνται με ποιοτικά (ιδιότητες – ικανότητες) κριτήρια», με την κωδικοποίηση «e3»,
- → Η ερώτηση: «Στη συνάντηση συζητούνται και αξιολογούνται όλα τα σημαντικά μέρη της εργασίας κάθε εργαζόμενου», με την κωδικοποίηση «c2»,
- → Η ερώτηση: «Η αξιολόγηση βοηθάει να καταλαβαίνουν οι εργαζόμενοι τα δυνατά και αδύνατα σημεία τους στην διεκπεραίωση της εργασίας», με την κωδικοποίηση «d2»,
- → Η ερώτηση: «Η αξιολόγηση βοηθάει να καταλαβαίνουν οι εργαζόμενοι την συνεισφορά τους στην Επιχείρηση», με την κωδικοποίηση «d5» και
- → Η ερώτηση: «Η αξιολόγηση βοηθάει στο να καταλάβει ο εργαζόμενος τι είναι σημαντικό στη δουλειά του», με την κωδικοποίηση «d7».

Επιπλέον για να γίνουν περισσότερο κατανοητά τα αποτελέσματα χρησιμοποιούμε και τις εξής δημογραφικές μεταβλητές: 1. «Χρόνια εργασίας στην τράπεζα», με 3 επίπεδα (Κάτω

από 10, Από 11-25, Πάνω από 25), και 2. «Θέση», με 5 επίπεδα (Διευθυντής, Β. Διευθυντής, Α-Υπογραφή, Β-Υπογραφή, Υπάλληλος).

Τα αποτελέσματα παρουσιάζονται στο Γράφημα 5.2 που είναι ένα συμμετρικό *biplot* των 2 πρώτων αξόνων για τις κατηγορίες των στηλών, δηλαδή οι κύριες συντεταγμένες είναι κλιμακοποιημένες έτσι ώστε τα σημεία των στηλών με μεγάλη μάζα να μην επηρεάζουν σημαντικά το γράφημα (principal coordinates).

Γράφημα 5.2 : *Αναπαράσταση όλων των κατηγοριών στους δύο πρώτους άξονες.*



Τα αποτελέσματα της ανάλυσης για τις στήλες υπάρχουν στο Παράρτημα Γ (μάζες, χ^2 αποστάσεις, αδράνειες, συντεταγμένες, ποιότητα). Για την ανάλυση, η μεταβλητή «Χρόνια εργασίας στην τράπεζα» (*years*), τοποθετήθηκε σαν συμπληρωματική και συνεπώς δεν χρησιμοποιήθηκε για την κατασκευή των αξόνων και την εύρεση των συντεταγμένων, απλώς απεικονίζεται πάνω στο χάρτη για να αυξήσει την ερμηνευτικότητα και πιθανώς να μας δώσει επιπρόσθετες πληροφορίες για τη συμπεριφορά των βασικών μεταβλητών.

Η πολλαπλή ανάλυση αντιστοιχιών του πίνακα Burt δίνει 32 διαστάσεις εκ των οποίων μόνο 2 είναι μεγαλύτερες του 1/8 κριτηρίου (αφού έχουμε 8 μεταβλητές). Οι δύο πρώτες

κύριες αδράνειες είναι 0.2399 και 0.169 και ερμηνεύουν το 46.9% της ολικής αδράνειας. Όμως η αδράνεια που προκύπτει με τον τρόπο αυτό είναι κάπως υπερκτιμημένη αφού οι διαγώνιοι υποπίνακες του Burt έχουν μέγιστη αδράνεια (Υπενθυμίζουμε ότι υπάρχουν 64 υποπίνακες συνολικά). Για την ακρίβεια, η αδράνεια όλων των υποπινάκων εξαιρουμένης της διαγωνίου είναι 0.186 ενώ η αδράνεια του πίνακα Burt είναι 0.873 και το οποίο δείχνει την υπερεκτίμηση που υπάρχει στα δεδομένα. Για τον λόγο αυτό θα χρησιμοποιήσουμε τις προσαρμοσμένες βασικές αδράνειες και τις οποίες θα υπολογίσουμε και αυτές θα χρησιμοποιήσουμε για την περαιτέρω ανάπτυξη (εύρεση των συντεταγμένων, συνεισφορές, ποιότητα, κλπ).

Ο τρόπος που αλλάξαμε την κλίμακα για τις αδράνειες είναι κάπως αυθαίρετος όποτε το ποσοστό της αδράνειας που ερμηνεύεται από κάθε άξονα πρέπει να εξεταστεί με προσοχή. Έτσι τώρα οι 2 πρώτες βασικές αδράνειες είναι 0.174 και 0.107 αντίστοιχα, ενώ εξηγούν περίπου το 65.9% της συνολικής αδράνειας. Η αδράνεια του πίνακα των δεδομένων είναι 0.4265 και ως γνωστό είναι το άθροισμα όλων των ιδιοτιμών, ενώ το ποσοστό της αδράνειας δεν είναι 100% όπως αναμενόταν αλλά 84.8% και αυτό συμβαίνει γιατί η αδράνεια όταν χρησιμοποιούμε τις προσαρμοσμένες τιμές δεν αθροίζει στο 100% (Greenacre, 1993, 1998). Άρα 2 άξονες μπορούν να ερμηνεύσουν ικανοποιητικά τα δεδομένα μας και να τα αναπαραστήσουμε σε ένα διδιάστατο γράφημα.

Επιστρέφοντας στο Γράφημα 5.2 αρχικά βλέπουμε ότι οι αποστάσεις ανάμεσα στα σημεία που αντιπροσωπεύουν τις διάφορες κατηγορίες των μεταβλητών αποκαλύπτουν τις ομοιότητες ή διαφοροποιήσεις σε αυτές. Δηλαδή είναι ξεκάθαρο ότι οι κατηγορίες των μεταβλητών 2, 3, 4 είναι πολύ κοντά μεταξύ τους ενώ οι κατηγορίες 1 και 5 είναι απομακρυσμένες. Αυτό συμβαίνει για όλες τις μεταβλητές. Επίσης, παρατηρούμε ότι στο $1^{ο}$ από αριστερά τεταρτημόριο υπάρχουν σε μία μεγάλη ομάδα όλες οι μεταβλητές για την $1^{η}$ κατηγορία με τις {d2.1 και d7.1} σχεδόν να ταυτίζονται και άρα μπορούμε να πούμε ότι υπάρχει μεγάλη συσχέτιση μεταξύ τους. Οι κατηγορίες αυτές, όπως προκύπτει από τα αποτελέσματα, έχουν μικρές μάζες με αποτέλεσμα να μην είναι σημαντικές ενώ βρίσκονται και σε μεγάλη απόσταση από τον μέσο. Επίσης στο $2^{ο}$ από αριστερά τεταρτημόριο είναι μαζεμένες πολύ κοντά οι κατηγορίες 2 και 3 των διαφορετικών μεταβλητών όπου παρατηρείται και πάλι μια έντονη συσχέτιση μεταξύ των κατηγοριών διαφορετικών μεταβλητών. Για παράδειγμα, οι {d5.2, d2.3, e3.3} είναι πολύ συσχετισμένες. Επιπλέον βλέπουμε ότι η θέση «Υπάλληλος» ανήκει μέσα στην ομάδα αυτή και πιο πολύ σχηματίζει

ένα group με τις μεταβλητές {e3.2, c2.2, c2.3, e2.2, d7.3}, άρα η κατηγορία αυτή του προσωπικού σχετίζεται περισσότερο με τις παραπάνω απαντήσεις και ερωτήσεις (δηλαδή συμβαίνει/συναντιέται πιο «συχνά»).

Αντίστοιχα στο 3^{ο} από αριστερά τεταρτημόριο του χάρτη βλέπουμε την 4^{η} κατηγορία των διαφορετικών μεταβλητών και την {e2.3} της 3^{ης}. Βέβαια δεν μπορούμε να πούμε ότι η {e2.3} σχηματίζει ομάδα με την 4^{η} κατηγορία απλά μόνο ότι ανήκει στο τεταρτημόριο αυτό. Αντιθέτως οι διαφορετικές μεταβλητές της 4^{ης} κατηγορίας σχηματίζουν μια συμπαγή ομάδα πλην της {e2.4} που είναι περισσότερο απομακρυσμένη. Εδώ μπορούμε να πούμε ότι η κατηγορία αυτή των διαφορετικών μεταβλητών συσχετίζεται περισσότερο με τη θέση στην τράπεζα του «Β.Διευθυντής», όπου προφανώς συναντιέται περισσότερο.

Στο 4^{ο} τεταρτημόριο είναι συγκεντρωμένες οι μεταβλητές της 5^{ης} κατηγορίας όπου μπορούμε να πούμε ότι σχηματίζουν 2 group με τις {d2.5, d7.5, d5.5, e3.5, b1.5} και {c2.5, e2.5} αντίστοιχα. Επίσης για τις {c2.5, e2.5} βλέπουμε ότι είναι «ακραίες», δηλαδή εντελώς απομακρυσμένες ενώ έχουν και πολύ μικρές μάζες με αποτέλεσμα να μην έχουν μεγάλες συχνότητες και να μην επηρεάζουν την ανάλυση. Όσον αφορά κάποια από τις δημογραφικές μεταβλητές που να συσχετίζεται μαζί τους, από το παραπάνω συμμετρικό *biplot* δεν φαίνεται κάτι τέτοιο. Μια ερμηνεία που μπορούμε να δώσουμε είναι ότι οι κατηγορίες αυτές των μεταβλητών δεν είναι πολύ σημαντικές και πρέπει να απομακρυνθούν από μια μελλοντική έρευνα, παρόλα αυτά, τα σημεία τους στο χάρτη περιγράφονται ικανοποιητικά από τους άξονες (έχουν υψηλές σχετικές συνεισφορές).

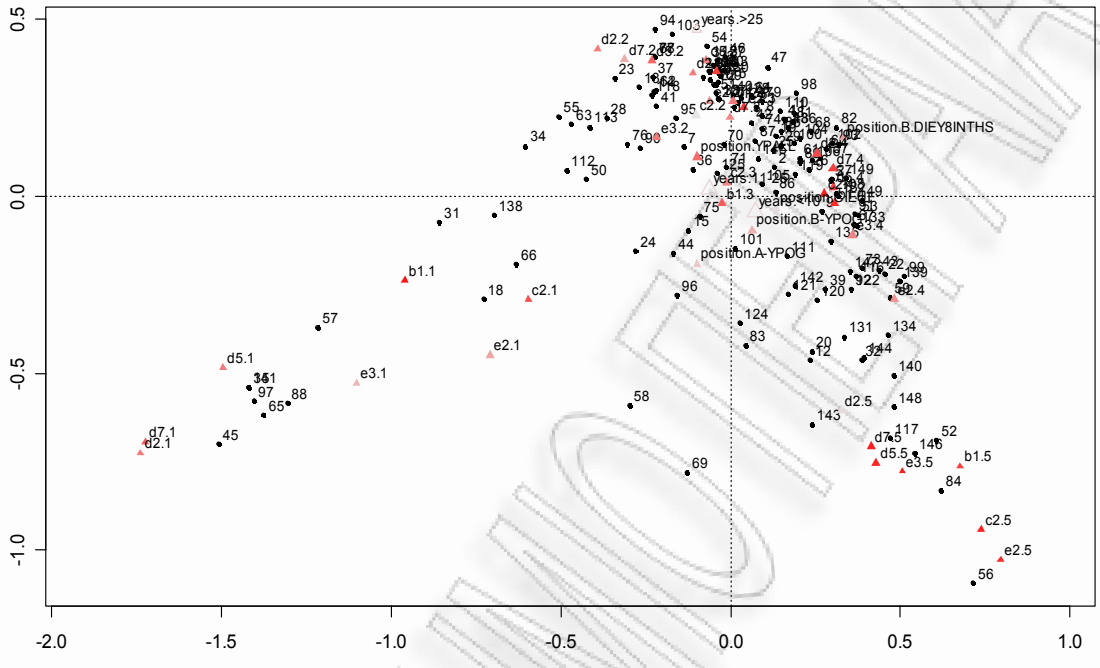
Η θέση του «Διευθυντής» είναι μεταξύ της 4^{ης} κατηγορίας των μεταβλητών και του μέσου ή κεντροειδές των κατηγοριών η δε συσχέτιση του είναι πολύ μεγάλη με τις {b1.3, c2.3} που από πλευράς σημαντικότητας έχουν ιδιαίτερη σημασία αφού αναφέρονται στην «ποιότητα» των αξιολογητών και της συνάντησης αξιολόγησης. Έτσι, κυρίως τα διευθυντικά στελέχη έχουν μια ουδέτερη άποψη για την αποτελεσματικότητα αυτών, αν αναλογιστούμε και την βαρύτητα που έχει η θέση αυτή. Οι θέσεις «Α και Β Υπογραφή» δεν συνεισφέρουν περισσότερο στην ερμηνεία των αποτελεσμάτων ενώ η απόσταση τους είναι σχετικά μικρή. Χωρίς να επηρεαστεί η αξιοπιστία της ανάλυσης θα μπορούσαμε να τις ενοποιήσουμε ή και ακόμα να μην τις λάβουμε υπόψη σε νέα έρευνα. Θα πρέπει εδώ να πούμε ότι για τις δημογραφικές μεταβλητές και την απεικόνιση των συσχετίσεων με τις υπόλοιπες ένα μη-συμμετρικό *biplot* θα μας καταδείξει ότι οι κατηγορίες αυτές σχεδόν ταυτίζονται και

συνεισφέρουν μόνο με την κατηγορία «Κάτω από 10» χρόνια εργασίας σε ένα κάπως «ουδέτερο» χώρο πάνω στο γράφημα.

Χρησιμοποιώντας στην ερμηνεία τη συμπληρωματική μεταβλητή «Χρόνια εργασίας στην τράπεζα» (*years*) βλέπουμε ότι αυτοί που δουλεύουν πάνω από 25 χρόνια στην τράπεζα συσχετίζονται περισσότερο με το group {d2.2, d7.2, d5.2, d5.3, d2.3, e3.3}, ενώ εκείνοι που δουλεύουν από 11 έως 25 χρόνια απαρτίζουν ένα group με τη θέση «Υπάλληλος» και ότι συνεπάγεται από πριν για αυτήν, πλέον τη σύσταση μιας νέας υποομάδας με τις {b1.3, c2.3} αλλά και τη θέση του «Διευθυντής» από πριν. Άρα βλέπουμε ότι δημιουργείται ένα ευρύτερο group ή 2 μικρότερα group αυτού. Τέλος, αυτοί που έχουν κάτω από 10 χρόνια προϋπηρεσίας απλά δημιουργούν μία ομάδα με τη θέση «Β Υπογραφή» χωρίς να μας παρέχουν επιπλέον πληροφορίες για τη συμπεριφορά των μεταβλητών της έρευνας.

Ένα άλλο στοιχείο είναι ότι παρατηρείται μία τάση της $1^{ης}$ κατηγορίας των διαφορετικών μεταβλητών του group {c2.1, e2.1, e3.1, b1.1 και d5.1} να κινηθεί προς το $2^{ο}$ τεταρτημόριο οπότε τα προφίλ των στηλών αυτών να μοιάσουν μεταξύ τους. Αυτό ίσως να είναι δείγμα ότι σε μελλοντική έρευνα για αντίστοιχο ή ίδιο αντικείμενο ενδεχομένως κάποιες από τις κατηγορίες των ερωτήσεων αυτών να χρειαστεί να ενοποιηθούν, αφού τα προφίλ τους θα είναι ίδια. Βέβαια παρόμοια συμπεράσματα μπορούν να προκύψουν και για κάποιες μεταβλητές της $3^{ης}$ κατηγορίας που «κινούνται» προς την $4^{η}$. Έτσι ένα νέο ερωτηματολόγιο δεν χρειάζεται να έχει κατά ανάγκη τις ίδιες (ίσες) κατηγορίες σε όλες τις μεταβλητές του (ερωτήσεις).

Γράφημα 5.3 : *Αναπαράσταση όλων των κατηγοριών και των παρατηρήσεων στους δύο πρώτους άξονες.*



Από το Γράφημα 5.3 βλέπουμε ότι το «νέφος» των παρατηρήσεων είναι συγκεντρωμένο κυρίως στην $4^{η}$ και $3^{η}$ κατηγορία των διαφορετικών μεταβλητών. Με βάση τα όσα είπαμε πριν προκύπτει ότι οι απαντήσεις του δείγματος κινούνται γύρω από αυτές τις κατηγορίες των μεταβλητών ενώ υπάρχει μεγάλη διασπορά των παρατηρήσεων για τις υπόλοιπες κατηγορίες. Βέβαια υπάρχουν ελάχιστες παρατηρήσεις στην $5^{η}$ κατηγορία. Συνεπώς βάσει της αρχικής υπόθεσης που κάναμε ότι υπάρχουν 4 βασικά σημεία και ένα υποσύνολο ερωτήσεων που μπορεί να απαντήσει στο ερώτημα για την αποτελεσματικότητα ή όχι του υπάρχοντος συστήματος αξιολόγησης της τράπεζας, από τα δεδομένα προκύπτει πως ναι αυτό μπορεί να συμβαίνει και το σύστημα να είναι αποτελεσματικό αλλά μόνο για την υπάρχουσα δομή του ερωτηματολογίου. Επιπλέον θα πρέπει να κρατήσουμε στα υπόψη την παρούσα κατάσταση όπως αυτή διατυπώνεται από τα γραφήματα και να τη συγκρίνουμε με νέα έρευνα για να δούμε εάν πράγματι προκύψουν μεταβολές της εικόνας στο μέλλον. Στην περίπτωση όμως που το ερωτηματολόγιο επανασχεδιαστεί και τεθούν υπόψη ορισμένες από τις παρατηρήσεις που έγιναν, ενδέχεται να προκύψουν άλλα συμπεράσματα αφού τα δεδομένα πλέον θα έχουν αλλάξει.

Κλείνοντας αναφέρουμε ότι η γενική εικόνα της έρευνας που συνοψίστηκε από το μικρότερο σύνολο των 8+1 μεταβλητών δεν αλλάζει όταν πάρουμε προς ανάλυση κάποιο άλλο συναφές υποσύνολο.

ΠΑΡΑΡΤΗΜΑ Α: Ερωτηματολόγιο της έρευνας

Το ερωτηματολόγιο αυτό έχει στόχο να ερευνήσει την αποτελεσματικότητα του συστήματος αξιολόγησης της Τράπεζας *ABC* .

Η συμπλήρωση του παρόντος ερωτηματολόγίου είναι **ΑΝΩΝΥΜΗ**

Σας παρακαλούμε να απαντήσετε σε όλες τις ερωτήσεις με σκέψη και ειλικρίνεια για να βοηθήσετε στη προσπάθεια της Διοίκησης της Τράπεζας να βελτιώσει το υπάρχον σύστημα αξιολόγησης

Ευχαριστούμε πολύ για τη συνεργασία

ΔΗΜΟΓΡΑΦΙΚΑ ΣΤΟΙΧΕΙΑ (βάλτε ένα √ στο κατάλληλο κουτάκι) :

	κάτω από 30	από 31-45	πάνω από 45
ΗΛΙΚΙΑ	☐	☐	☐

	Κάτω από 10	από 11- 25	πάνω από 25
ΧΡΟΝΙΑ ΕΡΓΑΣΙΑΣ ΣΤΗΝ ΤΡΑΠΕΖΑ	☐	☐	☐

	Λύκειο	Ανώτερη	Ανώτατη
ΕΚΠΑΙΔΕΥΤΙΚΟ ΕΠΙΠΕΔΟ	☐	☐	☐

Δίκτυο	Διοίκηση
☐	☐

	Δ/ντής	Υποδ/τής	Α-Υπογραφή	Β-Υπογραφή	Υπάλληλος
ΘΕΣΗ	☐	☐	☐	☐	☐

	Αθήνα	Υπόλοιπη Ελλάδα
ΓΕΩΓΡΑΦΙΚΗ ΠΕΡΙΟΧΗ	☐	☐

Α/Α	ΠΕΡΙΓΡΑΦΗ	ΔΙΑΦΩΝΩ ΑΠΟΛΥΤΩΣ	ΔΙΑΦΩΝΩ	ΟΥΤΕ ΣΥΜΦΩΝΩ ΟΥΤΕ ΔΙΑΦΩΝΩ	ΣΥΜΦΩΝΩ	ΣΥΜΦΩΝΩ ΑΠΟΛΥΤΩΣ
Α. Η ΔΙΑΔΙΚΑΣΙΑ – ΤΟ ΣΥΣΤΗΜΑ						
1	Η επιχείρηση στηρίζει σοβαρά το σύστημα αξιολόγησης	1	2	3	4	5
2	Η ανώτερη διοίκηση δίνει πρώτη το παράδειγμα στην σωστή αξιολόγηση	1	2	3	4	5
3	Στην επιχείρηση αξιολογούνται όλοι χωρίς εξαίρεση	1	2	3	4	5
4	Αν υπάρχουν διαφωνίες στην αξιολόγηση λύνονται αρχικά με συζήτηση	1	2	3	4	5
5	Οι εργαζόμενοι είναι καλά ενημερωμένοι για το σκοπό και την λειτουργία του συστήματος αξιολόγησης	1	2	3	4	5
6	Οι προϊστάμενοι – αξιολογητές ξέρουν να αξιολογούν σωστά	1	2	3	4	5
7	Το σύστημα εγγυάται αντικειμενικότητα και δικαιοσύνη στην αξιολόγηση	1	2	3	4	5
8	Όλοι παίρνουν πολύ στα σοβαρά την αξιολόγηση	1	2	3	4	5
9	Οι εργαζόμενοι ειδοποιούνται έγκαιρα π χ μια εβδομάδα πριν την συνάντησης αξιολόγησης	1	2	3	4	5
10	Αν υπάρχουν σοβαρές διαφωνίες στην αξιολόγηση , επιλύονται από ειδικό για το σκοπό αυτό όργανο	1	2	3	4	5
11	Οι αξιολογητές είναι εκπαιδευμένοι ώστε να αξιολογούν με αντικειμενικό και εποικοδομητικό τρόπο	1	2	3	4	5
12	Το σύστημα αξιολόγησης είναι μια ζωντανή διαδικασία και όχι μια γραφειοκρατική συμπλήρωση εντύπων	1	2	3	4	5
13	Οι προϊστάμενοι θεωρούν την αξιολόγηση των εργαζομένων τους σαν ένα σημαντικό μέρος των καθηκόντων τους	1	2	3	4	5
14	Ο αξιολογούμενος ενημερώνεται για την βαθμολογία του	1	2	3	4	5
15	Η επιχείρηση θεωρεί την αξιολόγηση των εργαζομένων σαν ένα από τα σημαντικά καθήκοντα των προϊσταμένων	1	2	3	4	5

Α/Α	ΠΕΡΙΓΡΑΦΗ	ΔΙΑΦΩΝΩ ΑΠΟΛΥΤΩΣ	ΔΙΑΦΩΝΩ	ΟΥΤΕ ΣΥΜΦΩΝΩ ΟΥΤΕ ΔΙΑΦΩΝΩ	ΣΥΜΦΩΝΩ	ΣΥΜΦΩΝΩ ΑΠΟΛΥΤΩΣ
Β. Η ΚΑΘΟΔΗΓΗΣΗ ΚΑΤΑ ΤΗΝ ΔΙΑΡΚΕΙΑ ΤΟΥ ΧΡΟΝΟΥ						
1	Κατά τη διάρκεια της χρονιάς υπάρχουν τακτικές συναντήσεις για βοήθεια και κατεύθυνση	1	2	3	4	5
2	Οι προϊστάμενοι δίνουν τακτικά πληροφόρηση για την πορεία της δουλειάς σε κάθε εργαζόμενο	1	2	3	4	5
3	Οι προϊστάμενοι δεν περιμένουν την αξιολόγηση για να εκφράσουν την άποψή τους για τους εργαζόμενους	1	2	3	4	5
4	Οι προϊστάμενοι κάνουν συγκεκριμένες υποδείξεις βελτίωσης όλη τη χρονιά	1	2	3	4	5
5	Οι προϊστάμενοι έχουν έγκαιρα προειδοποιήσει για τις συνέπειες της μέτριας απόδοσης	1	2	3	4	5
6	Οι προϊστάμενοι ενισχύουν και ενθαρρύνουν τους εργαζόμενους να επιτύχουν τους στόχους τους	1	2	3	4	5
7	Όταν κάποιος αποδίδει πολύ καλά ο προϊστάμενος του το βλέπει και του κάνει θετικά σχόλια επιβράβευσης	1	2	3	4	5
8	Όταν κάποιος δεν αποδίδει καλά ο προϊστάμενος του το βλέπει και τον βοηθάει παρεμβαίνοντας άμεσα με υποδείξεις	1	2	3	4	5
9	Οι εργαζόμενοι παίρνουν έγκαιρα πληροφόρηση σχετικά με την απόδοσή τους και έτσι είναι χρήσιμη	1	2	3	4	5
10	Οι προϊστάμενοι διαθέτουν την απαραίτητη επικοινωνιακή επάρκεια για να καθοδηγούν τους εργαζόμενους	1	2	3	4	5
Γ. Η ΣΥΝΑΝΤΗΣΗ ΑΞΙΟΛΟΓΗΣΗΣ						
1	Υπάρχει συνάντηση αξιολόγησης μεταξύ προϊσταμένου και υφισταμένου	1	2	3	4	5
2	Στη συνάντηση συζητιόνται και αξιολογούνται όλα τα σημαντικά μέρη της εργασίας κάθε εργαζόμενου	1	2	3	4	5
3	Αξιολογείται κάθε μέρος της εργασίας, ένα – ένα αναλυτικά	1	2	3	4	5
4	Τα σχόλια από τα μεριά του προϊσταμένου είναι συγκεκριμένα	1	2	3	4	5
5	Εντοπίζονται οι περιοχές που ο εργαζόμενος χρειάζεται βελτίωση	1	2	3	4	5

6	Ο εργαζόμενος ενισχύεται με θετικά λόγια για τις επιτυχίες του	1	2	3	4	5
7	Συμφωνείται εκπαιδευτική παρέμβαση	1	2	3	4	5
8	Συμφωνείται σχέδιο δράσης για την επόμενη περίοδο	1	2	3	4	5
9	Υπάρχει άνεση να μιλήσει κανείς και να πει τη γνώμη του	1	2	3	4	5
10	Η ατμόσφαιρα είναι φιλική	1	2	3	4	5
11	Δεν υπάρχουν διακοπές από τηλέφωνα ή επισκέπτες	1	2	3	4	5
12	Οι προϊστάμενοι αξιολογούν χωρίς να κάνουν άσχημα σχόλια και κριτική της προσωπικότητας του εργαζόμενου	1	2	3	4	5
13	Οι προϊστάμενοι είναι αντικειμενικοί και δίκαιοι	1	2	3	4	5
14	Η βαθμολόγησή είναι δίκαιη και αντικατοπτρίζει την πραγματική απόδοση του εργαζόμενου	1	2	3	4	5
15	Οι προϊστάμενοι είναι καλά προετοιμασμένοι πριν από τη συνάντηση	1	2	3	4	5
16	Οι εργαζόμενοι είναι καλά προετοιμασμένοι πριν από τη συνάντηση αξιολόγησης	1	2	3	4	5
17	Ο χρόνος που διατίθεται είναι επαρκής	1	2	3	4	5
18	Η συζήτηση με το προϊστάμενο είναι ειλικρινής	1	2	3	4	5
19	Ο προϊστάμενος συζητάει μαζί με τον εργαζόμενο τους λόγους που τον οδήγησαν στην συγκεκριμένη βαθμολογία	1	2	3	4	5
20	Οι προϊστάμενοι έχουν την απαραίτητη επικοινωνιακή επάρκεια που χρειάζεται για την σωστή αξιολόγηση	1	2	3	4	5
21	Δεν υπάρχουν εκπλήξεις από την μεριά του προϊσταμένου σχετικά με την βαθμολογία μου	1	2	3	4	5
22	Εμπιστεύομαι την κρίση του προϊσταμένου μου σχετικά με την απόδοσή μου	1	2	3	4	5

Α/Α	ΠΕΡΙΓΡΑΦΗ	ΔΙΑΦΩΝΩ ΑΠΟΛΥΤΩΣ	ΔΙΑΦΩΝΩ	ΟΥΤΕ ΣΥΜΦΩΝΩ ΟΥΤΕ ΔΙΑΦΩΝΩ	ΣΥΜΦΩΝΩ	ΣΥΜΦΩΝΩ ΑΠΟΛΥΤΩΣ
Δ. Η ΧΡΗΣΙΜΟΤΗΤΑ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ						
1	Το σύστημα αξιολόγησης είναι χρήσιμο για όλους	1	2	3	4	5
2	Η αξιολόγηση βοηθάει να καταλαβαίνουν οι εργαζόμενοι τα δυνατά και αδύνατα σημεία τους στην διεκπεραίωση της εργασίας	1	2	3	4	5
3	Η αξιολόγηση βοηθάει τους εργαζόμενους να σχεδιάσουν ένα πρόγραμμα προσωπικής βελτίωσης	1	2	3	4	5
4	Η αξιολόγηση έχει σαν αποτέλεσμα οι εργαζόμενοι να παρακολουθούν χρήσιμα για την εργασία τους εκπαιδευτικά προγράμματα	1	2	3	4	5
5	Η αξιολόγηση βοηθάει να καταλαβαίνουν οι εργαζόμενοι την συνεισφορά τους στην Επιχείρηση	1	2	3	4	5
6	Η αξιολόγηση βοηθάει στη βελτίωση της επικοινωνίας προϊστάμενου-υφισταμένου	1	2	3	4	5
7	Η αξιολόγηση βοηθάει στο να καταλάβει ο εργαζόμενος τι είναι σημαντικό στη δουλειά του	1	2	3	4	5
8	Η αξιολόγηση βοηθάει τους εργαζόμενους να βελτιώσουν την απόδοσή τους	1	2	3	4	5
9	Το σύστημα αξιολόγησης βοηθάει να ξεχωρίσουν οι πραγματικά ικανοί	1	2	3	4	5
10	Το σύστημα αξιολόγησης εντοπίζει τους εργαζόμενους που έχουν χαμηλή απόδοση	1	2	3	4	5
11	Το σύστημα αξιολόγησης συνδέεται με πρόσθετες αμοιβές	1	2	3	4	5
12	Το σύστημα αξιολόγησης συνδέεται ΟΥΣΙΑΣΤΙΚΑ με την εξέλιξη – καριέρα των εργαζομένων	1	2	3	4	5

Α/Α	ΠΕΡΙΓΡΑΦΗ	ΔΙΑΦΩΝΩ ΑΠΟΛΥΤΩΣ	ΔΙΑΦΩΝΩ	ΟΥΤΕ ΣΥΜΦΩΝΩ ΟΥΤΕ ΔΙΑΦΩΝΩ	ΣΥΜΦΩΝΩ	ΣΥΜΦΩΝΩ ΑΠΟΛΥΤΩΣ
Ε . ΤΑ ΚΡΙΤΗΡΙΑ ΑΞΙΟΛΟΓΗΣΗΣ						
1	Τα έντυπα της αξιολόγησης είναι απλά και κατανοητά	1	2	3	4	5
2	Οι εργαζόμενοι αξιολογούνται με ποσοτικά (στόχοι) κριτήρια	1	2	3	4	5
3	Οι εργαζόμενοι αξιολογούνται με ποιοτικά (ιδιότητες – ικανότητες) κριτήρια	1	2	3	4	5
4	Οι στόχοι αξιολόγησης συνδέονται με τη στρατηγική κατεύθυνση της επιχείρησης	1	2	3	4	5
5	Οι στόχοι είναι σχετικοί με τη δουλειά του καθένα	1	2	3	4	5
6	Οι στόχοι αξιολόγησης συμφωνούνται και δεν επιβάλλονται	1	2	3	4	5
7	Όταν διαφωνεί κανείς με τους στόχους που αναλαμβάνει μπορεί να το δηλώνει χωρίς συνέπειες	1	2	3	4	5
8	Όλοι είναι ενήμεροι για τους στόχους του τμήματος τους και της επιχείρησης συνολικά	1	2	3	4	5
9	Η νοοτροπία της επιχείρησης είναι να εργαζόμαστε υλοποιώντας στόχους και όχι απλά να διεκπεραιώνουμε εργασίες	1	2	3	4	5
10	Οι ιδιότητες-ικανότητες με τις οποίες αξιολογούνται οι εργαζόμενοι είναι ξεκάθαρες	1	2	3	4	5
11	Οι ιδιότητες- ικανότητες με τις οποίες αξιολογούνται οι εργαζόμενοι είναι σχετικές με το είδος της δουλειάς τους	1	2	3	4	5
12	Οι προϊστάμενοι καταλαβαίνουν και χρησιμοποιούν τα ποιοτικά κριτήρια (ιδιότητες) με τον ίδιο τρόπο	1	2	3	4	5
13	Τα έντυπα αξιολόγησης βοηθάνε τον αξιολογητή να κάνει σωστά τη δουλειά του	1	2	3	4	5

ΠΑΡΑΡΤΗΜΑ Β: Output από το SPSS για την Παραγοντική ανάλυση

Εικόνα 1: *Ο πίνακας με τα φορτία των παραγόντων μετά την περιστροφή.*

Rotated Component Matrix [a]

	Component			
	1	2	3	4
a1	,339	,089	,740	,091
a2	,215	,020	,731	,157
a4	,657	,146	,183	,211
a6	,476	,286	,416	,099
a7	,317	,132	,739	,114
a9	,490	,138	,381	,035
a10	,432	,253	,338	,171
a11	,477	,146	,579	,207
a12	,226	,212	,603	,129
a13	,464	,331	,333	,051
a15	,486	,137	,316	,196
b1	,458	,457	,113	,251
b2	,280	,731	,199	,071
b3	,131	,695	,047	,148
b4	,207	,784	,121	,166
b5	,381	,604	,097	,179
b6	,311	,736	,090	,163
b7	,327	,641	,131	,178
b8	,212	,793	,122	,198
b9	,365	,674	,199	,150
b10	,329	,649	,347	,160
c1	,783	,174	,182	,065
c2	,696	,338	,195	,120
c3	,552	,465	,278	,170
c4	,636	,375	,280	,100
c5	,560	,544	,133	,147
c6	,449	,599	,085	,089
c7	,164	,649	,359	,205
c8	,212	,589	,466	,137
c9	,548	,472	,187	,060
c10	,628	,331	,162	,131
c12	,503	,428	,281	,067
c13	,563	,428	,348	,317
c14	,574	,318	,438	,200
c15	,631	,382	,399	,141
c16	,611	,170	,397	,206
c17	,692	,209	,265	,017
c18	,616	,479	,240	,053
c19	,585	,486	,176	,139
c20	,524	,450	,427	,212
c21	,612	,301	,185	,174
c22	,538	,399	,271	,175
d1	,136	,121	,213	,734
d2	,142	,192	,142	,848
d3	,113	,183	,194	,852
d4	,019	,279	,188	,790
d5	,088	,279	,156	,847
d6	,215	,230	,121	,786
d7	,213	,130	,092	,857
d8	,246	,153	,151	,842
d9	,150	,012	,397	,753
d10	,108	,013	,404	,713
d11	-,024	,039	,503	,504
d12	-,053	,084	,572	,463
e2	,243	,114	,621	,206
e3	,328	,110	,485	,361
e4	,216	,169	,649	,336
e5	,213	,271	,645	,247
e6	,154	,400	,564	,139
e7	,293	,288	,451	,159
e8	,351	,557	,388	,109
e9	,251	,395	,541	,235
e10	,422	,203	,623	,260
e11	,206	,267	,534	,356
e12	,392	,325	,590	,219
e13	,279	,134	,651	,149

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 8 iterations.

Εικόνα 2: *Ο πίνακας με τις εταιρικότητες των μεταβλητών για το μοντέλο με 4 παράγοντες.*

Communalities

	Initial	Extraction
a1	1,000	,679
a2	1,000	,606
a4	1,000	,531
a6	1,000	,492
a7	1,000	,677
a9	1,000	,406
a10	1,000	,395
a11	1,000	,627
a12	1,000	,475
a13	1,000	,438
a15	1,000	,393
b1	1,000	,494
b2	1,000	,657
b3	1,000	,524
b4	1,000	,699
b5	1,000	,552
b6	1,000	,674
b7	1,000	,566
b8	1,000	,727
b9	1,000	,650
b10	1,000	,675
c1	1,000	,680
c2	1,000	,652
c3	1,000	,627
c4	1,000	,634
c5	1,000	,648
c6	1,000	,576
c7	1,000	,619
c8	1,000	,628
c9	1,000	,562
c10	1,000	,547
c12	1,000	,519
c13	1,000	,722
c14	1,000	,663
c15	1,000	,724
c16	1,000	,602
c17	1,000	,593
c18	1,000	,670
c19	1,000	,628
c20	1,000	,705
c21	1,000	,529
c22	1,000	,552
d1	1,000	,617
d2	1,000	,796
d3	1,000	,810
d4	1,000	,737
d5	1,000	,828
d6	1,000	,731
d7	1,000	,805
d8	1,000	,815
d9	1,000	,747
d10	1,000	,683
d11	1,000	,509
d12	1,000	,552
e2	1,000	,500
e3	1,000	,485
e4	1,000	,609
e5	1,000	,596
e6	1,000	,521
e7	1,000	,398
e8	1,000	,596
e9	1,000	,567
e10	1,000	,675
e11	1,000	,525
e12	1,000	,655
e13	1,000	,542

Extraction Method: Principal Component Analysis.

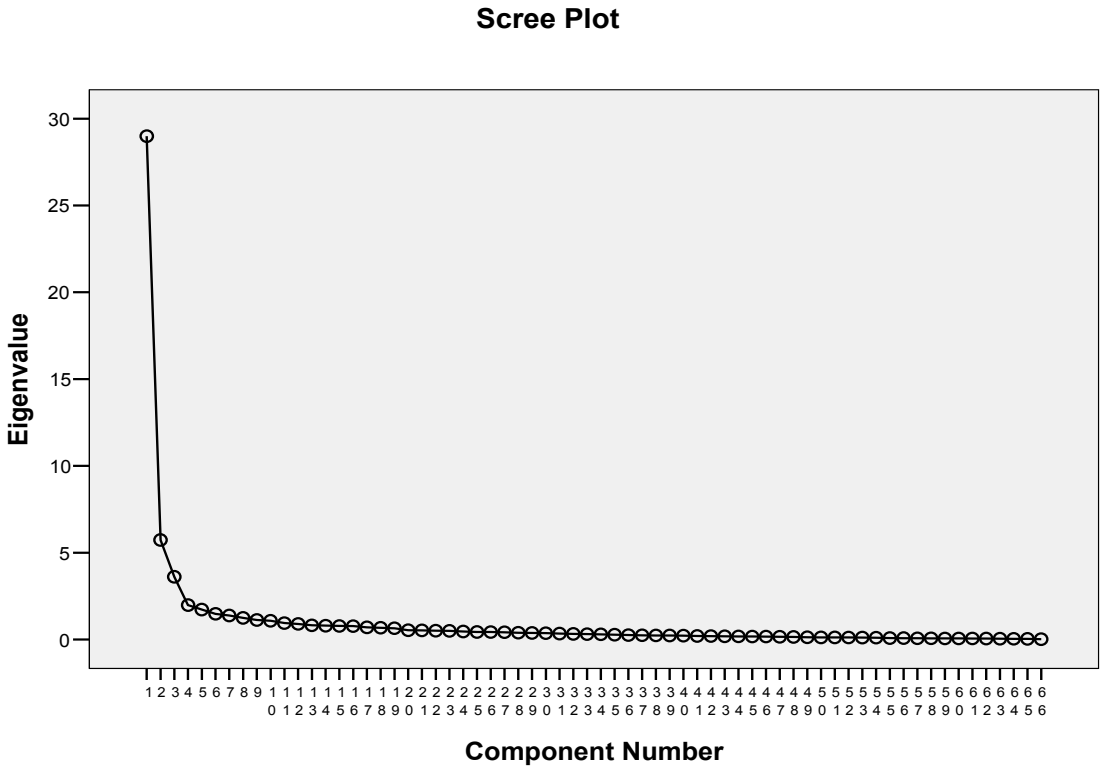
Εικόνα 3: *Ιδιοτιμές, ποσοστό της διακύμανσης που εξηγούν οι 4 παράγοντες πριν & μετά την περιστροφή.*

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	28,988	43,922	43,922	28,988	43,922	43,922	11,216	16,994	16,994
2	5,731	8,683	52,605	5,731	8,683	52,605	10,343	15,672	32,666
3	3,612	5,472	58,077	3,612	5,472	58,077	10,077	15,268	47,933
4	1,985	3,008	61,085	1,985	3,008	61,085	8,680	13,152	61,085
5	1,728	2,618	63,703						
6	1,481	2,244	65,948						
7	1,384	2,097	68,045						
8	1,249	1,893	69,937						
9	1,127	1,708	71,645						
10	1,079	1,634	73,279						
11	,948	1,436	74,715						
12	,900	1,363	76,078						
13	,819	1,240	77,318						
14	,799	1,211	78,529						
15	,784	1,188	79,718						
16	,775	1,174	80,892						
17	,702	1,063	81,955						
18	,676	1,025	82,979						
19	,654	,991	83,970						
20	,536	,813	84,783						
21	,530	,803	85,586						
22	,512	,776	86,361						
23	,498	,754	87,115						
24	,463	,701	87,816						
25	,436	,661	88,477						
26	,432	,654	89,131						
27	,419	,636	89,767						
28	,393	,596	90,362						
29	,380	,575	90,938						
30	,368	,557	91,495						
31	,344	,521	92,016						
32	,324	,491	92,507						
33	,314	,475	92,982						
34	,301	,456	93,438						
35	,280	,424	93,863						
36	,258	,391	94,253						
37	,244	,370	94,624						
38	,237	,359	94,983						
39	,229	,347	95,330						
40	,221	,335	95,665						
41	,207	,314	95,979						
42	,200	,303	96,282						
43	,188	,285	96,567						
44	,180	,273	96,841						
45	,174	,263	97,104						
46	,172	,261	97,365						
47	,158	,240	97,605						
48	,151	,229	97,834						
49	,133	,202	98,035						
50	,123	,187	98,222						
51	,122	,185	98,407						
52	,119	,180	98,587						
53	,111	,167	98,755						
54	,105	,160	98,915						
55	,087	,132	99,047						
56	,083	,126	99,173						
57	,075	,113	99,286						
58	,072	,108	99,395						
59	,065	,099	99,493						
60	,064	,097	99,590						
61	,061	,092	99,682						
62	,058	,087	99,769						
63	,049	,074	99,843						
64	,042	,063	99,906						
65	,037	,056	99,962						
66	,025	,038	100,000						

Extraction Method: Principal Component Analysis.

Εικόνα 4: *Το scree plot για τα δεδομένα.*



Εικόνα 5 : *Πίνακας με τους συντελεστές των μεταβλητών των factor σκορ.*

Component Score Coefficient Matrix

	Component			
	1	2	3	4
a1	-,008	-,044	,132	-,046
a2	-,032	-,043	,140	-,031
a4	,151	-,085	-,059	,026
a6	,041	-,015	,031	-,023
a7	-,022	-,030	,132	-,043
a9	,074	-,052	,027	-,025
a10	,045	-,018	,012	-,002
a11	,046	-,056	,063	-,011
a12	-,045	,006	,108	-,034
a13	,040	,002	,016	-,026
a15	,084	-,058	-,002	,010
b1	,048	,028	-,056	,024
b2	-,066	,136	,002	-,030
b3	-,083	,150	-,023	-,004
b4	-,085	,159	-,016	-,010
b5	,001	,082	-,045	,005
b6	-,042	,129	-,037	-,004
b7	-,026	,101	-,029	-,001
b8	-,085	,159	-,019	-,004
b9	-,029	,104	-,015	-,012
b10	-,055	,102	,026	-,023
c1	,182	-,093	-,063	,000
c2	,127	-,038	-,054	,001
c3	,052	,017	-,020	-,004
c4	,091	-,018	-,024	-,012
c5	,060	,037	-,056	,001
c6	,022	,073	-,049	-,010
c7	-,105	,128	,046	-,018
c8	-,096	,107	,073	-,036
c9	,060	,024	-,033	-,017
c10	,112	-,029	-,055	,005
c12	,042	,020	-,004	-,022
c13	,055	-,001	-,016	,019
c14	,064	-,027	,015	-,006
c15	,074	-,018	,003	-,015
c16	,105	-,072	,001	,005
c17	,137	-,067	-,027	-,018
c18	,072	,014	-,028	-,022
c19	,071	,018	-,047	-,001
c20	,028	,015	,017	-,009
c21	,110	-,035	-,050	,012
c22	,060	,002	-,020	,000
d1	,003	-,024	-,028	,114
d2	,003	-,011	-,057	,138
d3	-,011	-,009	-,041	,134
d4	-,055	,034	-,025	,118
d5	-,030	,020	-,047	,132
d6	,020	-,011	-,067	,128
d7	,042	-,039	-,079	,147
d8	,040	-,039	-,067	,139
d9	,002	-,057	,015	,106
d10	-,012	-,048	,025	,098
d11	-,072	-,010	,085	,048
d12	-,098	,007	,109	,032
e2	-,024	-,025	,104	-,017
e3	,022	-,046	,046	,025
e4	-,044	-,012	,103	,002
e5	-,063	,019	,110	-,018
e6	-,093	,067	,106	-,037
e7	-,018	,014	,059	-,017
e8	-,039	,076	,038	-,031
e9	-,059	,045	,079	-,016
e10	,015	-,035	,076	-,008
e11	-,048	,015	,074	,012
e12	-,011	,004	,075	-,018
e13	-,021	-,024	,112	-,030

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

ΠΑΡΑΡΤΗΜΑ Γ: Αποτελέσματα από την πολλαπλή ανάλυση αντιστοιχιών για το ερωτηματολόγιο

Principal inertias (eigenvalues):

dim	value	%	cum%	scree plot
1	0.173848	40.8	40.8	************************
2	0.107314	25.2	65.9	***************
3	0.035238	8.3	74.2	*****
4	0.021169	5.0	79.1	***
5	0.010823	2.5	81.7	**
6	0.007076	1.7	83.3	*
7	0.003437	0.8	84.2	
8	0.001367	0.3	84.5	
9	0.000730	0.2	84.6	
10	0.000397	0.1	84.7	
11	0.000114	0.0	84.8	
12	00000000	0.0	84.8	
	--------	-----		
Total:	0.426503	<NA>		

Αποτελέσματα για τις στήλες:

Columns:*

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	pst.A-YPOG	14	453	8	-99	79	0	-192	374	2
2	pst.B-YPOG	40	716	4	64	179	0	-98	537	1
3	pst.DIEY8INTHS	8	833	6	330	636	2	163	197	1
4	pst.CIEUF	10	187	5	133	173	0	-34	14	0
5	pst.YPALL	52	1532	4	-100	606	1	109	926	2
6	b1.1	16	2217	31	-960	2055	35	-239	162	3
7	b1.2	39	1523	10	7	1	0	267	1522	8
8	b1.3	27	26	5	-25	13	0	-21	13	0
9	b1.4	37	1488	10	305	1479	8	-20	8	0
10	b1.5	7	2273	15	675	864	7	-763	1408	12
11	c2.1	14	1856	16	-597	1423	12	-292	433	4
12	c2.2	50	1994	8	-103	277	1	227	1717	8
13	c2.3	25	75	3	-12	5	0	38	70	0
14	c2.4	27	1024	9	275	1024	5	7	1	0
15	c2.5	8	2414	25	737	785	11	-942	1629	23
16	e2.1	20	2261	30	-710	1498	24	-449	762	12
17	e2.2	35	917	16	-61	37	0	267	880	8
18	e2.3	40	980	14	39	19	0	249	962	7
19	e2.4	27	1995	20	481	1359	15	-292	636	7
20	e2.5	3	1088	20	794	348	4	-1027	740	8
21	e3.1	13	2114	41	-1103	1636	37	-529	478	11
22	e3.2	22	621	13	-217	359	2	165	262	2
23	e3.3	36	1480	17	-42	16	0	352	1464	14
24	e3.4	48	1834	16	358	1636	15	-111	198	2
25	e3.5	6	1285	20	505	320	4	-777	965	11
26	d7.1	8	2146	55	-1724	1777	54	-696	369	11
27	d7.2	26	1688	19	-313	578	6	385	1110	12
28	d7.3	28	816	9	-4	0	0	221	816	4
29	d7.4	44	1045	18	301	962	9	78	83	1
30	d7.5	20	2172	33	414	460	8	-708	1712	31
31	d5.1	10	2128	52	-1496	1878	54	-484	250	7

```
32 |           d5.2 |   22 1396   16 |  -231  312    3 |   382 1085   10 |
33 |           d5.3 |   28 1741   13 |   -72   48    0 |   382 1693   12 |
34 |           d5.4 |   45 1061   17 |   302 1052   10 |    26   10    0 |
               name   mass  qlt  inr     k=1  cor  ctr     k=2  cor  ctr
35 |           d5.5 |   20 2061   39 |   428  416    9 |  -753 1644   35 |
36 |           d2.1 |    8 2101   58 | -1739 1719   55 |  -726  382   12 |
37 |           d2.2 |   16 1360   19 |  -393  562    6 |   415  798    8 |
38 |           d2.3 |   22 1390   11 |  -112  105    1 |   346 1285    8 |
39 |           d2.4 |   55 1190   16 |   253  924    8 |   120  265    2 |
40 |           d2.5 |   25 2034   31 |   325  375    6 |  -606 1660   28 |
41 |   (*)yrs.11-25 |  489   63   62 |   -64   58 <NA> |    19    5 <NA> |
42 |     (*)yrs.<10 |  484  102   62 |    72   73 <NA> |   -45   29 <NA> |
43 |     (*)yrs.>25 |   27  109  103 |  -100    5 <NA> |   469  105 <NA> |
```

∗Οι στήλες «qlt», «inr», «k=1, k=2», «cor» και «ctr», αναφέρονται κατά σειρά στην «Ποιότητα» της παρουσίασης, στην αδράνεια που ερμηνεύει κάθε σημείο, στις κύριες συντεταγμένες (principal coordinates) των σημείων για τους 2 πρώτους άξονες, στις σχετικές συνεισφορές και στις απόλυτες συνεισφορές. Οι ποσότητες αυτές είναι πολλαπλασιασμένες επί 1000 είτε αναφέρονται σε permills (thousandths).

Οι ιδιόμορφες τιμές:

```
[1] 0.416951169 0.327588402 0.187716910 0.145496335 0.104032464
[6] 0.084119834 0.058623571 0.036972249 0.027024497 0.019922237
[11]0.010663453 0.000639284
```

Οι μάζες των στηλών:

```
pos.A-YPOG    pos.B-YPOG    pos.B.DIEY8     pos.CIEUF     pos.YPALL
0.014326346   0.040345518   0.008427262   0.010112715   0.052038344
       b1.1          b1.2          b1.3          b1.4          b1.5
0.016011798   0.038554725   0.026861898   0.037079954   0.006741810
       c2.1          c2.2          c2.3          c2.4          c2.5
0.014326346   0.049510165   0.025281787   0.026967239   0.008427262
       e2.1          e2.2          e2.3          e2.4          e2.5
0.020225429   0.035394501   0.039608132   0.026861898   0.002528179
       e3.1          e3.2          e3.3          e3.4          e3.5
0.012640893   0.021910882   0.036237227   0.047930054   0.005899084
       d7.1          d7.2          d7.3          d7.4          d7.5
0.007584536   0.026124513   0.027809965   0.043505741   0.020225429
       d5.1          d5.2          d5.3          d5.4          d5.5
0.010112715   0.021910882   0.027809965   0.045191194   0.020225429
       d2.1          d2.2          d2.3          d2.4          d2.5
0.007584536   0.016011798   0.021910882   0.054566523   0.025176446
(*)years.11-25  (*)yrs.<10   (*)yrs.>25
0.489486964   0.483599664   0.026913373
```

χ^2 απόσταση κάθε κατηγόριας από το μέσο ή κεντροειδές:

```
 pos.A-YPOG    pos.B-YPOG    pos.B.DIEY8     pos.CIEUF     pos.YPALL
  1.1088382     0.5564281     1.4452013     1.2791392     0.4578752
       b1.1          b1.2          b1.3          b1.4          b1.5
  1.3387780     0.6368484     0.7426158     0.6542723     1.8161000
       c2.1          c2.2          c2.3          c2.4          c2.5
  1.2209209     0.5205891     0.7399346     0.7807876     1.7843529
       e2.1          e2.2          e2.3          e2.4          e2.5
  1.1633576     0.7300039     0.6646366     0.9019932     3.1405468
```

```
      e3.1          e3.2          e3.3          e3.4          e3.5
 1.6343380     0.9295292     0.7298884     0.6047956     2.0485656
      d7.1          d7.2          d7.3          d7.4          d7.5
 2.3331401     0.9105793     0.7728321     0.6567061     1.1951829
      d5.1          d5.2          d5.3          d5.4          d5.5
 1.9797972     0.9737606     0.8164674     0.6343700     1.2528203
      d2.1          d2.2          d2.3          d2.4          d2.5
 2.3717755     1.1943321     0.9152206     0.5540127     1.0426707
 (*)years.11-25  (*)yrs.<10   (*)yrs.>25
 0.2653483       0.2659978     1.4507078
```

Αδράνεια των στηλών:

```
pos.A-YPOG    pos.B-YPOG    pos.B.DIEY8     pos.CIEUF     pos.YPALL
0.004320402   0.002182462   0.003462451   0.002497695   0.002045273
       b1.1          b1.2          b1.3          b1.4          b1.5
0.017228393   0.005521381   0.002910271   0.005580449   0.008518506
       c2.1          c2.2          c2.3          c2.4          c2.5
0.008595804   0.004550480   0.001573809   0.004783912   0.014011537
       e2.1          e2.2          e2.3          e2.4          e2.5
0.016315761   0.008755709   0.007797240   0.010956976   0.011001755
       e3.1          e3.2          e3.3          e3.4          e3.5
0.022536806   0.006909030   0.009382364   0.009026289   0.011278405
       d7.1          d7.2          d7.3          d7.4          d7.5
0.030411216   0.010630661   0.005098614   0.009800723   0.018050788
       d5.1          d5.2          d5.3          d5.4          d5.5
0.028887833   0.009017109   0.007302739   0.009382834   0.021312202
       d2.1          d2.2          d2.3          d2.4          d2.5
0.031986858   0.010532768   0.006248052   0.009078767   0.017020500
(*)years.11-25  (*)yrs.<10   (*)yrs.>25
0.034464638   0.034217013   0.056640625
```

Έλεγχος της σημαντικότητας των κατηγοριών των μεταβλητών στους πρώτους άξονες[**]

```
                Dim 1       Dim 2       Dim 3
11-25         1.52935065 -0.5354075 -3.4368166
<10          -1.69914061  1.2500864  2.5256914
>25           0.43757213 -2.2396173  2.6602846
A-YPOG        0.78667268  1.6484014 -1.0546514
B-YPOG       -0.86847694  1.6328985 -0.3067405
B.DIEY8INTHS -1.66499384 -1.2298419 -2.6768145
CIEUF        -0.71066550 -0.2359186  2.9522061
YPALL         1.52628702 -1.8680509  0.6487992
b1_1          7.41749093  2.2909603  0.8163823
b1_2         -0.02010262 -4.2130582  1.4279562
b1_3          0.31490393  0.2529822 -0.8559024
b1_4         -4.03464202  0.2334199 -1.8780639
b1_5         -3.40663357  4.2845006  1.1122030
c2_1          4.31392710  2.6164475  2.2826289
c2_2          1.84781065 -4.4240895  0.8830431
c2_3          0.16768895 -0.3382992 -3.0240064
c2_4         -3.01898314 -0.1837052 -0.9918812
c2_5         -4.16405106  5.9109020  1.9354584
e2_1          6.35816814  4.6848066 -1.3740651
```

e2_2	0.83527473	-4.0528028	5.2781150
e2_3	-0.47838063	-3.9126699	-3.3932865
e2_4	-5.19092522	3.5992395	-1.4774120
e2_5	-2.47747093	3.4718359	2.6997604
e3_1	7.54059491	4.3350912	-2.7485204
e3_2	2.06323157	-1.8215856	5.0041209
e3_3	0.66812717	-5.3290107	-1.1163787
e3_4	-5.80100070	2.0120773	-2.2444697
e3_5	-2.48383051	4.1138365	2.7843305
d7_1	8.87131733	4.3957234	-2.5177847
d7_2	3.33826942	-4.7250337	5.2114328
d7_3	0.03759994	-2.7192711	0.4099594
d7_4	-4.40101228	-1.4233349	-6.1551260
d7_5	-3.81157511	7.2697721	3.3235782
d5_1	8.95327171	3.6335998	-1.5491225
d5_2	2.21092285	-4.3430035	5.7101828
d5_3	0.84479532	-4.6679887	-0.5083461
d5_4	-4.53291670	-0.5422480	-6.7608058
d5_5	-3.97791063	7.7605346	4.5935830
d2_1	8.92560855	4.5914969	-2.6609288
d2_2	3.11874217	-3.8943174	6.0053028
d2_3	0.98855596	-3.5633720	1.4892990
d2_4	-4.38096356	-2.6442027	-7.0698282
d2_5	-3.45348405	7.1389937	3.8504488

** Υπολογίζονται βάσει τη σχέση: $\text{coord}_{jk}\sqrt{n_j\frac{n-1}{n-n_j}}$, όπου χρειαζόμαστε την συντεταγμένη της j κατηγορίας στον k άξονα, n είναι το μέγεθος του δείγματος και n_j είναι ο αριθμός των παρατηρήσεων στη j κατηγορία. Η ποσότητα αυτή προσεγγιστικά ακολουθεί την τυπική κανονική κατανομή, οπότε για να δούμε ποια κατηγορία είναι σημαντική σε ποιον άξονα ζητάμε σχετικά μεγάλες τιμές (σε απόλυτη τιμή) για την συνάρτηση αυτή.

Βιβλιογραφία

Ελληνική

Καρλής, Δ. (2005). *Πολυμεταβλητή Στατιστική Ανάλυση*, Εκδόσεις Σταμούλης, Αθήνα.

Κατέρη, Μ. (2006). *Ανάλυση Διακριτών Δεδομένων*, Πανεπιστημιακές Παραδόσεις, Πανεπιστήμιο Πειραιά.

Κούτρας, Μ. (2005). *Εφαρμοσμένη Πολυμεταβλητή Ανάλυση*: *Ανάλυση κατά Συστάδες*, Πανεπιστημιακές Παραδόσεις, Πανεπιστήμιο Πειραιά.

Μπόρα-Σέντα, Ε. και Μωϋσιάδης, Χ. (1997). *Εφαρμοσμένη Στατιστική*, Εκδόσεις ΖΗΤΗ, Θεσσαλονίκη.

Χατζηκωνσταντινίδης, Ε (2005). *Γενικευμένα Γραμμικά Μοντέλα*, Πανεπιστημιακές Παραδόσεις, Πανεπιστήμιο Πειραιά.

Ξένη

Adejumo, A.O., Heumann, C. and Toutenburg, H. (2004). A review of agreement measure as a subset of association measure between raters, *Department of Statistics*, *Munich*, 1-43.

Agresti, A. (1983). A survey of strategies for modeling cross-classifications having ordinal variables, *Journal of the American Statistical Association*, **78**, 184-198.

Agresti, A., Chuang, C., and Kezouh, A. (1987). Order-restricted score parameters in association models for contingency tables, *Journal of the American Statistical Association*, **82**, 619-623.

Agresti, A. (2001). *Categorical Data Analysis*, 2nd ed., John Wiley, New York.

Agresti, A. and Liu, I. (2001). Strategies for modeling a categorical variable allowing multiple category choices, *Sociological Methods and Research*, **29**, 403-434.

Alba, D.R. (1987). Interpreting the parameters of log-linear models, *Sociological Methods and Research*, **16**, 45-77.

Aldenderfer, M.S. and Blashfield, R.K. (1978). Computer programs for performing hierarchical cluster analysis, *Applied Psychological Measurement*, **2**, 403-411.

Anderberg, M.R. (1973). *Cluster Analysis for Applications*, Academic Press, New York.

Anderson, C.J. (1996). The analysis of three-way contingency tables by three-mode association models, *Psychometrika*, **61**, 465-483.

Bartholomew, D.J., Steele, F., Moustaki, I. and Galbraith, J.I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists*, Chapman & Hall/CRC.

Becker, M.P. (1989). Models for the analysis of association in multivariate contingency tables, *Journal of the American Statistical Association*, **84**, 1014-1019.

Becker, M.P. (1990). Algorithm AS 253: Maximum likelihood estimation of the RC(M) association model, *Applied Statistics*, **39**, 152-167.

Becker, M.P. and Clogg, C.C. (1989). Analysis of sets of two-way contingency tables using association models, *Journal of the American Statistical Association*, **84b**, 142-151.

Beckstead, J.W. (2002). Using hierarchical cluster analysis in nursing research, *Western Journal of Nursing Research*, **24**, 307-319.

Beh, E.J. (1998). A comparative study of scores for correspondence analysis with ordered categories, *Biometrical Journal*, **40**, 413-429.

Beh, E.J. (2007). Simple correspondence analysis of nominal-ordinal contingency tables, *Journal of Applied Mathematics and Decision Sciences*, 2008, 1-17.

Beh, E.J. and Davy, P.J. (2004). A non-iterative alternative to ordinal log-linear models, *Journal of Applied Mathematics and Decision Sciences*, **8**, 67-86.

Blasius, J. and Greenacre, M. (1998). *Visualization of Categorical Data,* Academic Press, San Diego.

Bradburn, N., Sudman, S. and Wansink, B. (2004). *Asking Questions*: *The Definite Guide to Questionnaire Design – for Market Research, Political Polls*, *and Social and Health Questionnaires*, Jossey – Bass, San Francisco.

Breen, R. (2008). Statistical models of association for comparing cross-classifications, *Sociological Methods and Research*, **36**, 442-461.

Carlier, A. and Kroonenberg, P.M. (1996). Decompositions and biplots in three-way correspondence analysis, *Psychometrika*, **61**, 355-373.

Chambers, J.M. and Hastie, T.J. (1996). *Statistical Models in S*, Chapman & Hall, London.

Cheung, M.L. and Chan, W. (2005). Classifying correlation matrices into relatively homogeneous subgroups: A cluster analytic approach, *Educational and Psychological Measurement*, **65**, 954-979.

Choulakian, V. (1988). Exploratory analysis of contingency tables by loglinear formulation and generalizations of correspondence analysis, *Psychometrika*, **53**, 235-250.

Clausen, S.E. (1998). *Applied Correspondence Analysis*: *An introduction*, Sage Publications, California.

Clogg, C.C. (1982). Some models for the analysis of association in multiway cross-classifications having ordered categories, *Journal of the American Statistical Association*, **77**, 803-815.

Clogg, C.C. (1982). Using association models in sociological research: Some examples, *The American Journal of Sociology*, **88**, 114-134.

Cox, T.F. and Cox, M.A.A. (2001). *Multidimensional Scaling*, 2nd ed., Chapman & Hall/CRC.

Dilts, D., Khamala, J. and Plotkin, A. (1995). Using cluster analysis for medical resource decision making, *Medical Decision Making*, **15**, 333-347.

Etzioni, R.D., Fienberg, S.E., Gilula, Z. and Haberman, S.J. (1994). Statistical models for the analysis of ordered categorical data in public health and medical research, *Statistical Methods in Medical Research*, **3**, 179-204.

Everitt, B.S. (2005). *An R and S-Plus Companion to Multivariate Analysis*, Springer-Verlag, London.

Faust, K. and Wasserman, S. (1993). Correlation and association models for studying measurements on ordinal relations, *Sociological Methodology*, **23**, 177-215.

Galindo-Garre, F. and Vermunt, J.K. (2004). The order-restricted association model: Two estimation algorithms and issues in testing, *Psychometrika*, **69**, 641-654.

Gilula, Z. (1986). Grouping and association in contingency tables: An exploratory canonical correlation approach, *Journal of the American Statistical Association*, **81**, 773-779.

Goodman, L.A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories, *Journal of the American Statistical Association*, **76**, 320-334.

Goodman, L.A. (1981c). Criteria for determining whether certain categories in a cross-classification table should be combined, with special reference to occupational categories in an occupational mobility table, *The American Journal of Sociology*, **87**, 612-650.

Goodman, L.A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries, *The Annals of Statistics*, **13**, 10-69.

Goodman, L.A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables, *International Statistical Review*, **54**, 243-270.

Goodman, L.A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data, *Journal of the American Statistical Association*, **86**, 1085-1111.

Goodman, L.A. (2002). Contributions to the statistical analysis of contingency tables: Notes on quasi-symmetry, quasi-independence, log-linear models, log-bilinear models, and correspondence analysis models, *Annales de la faculté des sciences de Toulouse Sér. 6*, **11**, 525-540.

Greenacre, M. and Hastie, T. (1987). The geometric representation of correspondence analysis, *Journal of the American Statistical Association*, **82**, 437-447.

Greenacre, M. (1994). *Theory and Applications of Correspondence Analysis*, Academic Press, New York.

Greenacre, M. and Pardo, R. (2006). Subset correspondence analysis: Visualizing relationships among a selected set of response categories from a questionnaire survey, *Sociological Methods and Research*, **35**, 193-218.

Harman, H.H. (1976). *Modern Factor Analysis*, 3rd ed., The University of Chicago Press, Chicago and London.

Hardle, W. and Simar, L. (2003). *Applied Multivariate Statistical Analysis*, Springer-Verlag, Heidelberg.

Hartigan, J.A. (1972). Direct clustering of a data matrix, *Journal of the American Statistical Association*, **67**, 123-129.

Hildebrand, D.K., Laing, J.D. and Rosenthal, H. (1977). *Analysis of Ordinal Data*, Sage Pub, London.

Hunter, M.A. and Takane, Y. (2002). Constrained principal component analysis: Various applications, *Journal of Education and Behavioral Statistics*, **27**, 105-145.

Ishii-Kuntz, M. and Coltrane, S. (1992). Predicting the sharing of household labor: Are parenting and housework distinct ?, *Sociological Perspectives*, **35**, 629-647.

Jackson, J.E. (1991). *A User's Guide to Principal Components*, John Wiley & Sons, Chicago.

Jobson, J.D. (1992). *Applied Multivariate Data Analysis Vol.2*: *Categorical and Multivariate Methods*, Springer, New York.

Joliffe, IT and Morgan, Bjt (1992). Principal component analysis and exploratory factor analysis, *Statistical Methods in Medical Research*, **1**, 69-95.

Jurowski, C. and Reich, A.Z. (2000). An explanation and illustration of cluster analysis for identifying hospitality market segments, *Journal of Hospitality & Tourism Research*, **24**, 67-91.

Kamo, Y. (1988). Determinants of household division of labor: Resources, power and ideology, *Journal of Family Issues*, **9**, 177-200.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data*: *An Introduction to Cluster Analysis*, John Wiley & Sons, New York.

Knalf, G.J. and Grey, M. (2007). Factor analysis model through likelihood cross-validation, *Statistical Methods in Medical Research*, **16**, 77-102.

Krause, A. and Olson, M. (2000). *The Basics of S and S-Plus*, 2nd ed., Springer, New York.

Kuanli, A.H., Guynes, C.S. and Pavur, R.J. (1996). *Introduction to Business Statistics*: *A Computer Integrated*, *Data Analysis Approach*, 4th ed., West Publishing Company, St.Paul.

Lee, H B. and MacQueen, J.B. (1980). A K-Means cluster analysis computer program with cross-tabulations and next-nearest-neighbor analysis, *Educational and Psychological Measurement*, **40**, 133-138.

Lee, S. and Xu, L. (2003). Case-deletion diagnostics for factor analysis models with continuous and ordinal categorical data, *Sociological Methods and Research*, **31**, 389-419.

Leech, N.L., Barrett, K.C. and Morgan, G.A. (2005). *SPSS for Intermediate Statistics*: *Use and Interpretation*, 2nd ed., Lawrence Erlbaum Associates, London.

Lian, B. and Young, W.C. (2001). On the measurement of concordance among variables and its application, *Journal of Educational and Behavioral Statistics*, **26**, 431-442.

Liou, M., Cheng, P.E. and Li, M.Y. (2001). Estimating comparable scores using surrogate variables, *Applied Psychological Measurement*, **25**, 197-207.

McEvoy, P. and Richards, D. (2001). Using log-linear models to analyze categorical data, *Nursing Time Research*, **6**, 867-875.

McQuitty, L.L. and Clark, J.A. (1968). Clusters from iterative, intercolumnar correlation analysis, *Educational and Psychological Measurement*, **28**, 211-238.

Milligan, G.W. and Cooper, M. (1987). Methodology review: Clustering methods, *Applied Psychological Measurement*, **11**, 329-354.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R*, Chapman & Hall/CRC.

Murrell, P. (2006). *R Graphics*, Chapman & Hall/CRC.

Page, W.F. (1977). Interpretation of Goodman's log-linear model effects: An odds ratio approach, *Sociological Methods and Research*, **5**, 419-435.

Pannekoek, J. (1985). Log-multiplicative models for multiway tables, *Sociological Methods and Research*, **14**, 137-153.

Powers, D.A. and Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis*, Academic Press, New York.

Rencher, A.C. (2002). *Methods of Multivariate Analysis*, 2nd ed., Wiley & Sons, New York.

Roux, B.L. and Rouanet, H.G. (2004). *Geometric Data Analysis*: *From Correspondence Analysis to Structured Data Analysis*, Kluwer Academic, New York.

Seo, J. and Gordish-Dressman, H. (2007). Exploratory data analysis with categorical variables: An improved rank-by-feature framework and a case study, *International Journal of Human-Computer Interaction*, **23**, 287-314.

Siciliano, R. and Mooijaart, Ab (1997). Three-factor association models for three-way contingency tables, *Computational Statistics and Data Analysis*, **24**, 337-356.

Simonoff, J.S. (2003). *Analyzing Categorical Data*, Springer-Verlag, New York.

Timm, N.H. (2002). *Applied Multivariate Analysis*, Springer-Verlag, New York.

Upton, Graham J.G. and Fingleton, B. (1989). *Spatial Data Analysis by Example Vol.2*: *Categorical and Directional Data*, John Wiley & Sons, Chichester.

Urban, G.D. and McDaniel, M.A. (1990). Factor and cluster analyses of the special assignment battery, *Educational and Psychological Measurement*, **50**, 663-671.

Van der Heijden, P.J. and de Leeuw, J. (1985). Correspondence analysis used complementary to log-linear analysis, *Psychometrika*, **50**, 429-447.

Van der Heijden, P.J., de Falguerolles, A. and de Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis, *Applied Statistics*, **38**, 249-292.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*, 4th ed., Springer-Verlag, New York.

Wong, R.S. (2001). Multidimensional association models: A multilinear approach, *Sociological Methods and Research*, **30**, 197-240.

Ziberna, A., Kejzar, N. and Golob, P. (2004). A comparison of different approaches to hierarchical clustering of ordinal data, *Metodoloski zvezki*, **1**, 57-73.