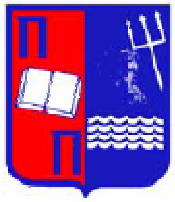
Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ



Τ Μ Η Μ Α Σ Τ Α Τ Ι Σ Τ Ι Κ Η Σ

Κ Α Ι Α Σ Φ Α Λ Ι Σ Τ Ι Κ Η Σ Ε Π Ι Σ Τ Η Μ Η Σ

Μ Ε Τ Α Π Τ Υ Χ Ι Α Κ Ο Π Ρ Ο Γ Ρ Α Μ Μ Α

Σ Π Ο Υ Δ Ω Ν

Σ Τ Η Ν Ε Φ Α Ρ Μ Ο Σ Μ Ε Ν Η Σ Τ Α Τ Ι Σ Τ Ι Κ Η

ΑΝΑΛΥΣΗ ΠΟΛΥΔΙΑΣΤΑΤΩΝ ΠΙΝΑΚΩΝ ΣΥΝΑΦΕΙΑΣ ΜΕ ΔΙΑΤΑΞΙΜΕΣ ΜΕΤΑΒΛΗΤΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Μαρία Α. Παπαδοπούλου

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
Ιούνιος 2007

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:

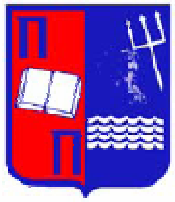
-…….(Επιβλέπων)

-…….

-…….

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

ANALYSIS OF MULTIWAY CONTINGENCY TABLES WITH ORDINAL VARIABLES OF CLASSIFICATION

By
Maria A. Papadopoulou

MSc Dissertation

Submitted to the Department of Statistics and Insurance Science of the University in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
June 2007

Στον Στέφανο, τη Νίκη, τους γονείς μου,

τη γιαγιά μου και τον παππού μου

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς την κα Κατέρη Μαρία, Επίκουρη Καθηγήτρια του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς για την πολύ σημαντική καθοδήγηση, βοήθεια, υποστήριξη και υπομονή κατά τη διάρκεια υλοποίησης της παρούσας εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τον κο Πολίτη Κωνσταντίνο, Επίκουρο Καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς και τον κο Πιτσέλη Γεώργιο, Επίκουρο Καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, για την συμμετοχή τους στην τριμελή επιτροπή.

Τέλος, θα ήθελα να ευχαριστήσω όλους τους φίλους μου για την αμέριστη συμπαράστασή τους.

Παπαδοπούλου Μαρία

Πειραιάς

Ιούνιος 2007

Περίληψη

Μετά το τέλος του 2^{ου} παγκόσμιου πολέμου, ξεκίνησε η ανάπτυξη των μεθόδων ανάλυσης δεδομένων. Στη δεκαετία του ΄60, άρχισε η μελέτη των (διατάξιμων) κατηγορικών δεδομένων σε πολυδιάστατους πίνακες συνάφειας και συγκεκριμένα η προσπάθεια μοντελοποίησής των, παραμετρικά και απαραμετρικά. Πρωτεργάτης σε αυτή την προσπάθεια θεωρείται ο Leo Goodman. Σημαντική εργασία έχει γίνει επίσης και στο πεδίο των γραφικών αναπαραστάσεων της δομής και των σχέσεων που διέπουν τα κατηγορικά δεδομένα. Στην παρούσα εργασία, θα ασχοληθούμε με την ανάλυση (διατάξιμων) κατηγορικών δεδομένων για πολυδιάστατους πίνακες συνάφειας.

Στο πρώτο κεφάλαιο, παρατίθεται μια ιστορική αναδρομή στην ανάλυση κατηγορικών δεδομένων και στο δεύτερο κεφάλαιο παρουσιάζονται γραφικές αναπαραστάσεις των κατηγορικών δεδομένων με την βοήθεια των μωσαϊκών αλλά και των γραφικών αναπαραστάσεων συνάφειας. Στο τρίτο κεφάλαιο παρουσιάζονται τα λογαριθμογραμμικά μοντέλα συνάφειας για τρεις μεταβλητές, καθώς και ένας αλγόριθμος επιλογής βέλτιστου μοντέλου. Στο τέταρτο κεφάλαιο ορίζονται τα γραφήματα συνάφειας και δίνεται μια σύντομη αναφορά στα γραφικά μοντέλα. Στο πέμπτο κεφάλαιο παρουσιάζονται τα μοντέλα λανθανουσών μεταβλητών και στο έκτο κεφάλαιο η ανάλυση ομοιογένειας (ή πολλαπλή ανάλυση αντιστοιχιών) και δίνεται μια εφαρμογή αυτής βήμα προς βήμα.

Abstract

The development of methods of data analysis had started after the end of the second world war. The beginning of the study of (ordinal) categorical data and especially an attempt of modeling them, parametrically and non parametrically, is placed at sixties. Leo Goodman is considered as a main contributor toward this direction. Serious effort has been made in the field of graphical representations of the structure and the relationships underlying categorical data. In this thesis, we focus our interest on the analysis of (ordinal) categorical data for multiple-way contingency tables.

In the first chapter, a history of analysis of categorical data is represented and in the second chapter, we build graphical representations for categorical data, i.e. mosaics and association plots. In the third chapter, three variables loglinear association models are presented and also, an algorithm for model selection. In the forth chapter, we define association graphs and a brief report for graphical models is given. In the fifth chapter, we discuss latent variables models and in the sixth chapter, we present homogeneity analysis (or multiple correspondence analysis) and a detailed example.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1^{ο}

Ιστορική Αναφορά στην Ανάλυση Κατηγορικών Δεδομένων

1.0 Εισαγωγή

Η ιστορική αναφορά στην ανάλυση κατηγορικών δεδομένων (*Categorical Data Analysis-CDA*) θεωρείται απαραίτητη διότι είναι πολύ σημαντικό για έναν ερευνητή να γνωρίσει και να κατανοήσει την εξέλιξη των μεθόδων ανάλυσης των κατηγορικών δεδομένων. Κατά αυτόν τον τρόπο, θα συνειδητοποιήσει πώς έφτασαν στα χέρια του αυτές οι μέθοδοι ανάλυσης που χρησιμοποιεί ακόμα και σήμερα και τελικά θα μπορέσει να συλλάβει το τι ακριβώς θα πρέπει να περιμένει στο μέλλον. Και όλα αυτά διότι ο τομέας της ανάλυσης των κατηγορικών δεδομένων είναι δυναμικός και διαρκώς εξελισσόμενος, εξαιτίας και των εφαρμογών του σε άλλα ερευνητικά πεδία, όπως θα διαπιστώσουμε στη συνέχεια.

1.1 Το παρελθόν

Όπως αναφέρει ο A.Agresti στο βιβλίο του *Categorical Data Analysis* (2002), μετά το τέλος του 2^{ου} παγκοσμίου πολέμου, άρχισε η ανάπτυξη της θεωρητικής θεμελίωσης των πινάκων συνάφειας. Ο H.Cramer διατύπωσε ποικίλες εκφράσεις όσον αφορά σε κατανομές για μεγάλα δείγματα και ο C.R.Rao (1957, 1963) πραγματοποίησε σχετική εργασία με αυτές.

Το 1949, ο στατιστικός Neyman, ο οποίος είχε ήδη παρουσιάσει σημαντική εργασία στον έλεγχο υποθέσεων και στις μεθόδους εκτίμησης διαστημάτων εμπιστοσύνης, μαζί με τον E.S.Pearson, παρουσίασε την οικογένεια βέλτιστων ασυμπτωτικά κανονικών εκτιμητών (*Best Asymptotically Normal-BAN*), η οποία περιλαμβάνει εκτιμητές σταθμισμένων ελαχίστων τετραγώνων. Η απλότητα στους υπολογισμούς, τους κάνει να συγκρίνονται με τους εκτιμητές μέγιστης πιθανοφάνειας (*Maximum Likelihood-ML*) και αποτελεί μια σημαντική διαπίστωση πριν την επινόηση των σύγχρονων υπολογιστικών συστημάτων.

Στις αρχές της δεκαετίας του 1950, ο W.Cochran εξέδωσε εργασίες που αφορούσαν σε μια ποικιλία από σημαντικά θέματα στην ανάλυση κατηγορικών δεδομένων: μοντελοποίησε Poisson και διωνυμικές αποκρίσεις με μετασχηματισμούς διακύμανσης(1940), αναγνώρισε και μελέτησε τρόπους που ασχολούνται με την υπερμεταβλητότητα (*overdispersion*) (1943)

και παρουσίασε μια γενίκευση του ελέγχου McNemar (*Cochran's Q*) για σύγκριση ποσοστών κατά ζεύγη (1950). Το κλασικό άρθρο του (1954) εισάγει νέες μεθοδολογίες για την εφαρμοσμένη στατιστική. Επίσης, έδωσε κατευθυντήριες γραμμές για X^2 προσεγγίσεις σε μεγάλα δείγματα ώστε να «δουλέψουν» καλά για τη στατιστική συνάρτηση X^2 και τόνισε τη σημασία της διαμέρισης του στατιστικού X^2 σε συνιστώσες. Τέλος, ο Cochran πρότεινε έναν έλεγχο για υπό συνθήκη ανεξαρτησία σε διάφορους 2×2 πίνακες συνάφειας, που σχετίζεται πολύ με τον έλεγχο των Mantel-Haenszel (1959).

Η εργασία του Bartlett για τη δομή αλληλεπίδρασης σε $2\times2\times2$ πίνακες συνάφειας είχε μάλλον μικρό αντίκτυπο για τα επόμενα 20 χρόνια, αλλά από τα μέσα της δεκαετίας του 1950 και αρχές της δεκαετίας του ΄60, η εργασία του Bartlett επεκτάθηκε με διάφορους τρόπους για πίνακες πολλαπλής εισόδου, από τους: Darroch (1962), Good (1963), Goodman (1964b), Plackett (1962), Roy&Kastenbaum (1956) και Roy&Mitra (1956). Τα άρθρα αυτά, καθώς και τα άρθρα του M.W.Birch (1963, 1964a,b, 1965) αποτελούν την γένεση της έρευνας για τα λογαριθμογραμμικά μοντέλα κατά τα έτη 1965 έως 1975. Ο Birch έδειξε με ποιο τρόπο λαμβάνουμε εκτιμητές μέγιστης πιθανοφάνειας των συχνοτήτων των κελιών σε πίνακες τριπλής εισόδου υπό από διάφορες συνθήκες και έδειξε την ισοδυναμία τους για Poisson και πολυωνυμικά δείγματα. Μαζί με τον Watson (1959), επέκτειναν τα θεωρητικά αποτελέσματα των Cramer και Rao για κατανομές μεγάλων δειγμάτων για μοντέλα πινάκων συνάφειας. Ο Mantel συζήτησε πρώιμα αποτελέσματα και διασαφήνισε την τυποποίηση των λογαριθμογραμμικών μοντέλων. Ο Caussinus (1966) παρουσίασε την quasi-συμμετρία σε τετραγωνικούς πίνακες.

Τις επόμενες δεκαετίες, πραγματοποιήθηκε αρκετή έρευνα πάνω στα λογαριθμογραμμικά μοντέλα και στα μοντέλα logit σε τρία πανεπιστήμια της Αμερικής: το πανεπιστήμιο του Σικάγο, το Χάρβαρντ και το πανεπιστήμιο της Βόρειας Καρολίνας. Στο πανεπιστήμιο του Σικάγο, ο L.Goodman έγραψε μια σειρά από άρθρα που αφορούσαν στα εξής θέματα: διαμερισμός του X^2, μοντέλα για τετραγωνικούς πίνακες (***quasi***-ανεξαρτησία), stepwise διαδικασία για logit και λογαριθμογραμμικά μοντέλα, παραγωγίσιμες ασυμπτωτικά διασπορές, εκτιμητές μέγιστης πιθανοφάνειας λογαριθμογραμμικών παραμέτρων, μοντέλα λανθανουσών μεταβλητών (***latent variables***), μοντέλα συσχέτισης, ανάλυση αντιστοιχιών και μοντέλα συνάφειας. Ο Goodman (1969b) έγραψε επίσης άρθρα για περιοδικά κοινωνικών επιστημών, που συνέβαλλαν ουσιαστικά στο να γίνουν δημοφιλείς λογαριθμογραμμικές και logit μέθοδοι για σχετικές εφαρμογές. Τα τελευταία 50 χρόνια, ο Goodman ήταν ο πιο σημαντικός ερευνητής στην ανάπτυξη μεθόδων ανάλυσης κατηγορικών δεδομένων. Το πεδίο

της στατιστικής του οφείλει τεράστια ευγνωμοσύνη για το σημαντικό και τεράστιο όγκο της εργασίας του.

Μαθητές του Goodman στο Πανεπιστήμιο του Σικάγο με τις εργασίες τους, συνέβαλλαν επίσης σημαντικά στον τομέα της στατιστικής, όπως ο Haberman το 1970 όπου ολοκλήρωσε τη διδακτορική διατριβή του με το αναπτύξει και να μελετήσει τα λογαριθμικά μοντέλα και τα λογαριθμογραμμικά μοντέλα για διατάξιμες μεταβλητές ταξινόμησης. Μερικά από τα θέματα που ασχολήθηκε ήταν η ανάλυση καταλοίπων, η ύπαρξη εκτιμητών μέγιστης πιθανοφάνειας, τα λογαριθμογραμμικά μοντέλα για διατάξιμες μεταβλητές και θεωρητικά μοντέλα στα οποία ο αριθμός των παραμέτρων αυξάνει καθώς αυξάνει το μέγεθος του δείγματος. Ο C.Clogg ακολούθησε τα βήματα στην εργασία του Goodman, όσον αφορά στις κοινωνικές επιστήμες, στα μοντέλα συνάφειας, στη δημογραφία, στην απογραφή, στα μοντέλα για ποσοστά και άλλα παρόμοια θέματα.

Ταυτόχρονα με την εργασία του Goodman, σχετική έρευνα σε μεθόδους μέγιστης πιθανοφάνειας για λογαριθμογραμμικά και logit μοντέλα πραγματοποιήθηκε στο Χάρβαρντ από τους μαθητές του F.Mosteller, όπως ο S.Fienberg, και του W.Cochran. Μεγάλο μέρος της εργασίας αυτής εμπνεύστηκε από τα προβλήματα που προέκυψαν στην ανάλυση πολυμεταβλητών συνόλων δεδομένων στη μελέτη για μια αναισθητική ουσία (***National Halothane Study***). Αυτή η μελέτη εξετάζει εάν η ουσία αυτή είναι πιο πιθανή να προκαλέσει θάνατο από άλλα αναισθητικά λόγω βλάβης στο συκώτι. Μια αναφορά από τον Mosteller (1968) στην American Statistical Association περιγράφει πρώιμες χρήσεις των λογαριθμογραμμικών μοντέλων για εξομάλυνση πολυδιάστατων διακριτών συνόλων δεδομένων. Ο Fienberg και οι δικοί του μαθητές επέκτειναν την εργασία του Mosteller. Με το βιβλίο του «Διακριτή Πολυμεταβλητή Ανάλυση» (***Discrete Multivariate Analysis***) (1975) συνέβαλε κατά πολύ στην παρουσίαση των λογαριθμογραμμικών μοντέλων στον ευρύτερο κόσμο της στατιστικής και παραμένει ακόμη μια άριστη αναφορά.

Έρευνα στην Βόρεια Καρολίνα από τον G.Koch και διάφορους μαθητές και συνεργάτες του, επηρέασε σε μεγάλο βαθμό τις βιο-ιατρικές επιστήμες. Ο Koch ανέπτυξε μεθόδους σταθμισμένων ελαχίστων τετραγώνων για μοντέλα κατηγορικών δεδομένων και, το 1969, το άρθρο του μαζί με τους J.Grizzle και F.Starmer έκανε δημοφιλή αυτήν την προσέγγιση. Ο Koch και οι συνάδελφοί του την επέκτειναν σε επόμενα άρθρα, σε μια εντυπωσιακή ποικιλία προβλημάτων, συμπεριλαμβανομένων προβλημάτων για τα οποία η μέθοδος μέγιστης πιθανοφάνειας είναι περίεργο να χρησιμοποιηθεί, όπως η ανάλυση κατηγορικών δεδομένων επαναλαμβανόμενων μετρήσεων (1977). Το 1966, ο V.Bhapkar

έδειξε ότι ο εκτιμητής σταθμισμένων ελαχίστων τετραγώνων ταυτίζεται συχνά με τον εκτιμητή του Neyman(*Neyman's minimum modified chi-squared estimator*).

Αρχικά, η βιβλιογραφία στα λογαριθμογραμμικά μοντέλα αντιμετώπιζε όλες τις ταξινομήσεις ως ποιοτικές. Ο Haberman (1974b) και ο Simon (1974) μελέτησαν τον τρόπο χειρισμού της διαταξιμότητας των μεταβλητών στα λογαριθμογραμμικά μοντέλα. Αυτή η εργασία επεκτάθηκε σε διάφορα άρθρα από τον Goodman (1979a, 1981a, b, 1983, 1985, 1986). Οι επεκτάσεις αυτές περιλαμβάνουν τα μοντέλα συνάφειας (***association models***), που αντικαθιστούν τα διατεταγμένα σκορς στα λογαριθμογραμμικά μοντέλα με παραμέτρους. Ο Goodman επίσης ασχολήθηκε (1985, 1986, 1996) και με τα μοντέλα συσχέτισης (***correlation models***), αλλά και μελέτησε μια προοπτική που στηρίζεται σε μεθόδους της ανάλυσης αντιστοιχιών (***correspondence analysis***). Στην ανάλυση αντιστοιχιών και την πολλαπλή ανάλυση αντιστοιχιών ή ανάλυση ομοιογένειας θα αναφερθούμε λεπτομερώς στο κεφάλαιο 6. Ωστόσο, τα μοντέλα συνάφειας, αλλά και τα μοντέλα (κανονικής) συσχέτισης δεν χρησιμοποιήθηκαν ευρέως, όπως συνέβη στα μοντέλα παλινδρόμησης.

Τέλος, συγκεκριμένα λογαριθμογραμμικά μοντέλα με δομή υπό συνθήκη ανεξαρτησίας, παρέχουν γραφικά μοντέλα (***graphical models***) για πίνακες συνάφειας, τα οποία σχετίζονται και με τα γραφήματα συνάφειας (***association graphs***). Με αυτά ασχολήθηκαν πρώτοι οι Darroch et all(1980). Επίσης, τα μοντέλα λανθανουσών μεταβλητών συνδέθηκαν αρκετά με τα γραφικά μοντέλα, κυρίως από τους Wermunt&Cox (1998), Anderson&Bockenholt (2000), Anderson&Vermunt (2000), Anderson (2002).

1.2 Το παρόν

Η πιο ενεργή περιοχή της νεότερης έρευνας στην ανάλυση κατηγορικών δεδομένων, κυρίως την τελευταία δεκαετία, αφορά στην μοντελοποίηση των δεδομένων κατά συστάδες (***cluster analysis***), όπως συμβαίνει στις διαχρονικές (***longitudinal***) μελέτες και σε άλλες μορφές επαναλαμβανόμενων μετρήσεων. Εδώ, υπάρχει μια ποικιλία από μεθόδους μοντελοποίησης, καθώς μετράται η συσχέτιση μεταξύ των αποκρίσεων στην ίδια συστάδα. Στις μεθόδους αυτές χρησιμοποιούνται εκτιμητές μέγιστης πιθανοφάνειας, αλλά η προσαρμογή τους σε ορισμένα μοντέλα είναι αρκετά δύσκολη υπολογιστικά.

Τα λογαριθμογραμμικά μοντέλα και τα μοντέλα λογιστικής παλινδρόμησης, όπως προαναφέρθηκε, μελετήθηκαν από τις αρχές της δεκαετίας το'60. Ωστόσο, στη δεκαετία του '80 ιδιαίτερη προσοχή δόθηκε στα αθροιστικά (***cumulative***) logit μοντέλα από τους McCullach (1980) και Goodman (1979). Επίσης, για διατάξιμες μεταβλητές αναπτύχθηκε το logit μοντέλο διαδοχικών κατηγοριών (***adjacent categories***) από τον Simon (1974) και Goodman (1983). Ο

Becker (1990) πρότεινε εκτιμητές μέγιστης πιθανοφάνειας για γενικευμένα μοντέλα συνάφειας και ο Becker (1989), αλλά και οι Becker&Clogg (1989) πρόσθεσαν συμμεταβλητές στα μοντέλα συνάφειας. Οι Bartolucci&Forcina (2003) επέκτειναν τα μοντέλα συνάφειας με το μοντελοποιήσουν ταυτόχρονα τις περιθώριες κατανομές με διάφορα διατάξιμα logit μοντέλα.

Ακόμη, αναπτύχθηκαν μέθοδοι εκτίμησης που στηρίζονται στις γενικευμένες εξισώσεις εκτίμησης (*Generalized Estimating Equations-GEE*), οι οποίες αφορούσαν αρχικά σε περιθώρια μοντέλα (*marginal models*) με μονομεταβλητές κατανομές, αλλά επεκτάθηκαν και σε άλλα αθροιστικά logit μοντέλα (Lipsitz και άλλοι, 1994) και αθροιστικά probit μοντέλα (Toledano&Gatsonis, 1996) για επαναλαμβανόμενες διατάξιμες αποκριτικές μεταβλητές. Πιθανές επεκτάσεις των περιθωριακών μοντέλων ώστε να περιλαμβάνουν και πολυμεταβλητές περιπτώσεις στην ουσία υλοποιούνται με τη μέθοδο *GEE* (βλέπε Qu et all, 1995). Μια προσέγγιση γενικευμένων γραμμικών mixed μοντέλων (*Generalized Linear Mixed Models-GLMMs*), μπορούμε να αναζητήσουμε στους Hedeker&Gibbons (1994) και στον Fielding (1999).

Σήμερα, ένας από τους σημαντικότερους ερευνητές στον τομέα της ανάλυσης κατηγορικών δεδομένων, θεωρείται ο Alan Agresti, του οποίου το βιβλίο *Categorical Data Analysis, Wiley* (2002), αποτελεί σημείο αναφοράς για οποιονδήποτε μελετητή.

Τέλος, η ανάπτυξη της μπεϋζιανής (*bayesian*) προσέγγισης στην ανάλυση κατηγορικών δεδομένων είναι μια συνεχώς αναπτυσσόμενη και ενεργή περιοχή, αλλά η πολυπλοκότητα των παραμέτρων περιπλέκει την μπεϋζιανή μοντελοποίηση. Ωστόσο, για πρώιμη χρήση της μπεϋζιανής εκτίμησης παραμέτρων, μπορούμε να αναζητήσουμε στον Good (1965) και τον Lindley (1964). Το σχετικό άρθρο του Good εξελίχθηκε προφανώς από την εργασία του κατά τη διάρκεια του $2^{ου}$ παγκοσμίου πολέμου όπου μαζί με τον Alan Turing στο Bletchley Park της Αγγλίας «έσπαγαν» κωδικούς των Ναζί. Μπεϋζιανές προσεγγίσεις μπορούμε να αναζητήσουμε στους Tan et all(1999), Chen&Dey(2000) και Qiu et all(2002). Οι Johnson&Albert(1999) επικεντρώθηκαν σε μπεϋζιανές προσεγγίσεις σχετικά με διατάξιμες αποκριτικές μεταβλητές. Για επιπλέον μπεϋζιανές προσεγγίσεις, βλέπε Chipman&Hamada(1996), Lang(1999), Johnson(1996), Albert&Chib (1993), Cowles et all(1996), Chib&Greenberg(1998), Qu&Tan(1998), Bradlow&Zaslavsky(1999), Tan et all(1999), Chen&Shao(1999), Chib(2000), Ishwaran&Gatsonis(2000), Ishwaran(2000), Xie et all(2000), Rossi et all(2001), Biswas&Das (2002), Webb&Forster(2004), Agresti&Hitchcock(2004).

1.3 Το μέλλον

Το να προβλέπει κανείς το μέλλον είναι πάντοτε ριψοκίνδυνο. Ωστόσο, ο μεγαλύτερος όγκος εργασίας σχετικά με τα (διατάξιμα) κατηγορικά δεδομένα είναι πολύ πιθανό να επικεντρωθεί σε υπολογιστικές μεθόδους, όπως στα γενικευμένα γραμμικά mixed μοντέλα. Άλλο φλέγον θέμα είναι η ανάπτυξη αλγοριθμικών μεθόδων για τεράστια σύνολα δεδομένων με μεγάλο αριθμό μεταβλητών, που βρίσκεται πολύ μακριά από την παραδοσιακή μοντελοποίηση. Τέτοιες μέθοδοι, που συχνά αναφέρονται ως data mining, ασχολούνται με το χειρισμό περίπλοκων δομών δεδομένων, πλεονεκτώντας στη δύναμη της πρόβλεψης, αλλά θυσιάζοντας την απλότητα και την ερμηνεία της δομής. Σημαντικές περιοχές εφαρμογής αυτών είναι η γενετική, όπως η ανάλυση διακριτών DNA ακολουθιών με τη μορφή πολύ μεγάλης διάστασης πινάκων συνάφειας και εφαρμογές σε επιχειρήσεις όπως credit scoring και μεθόδους με δομή «δένδρου» (tree) για την πρόβλεψη μελλοντικής συμπεριφοράς πελατών.

Σύμφωνα με την άποψη της Μ.Κατέρη (βλέπε Liu&Agresti, Discussion, 2005), το μέλλον των διατάξιμων κατηγορικών δεδομένων βρίσκεται στην ανάλυση των επαναλαμβανόμενων μετρήσεων ή γενικότερα σε συσχετισμένα δεδομένα. Επίσης, αναμένεται να αναπτυχθούν μη παραμετρικές μέθοδοι, ιδιαίτερα για προβλήματα υψηλότερης διάστασης, όπου να μπορούν να συγκριθούν ως προς τη δυναμική με τις αντίστοιχες παραμετρικές μεθόδους. Ωστόσο, το μέλλον των κατηγορικών δεδομένων και γενικότερα της στατιστικής τοποθετείται στην μπεϋζιανή ανάλυση, η οποία απαιτεί μικρού μεγέθους δείγματα που σημαίνει λιγότερες υπολογιστικές απαιτήσεις.

ΚΕΦΑΛΑΙΟ 2^{ο}

Γραφικές Απεικονίσεις Κατηγορικών Δεδομένων

2.0 Εισαγωγή

Οι στατιστικές μέθοδοι για κατηγορικά δεδομένα, όπως τα λογαριθμογραμμικά μοντέλα και η λογιστική παλινδρόμηση, είναι τεχνικές ανάλυσης δεδομένων ανάλογες της ανάλυσης διακύμανσης και των μεθόδων παλινδρόμησης για συνεχείς αποκριτικές μεταβλητές. Τα κατηγορικά δεδομένα αναπαρίστανται πιο συχνά σε πίνακες συνάφειας και οι αναλύσεις που χρησιμοποιούν λογαριθμογραμμικά μοντέλα και λογιστική παλινδρόμηση αναπαρίστανται πιο συχνά με όρους εκτίμησης παραμέτρων.

Διάφορα σχήματα για γραφική αναπαράσταση πινάκων συνάφειας στηρίζονται στο γεγονός ότι αν οι μεταβλητές των γραμμών και των στηλών είναι ανεξάρτητες, οι αναμενόμενες συχνότητες των κελιών είναι το γινόμενο των αθροισμάτων των γραμμών και των στηλών διαιρεμένο με το συνολικό άθροισμα. Τότε, κάθε κελί του πίνακα συνάφειας μπορεί να αναπαρασταθεί ως ένα ορθογώνιο του οποίου το εμβαδόν δείχνει τη συχνότητα του κελιού ή την απόκλιση από την ανεξαρτησία. Παρόλο που οι τεχνικές γραφικής απεικόνισης είναι κοινό επακόλουθο της ανάλυσης διακύμανσης και της ανάλυσης παλινδρόμησης, οι μέθοδοι για την γραφική αναπαράσταση δεδομένων σε πίνακα συνάφειας δεν χρησιμοποιούνται στην πράξη τόσο ευρέως.

Από την δεκαετία του ΄90, οι Hartigan and Kleiner, ο Friendly et all, ανέπτυξαν καινοτόμες μεθόδους απεικόνισης κατηγορικών δεδομένων, σχεδιασμένες ώστε να προσφέρουν απεικονίσεις ανάλογες με αυτές που χρησιμοποιούνται στα συνεχή δεδομένα όπου η φυσική ερμηνεία είναι άμεση. Στην παρούσα εργασία, θα ασχοληθούμε κυρίως με τα μωσαϊκά (*mosaic plots*), αλλά θα αναφερθούμε και στις γραφικές αναπαραστάσεις συνάφειας (*association plots*).

2.1 Μωσαϊκά

Τα μωσαϊκά αναπαριστούν τις συχνότητες ενός πίνακα συνάφειας με ορθογώνια πλακίδια (*tiles*), των οποίων το μέγεθος είναι ανάλογο της συχνότητας του κελιού. Αυτή η γραφική απεικόνιση για τα κατηγορικά δεδομένα γενικεύεται εύκολα σε πίνακες πολλαπλής εισόδου. Τα μωσαϊκά εισήχθησαν από τους Hartigan and Kleiner (1981, 1984), οι οποίοι

υποστήριξαν ότι το μοτίβο των ορθογωνίων πλακιδίων που υπάρχουν σε αυτά είναι χρήσιμο για να διατυπωθούν υποθέσεις, να γίνουν οπτικές συγκρίσεις μεταξύ τμημάτων (υποπίνακες) ενός πίνακα συνάφειας και να δοθεί έμφαση σε κελιά με μεγάλες ή μικρές συχνότητες.

Ο Friendly (1994) επέκτεινε την χρήση των μωσαϊκών ως ένα εργαλείο ανάλυσης δεδομένων με τον εξής τρόπο: για συγκεκριμένη απεικόνιση, προσάρμοσε ένα βασικό μοντέλο ανεξαρτησίας ή μερική ανεξαρτησίας και χρησιμοποίησε χρώμα και σκίαση των ορθογώνιων πλακιδίων για να αποδώσει την αποχώρηση από την ανεξαρτησία. Για μη διατάξιμες μεταβλητές, για να δώσει μια ιδέα για το μοτίβο της συνάφειάς τους, αναδιέταξε τις κατηγορίες, κυρίως βασιζόμενος στην Singular Value Decomposition (*SVD*) των καταλοίπων της ανεξαρτησίας. Για πολλαπλούς πίνακες εισόδου, θεώρησε χρήσιμο να εξετάσει μια ακολουθία από μωσαϊκά για περιθωριακούς υποπίνακες ως διαδοχικές μεταβλητές. Παρόλο που οποιοδήποτε λογαριθμογραμμικό μοντέλο μπορεί να προσαρμοστεί σε ολόκληρο τον πίνακα, μια κλάση από μοντέλα της από κοινού ανεξαρτησίας παρέχει μια γραφική αναπαράσταση μια διαμέρισης του ολικού λόγου πιθανοφάνειας G^2 της πλήρης ανεξαρτησίας σε ολόκληρο τον πίνακα, σε μέρη που συμβάλλουν στη διατύπωση υποθέσεων για τους περιθωριακούς υποπίνακες.

Στόχος των γραφικών αυτών μεθόδων είναι:

- Να παρέχουν μεθόδους απεικόνισης για διερεύνηση της δομής των κατηγορικών δεδομένων και προσαρμογή μοντέλων συγκρίσιμη με αυτή που χρησιμοποιείται στα συνεχή δεδομένα.
- Να υλοποιούν τις μεθόδους αυτές με ευανάγνωστο λογισμικό.

2.1.1 Παράδειγμα μωσαϊκού για πίνακα συνάφειας διπλής εισόδου

Θα παρουσιάσουμε σύντομα την γραφική απεικόνιση με μωσαϊκά για δισδιάστατους πίνακες συνάφειας, ώστε να γίνει πιο εύκολα κατανοητή η γενίκευσή τους στη συνέχεια για πολυδιάστατους πίνακες συνάφειας. Για το σκοπό αυτό, θα χρησιμοποιήσουμε τα δεδομένα που συνδέουν το χρώμα μαλλιών και χρώμα ματιών 592 σπουδαστών στο μάθημα της στατιστικής (Πίνακας 2.1), τα οποία συλλέχθηκαν από τον Snee (1974) και αναλύθηκαν από τον Friendly (1994).

	Χρώμα Μαλλιών				
Χρώμα Ματιών	Μαύρα	Καστανά	Κόκκινα	Ξανθά	**Σύνολο**
Καφέ	68	119	26	7	**220**
Μπλε	20	84	17	94	**215**
Φουντουκί	15	54	14	10	**93**
Πράσινα	5	29	14	16	**64**
Σύνολο	**108**	**286**	**71**	**127**	**592**

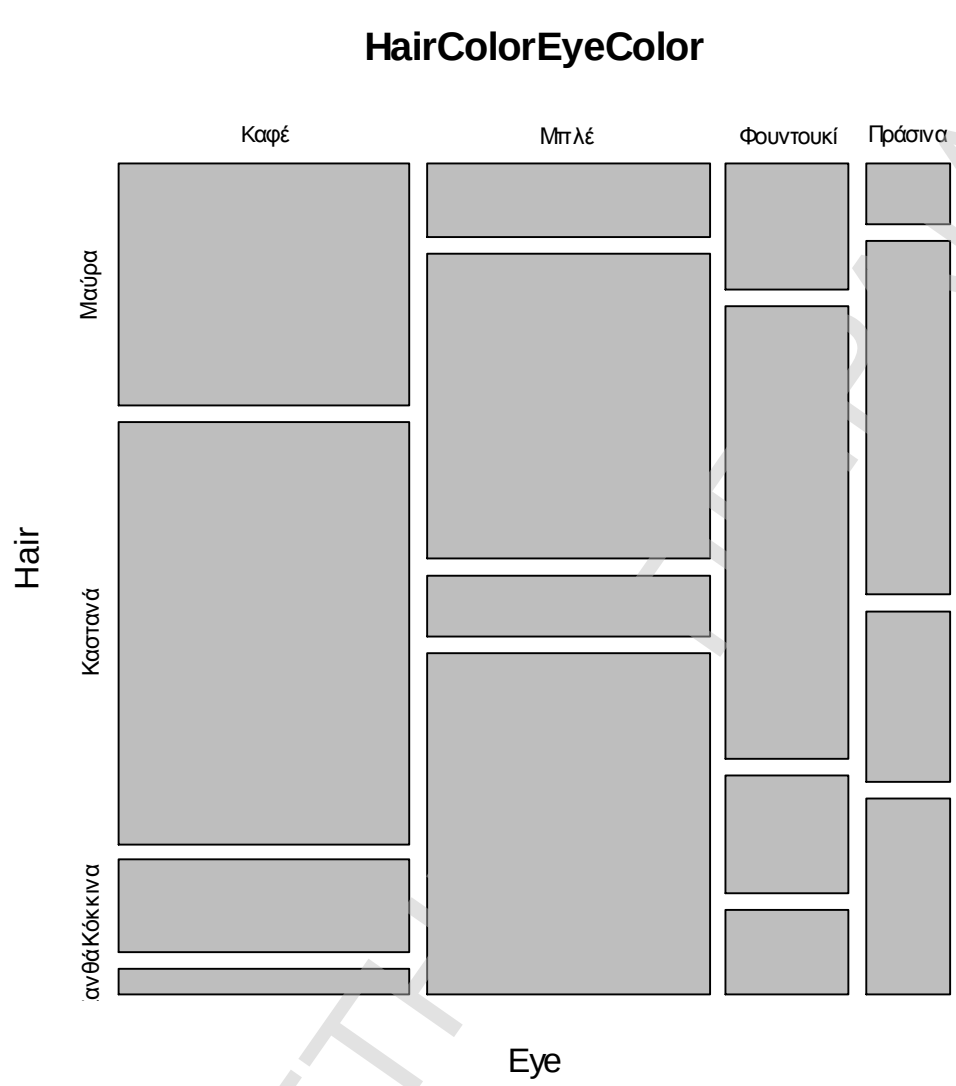
Πίνακας 2.1

Δεδομένα για χρώμα ματιών και χρώμα μαλλιών

Η στατιστική συνάρτηση X^2 του Pearson για τον έλεγχο καλής προσαρμογής του μοντέλου ανεξαρτησίας είναι 138.3 με 9 βαθμούς ελευθερίας, που δηλώνει ουσιαστική απομάκρυνση από την ανεξαρτησία. Το ζητούμενο, ωστόσο, είναι να κατανοήσουμε τη φύση της συνάφειας μεταξύ του χρώματος μαλλιών και ματιών. Για αυτό το σκοπό, ο Friendly κατασκεύασε και μελέτησε τα αντίστοιχα μωσαϊκά.

Στα μωσαϊκά αυτά, για οποιοδήποτε πίνακα συνάφειας, οι αναμενόμενες συχνότητες υπό την συνθήκη της ανεξαρτησίας αναπαριστώνται με ορθογώνια των οποίων τα μήκη είναι ανάλογα με τη συνολική συχνότητα της κάθε στήλης και τα πλάτη είναι ανάλογα με την υπό συνθήκη συχνότητα για κάθε γραμμή δοθείσης της συχνότητα της κάθε στήλης. Το εμβαδόν κάθε ορθογωνίου είναι ανάλογο με τη συχνότητα κάθε κελιού του πίνακα συνάφειας.

Η μορφή του γραφήματος που δικαιολογεί την ονομασία μωσαϊκό είναι αυτή του γραφήματος 2.1, και μοιάζει με ένα διαιρεμένο ραβδόγραμμα. Τα μήκη των ορθογωνίων εξακολουθούν και είναι ανάλογα με τη συνολική συχνότητα της κάθε στήλης, αλλά τα πλάτη είναι ανάλογα με την υπό συνθήκη συχνότητα για κάθε γραμμή (χρώμα ματιών) δοθείσης της συχνότητα της κάθε στήλης (χρώμα μαλλιών), έτσι ώστε το εμβαδόν του κάθε ορθογωνίου να είναι ανάλογο με τη συχνότητα του κελιού και η ανεξαρτησία να φαίνεται γραφικά όταν όλα τα ορθογώνια έχουν το ίδιο πλάτος.



Γράφημα 2.1

Μωσαϊκό για τις μεταβλητές Hair, Eye.

Όπως διαπιστώνουμε από το γράφημα 2.1, το χρώμα ματιών και το χρώμα μαλλιών δεν είναι ανεξάρτητες μεταβλητές. Επίσης, η πιο διαδεδομένη κατηγορία, όσοι έχουν χρώμα μαλλιών καστανό, έχουν κυρίως καφέ ή φουντουκί χρώμα ματιών και η λιγότερο διαδεδομένη κατηγορία, οι ξανθοί έχουν κυρίως καφέ ή μπλε μάτια. Επίσης, όσοι έχουν πράσινα μάτια, συνήθως έχουν και καστανά μαλλιά.

2.1.2 Γενίκευση του μωσαϊκού σε πίνακα πολλαπλής εισόδου

Η παραπάνω μορφή του μωσαϊκού γενικεύεται εύκολα για πολυδιάστατους πίνακες συνάφειας. Στη συνέχεια, θα περιγράψουμε με βήματα την κατασκευή ενός μωσαϊκού για πίνακα συνάφειας τριπλής εισόδου για τις μεταβλητές A, B, C με συχνότητες n_{ijk}, i=1,…, I, j=1, …, J, k=1,… , K. Η γενίκευση σε πίνακες ν-οστής εισόδου είναι άμεση. Η δομή του μωσαϊκού στηρίζεται στη διαταξιμότητα των μεταβλητών. Η διαίρεση σε ορθογώνια συνήθως εναλλάσσεται κατακόρυφα και οριζόντια με την παρακάτω σειρά:

1. Πρώτα, η διαθέσιμη περιοχή του γραφήματος διαιρείται σε κατακόρυφες λωρίδες που το εμβαδόν τους είναι ανάλογο με τα περιθώρια αθροίσματα της μεταβλητής A, έτσι ώστε τα μήκη να είναι ανάλογα στη συχνότητα $n_{i..}$, όπου η τελεία συμβολίζει ότι το άθροισμα ως προς τον δείκτη αυτό.
2. Κάθε κατακόρυφη λωρίδα κατόπιν υποδιαιρείται οριζόντια ανάλογα με τις από κοινού συχνότητες με τη δεύτερη μεταβλητή, $n_{ij.}$. Αυτό σημαίνει ότι κάθε ορθογώνιο έχει πλάτος ανάλογο με την υπό συνθήκη συχνότητα της δεύτερης μεταβλητής δοθείσης της πρώτης μεταβλητής, $n_{ij.}/n_{i..}$ και εμβαδόν ανάλογο της $n_{ij.}$
3. Κάθε IJ ορθογώνιο διαιρείται καθέτως ανάλογα με τη $n_{ijk.}$, δίνοντας μήκη ανάλογα του $n_{ijk.}/n_{ij..}$, κ.ο.κ.

Στη συνέχεια, παραθέτουμε μερικές επιπλέον ιδιότητες που μπορεί να έχουν τα μωσαϊκά και να μας παρέχουν επιπρόσθετες πληροφορίες για τη δομή των δεδομένων και τη συνάφεια μεταξύ των μεταβλητών.

i. Διακεκομμένες γραμμές (*spacings*)

Η διαδικασία που περιγράφηκε παραπάνω δίνει ένα μωσαϊκό με IJK ορθογώνια χωρίς όμως διακεκομμένες γραμμές, όπου κελιά με μικρές συχνότητες είναι δύσκολο να εντοπιστούν. Σύμφωνα με τους Hartigan και Kleiner (1981), η καλύτερη λύση είναι τα ορθογώνια να διαχωρίζονται με διακεκομμένες γραμμές σε προηγούμενες υποδιαιρέσεις, ώστε να τονίζονται τα κελιά με μικρές συχνότητες. Για παράδειγμα, στον πίνακα τριπλής εισόδου με κατακόρυφο διαχωρισμό στις διαστάσεις I και K, οι διαιρέσεις της πρώτης μεταβλητής διαχωρίζεται ανάλογα με το 1/(I-1). Οι διαιρέσεις μεταξύ των επιπέδων της τρίτης μεταβλητής διαχωρίζονται ανάλογα με το 1/(IK-1).

ii. Προσαρμόζοντας λογαριθμογραμμικά μοντέλα

Όταν παρουσιάζονται τρεις ή περισσότερες μεταβλητές σε ένα μωσαϊκό, μπορούμε να προσαρμόσουμε διαφορετικά μοντέλα ανεξαρτησίας και να αναπαραστήσουμε τα κατάλοιπα

από τα μοντέλα αυτά. Οι αποκλίσεις μεταξύ των παρατηρηθέντων συχνοτήτων και των αναμενόμενων συχνοτήτων, που αναπαρίστανται με σκίαση, όπως θα αναφέρουμε στη συνέχεια στην περίπτωση iv, πολύ συχνά προτείνουν όρους που πρέπει να προστεθούν σε ένα μοντέλο ώστε να πετύχουμε καλύτερη προσαρμογή.

iii. Αναδιατάσσοντας κατηγορίες:

α. με τη βοήθεια της ανάλυσης αντιστοιχιών

Δοθείσης μιας διαταξιμότητας των μεταβλητών, οι κατηγορίες των ονοματικών μεταβλητών μπορούν να αναδιαταχθούν, αναδεικνύοντας αποκλίσεις από τα μοντέλα της από κοινού ανεξαρτησίας. Για τις μεταβλητές A, B, C, με στόχο τη διαίρεση του μωσαϊκού, οι κατηγορίες των μεταβλητών A και B μπορούν να αναδιαταχθούν από το (A, B) διάγραμμα, αυτές της μεταβλητής C μπορούν να αναδιαταχθούν από το (AB, C) διάγραμμα, κ.ο.κ. Η εμπειρία έδειξε ότι για μικρούς πίνακες οι διατάξεις που διαγωνιοποιούν το μοτίβο των υπολοίπων μπορούν συχνά να αποφασιστούν διαισθητικά.

Μια πιο γενική προσέγγιση βασίζεται σε ιδέες της ανάλυσης αντιστοιχιών (*Correspondence Analysis-CA*), που αναθέτει σκορ στις κατηγορίες. Συνεπώς, επανατακτοποιώντας τις κατηγορίες γραμμών-στηλών σύμφωνα με τα σκορ της ανάλυσης αντιστοιχιών για την πρώτη διάσταση, το μωσαϊκό παρέχει μια διάταξη ώστε να αποδώσει με τον καλύτερο τρόπο τη μορφή της συνάφειας.

β. αναδιατάσσοντας τα κατηγορικά δεδομένα

Το τρισδιάστατο περιστρεφόμενο γράφημα ήταν ένα από τα πρώτα δυναμικά γραφήματα που υλοποιήθηκαν, τα οποία επιτρέπουν στο χρήστη να εξερευνήσει τρισδιάστατα συνεχή δεδομένα από διαφορετικές γωνίες. Ο Cook, κ.α. (1995) γενίκευσε αυτήν την τεχνική για περισσότερες από τρεις μεταβλητές. Με τα κατηγορικά δεδομένα μπορεί να επιτευχθεί κάτι ανάλογο. Καθώς, η διάταξη των μεταβλητών σε ένα μωσαϊκό καθορίζει ποιες μεταβλητές θα απεικονιστούν δοθέντος άλλων μεταβλητών, αλλάζοντας τη διάταξή τους είναι παρόμοια διαδικασία με το να περιστρέφονται συνεχή δεδομένα στον τρισδιάστατο χώρο.

Η διάταξη των μεταβλητών απαιτεί γρήγορο και ευέλικτο χειρισμό. Εκτός από τον τρόπο της περιστροφής “με το χέρι”, υπάρχουν και δύο τρόποι αυτόματης εύρεσης της διάταξης των μεταβλητών:

(1)Βρίσκοντας τα κελιά με τη μεγαλύτερη συχνότητα.

Δοθέντος ενός συνδυασμού των επιπέδων των μεταβλητών, η διάταξη των μεταβλητών επιλέγεται ως εξής: η πρώτη μεταβλητή είναι η μεταβλητή περιέχει τις περισσότερες

παρατηρήσεις σε ένα συγκεκριμένο επίπεδο. Η δεύτερη μεταβλητή είναι η μεταβλητή, όπου η διασταύρωση της πρώτης και της τρίτης μεταβλητής σε συγκεκριμένα επίπεδά τους μεγιστοποιούνται, κ.ο.κ.

(2)Ελαχιστοποιώντας τον αριθμό των άδειων κελιών.

Πολύ συχνά υπάρχουν άδεια κελιά σε έναν πίνακα συνάφειας, πράγμα που σημαίνει και άδεια κελιά στα μωσαϊκά. Ένα πρώτο βήμα ώστε να ελαχιστοποιηθεί ο αριθμός των άδειων κελιών είναι να μην διαχωρίζονται άδεια κελιά, εάν ένα κελί είναι ήδη άδειο σε ένα υψηλότερο επίπεδο. Π.χ. δεν υπήρχαν παιδιά ή γυναίκες στο πλήρωμα του Τιτανικού, οπότε δεν αποκτάμε περισσότερη πληροφορία με το να διαχωρίσουμε αυτήν την άδεια ομάδα με βάση το φύλο.

iv. Κατάλοιπα(*Residuals*)-Χρώμα και Σκίαση στο μωσαϊκό

Για να γίνει το μωσαϊκό πιο περιεκτικό σε πληροφορία, προστίθεται σε αυτό χρώμα και σκίαση, όπου αποδίδεται η τυποποιημένη απόκλιση d_{ijk}(πρόκειται για τυποποιημένα κατάλοιπα Pearson) από την υπόθεση της ανεξαρτησίας. Στην παρούσα ανάλυση, όπως θα διαπιστώσουμε από τα παρακάτω γραφήματα, το μπλε χρώμα δηλώνει θετική απόκλιση και το κόκκινο χρώμα αρνητική απόκλιση. Τα ορθογώνια με πιο σκούρο χρώμα υποδηλώνουν μεγαλύτερη απόκλιση. Κελιά με τυποποιημένα κατάλοιπα Pearson $\left|d_{ijk}\right| < 2$ είναι χωρίς χρώμα, κελιά με $\left|d_{ijk}\right| \geq 2$ περιέχουν χρώμα και κελιά με $\left|d_{ijk}\right| \geq 4$ έχουν πιο σκούρο χρώμα.

Τα κατάλοιπα που χρησιμοποιούνται στην περαιτέρω ανάλυσή μας για να αποδώσουν την απόκλιση από την ανεξαρτησία είναι τα τυποποιημένα κατάλοιπα Pearson, έστω $d_{ijk} = \frac{n_{ijk} - \hat{m}_{ijk}}{\sqrt{\hat{m}_{ijk}}}$, έτσι ώστε η στατιστική συνάρτηση $X^2 = \sum_{i,j,k} d_{ijk}^2$. Όταν η προσαρμογή του μοντέλου που υποθέσαμε είναι καλή, τα κατάλοιπα αυτά ακολουθούν ασυμπτωτικά κανονική κατανομή με μέση τιμή 0 και διακύμανση μικρότερη του 1, με μέση τιμή της διακύμανσης να ισούται με (β.ε. καταλοίπων)/(αριθμός κελιών). Συνεπώς, όταν προσαρμόζονται μοντέλα πιο περίπλοκα από το μοντέλο της αμοιβαίας ανεξαρτησίας, η διακύμανση είναι σημαντικά μικρότερη του 1 και η χρήση συμβατικών γκαουσσιανών τιμών ± 2, ± 4 για τα κατάλοιπα είναι μάλλον αρκετά συντηρητικές, σύμφωνα με τον Agresti(1990) και αποτυγχάνουν να προτείνουν κάποια κελιά των οποίων η απόκλιση από την ανεξαρτησία πρέπει να ληφθεί σοβαρά υπόψη. Μια λύση που πρότεινε ο Friendly(1994) είναι η χρήση των προσαρμοσμένων καταλοίπων του Haberman(1973), όπου προκύπτουν από τα τυποποιημένα κατάλοιπα του

Pearson αν διαιρεθούν με την εκτιμηθείσα τυπική τους απόκλιση και ακολουθούν ασυμπτωτικά κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1.

Συνεπώς, το μωσαϊκό για έναν πίνακα πολλαπλής εισόδου δίνει πληροφορίες για την από κοινού κατανομή συχνοτήτων σε διάφορα επίπεδα:

- Οι συχνότητες των κελιών φαίνονται άμεσα από το εμβαδόν του κάθε ορθογωνίου.
- Κατά προσέγγιση σημαντικότητα των κατανομών των κελιών σε περίπτωση έλλειψη προσαρμογής σε συγκεκριμένο μοντέλο φαίνεται με τη σκίαση.

2.2 Παράδειγμα μωσαϊκού για πίνακα συνάφειας τριπλής εισόδου

Στον Πίνακα 2.2 περιέχονται τα δεδομένα που αφορούν στον αριθμό αυτοκτονιών στη δυτική Γερμανία, ταξινομημένα βάσει της ηλικίας(A), του φύλου(S) και της μεθόδου αυτοκτονίας(M). Τα δεδομένα συλλέχθηκαν από τον Heuer(1979, table 1) και μελετήθηκαν από τους van der Heijden και de Leeuw(1985), αλλά και άλλους.

		Μέθοδος (M)					
Φύλο(S)	**Ηλικία(A)**	**Poison**	**Gas**	**Hang**	**Drown**	**Gun**	**Jump**
M	**10-20**	1160	335	1524	67	512	189
M	**25-35**	2823	883	2751	213	852	366
M	**40-50**	2465	625	3936	247	875	244
M	**55-65**	1531	201	3581	207	477	273
M	**70-90**	938	45	2948	212	229	268
F	**10-20**	921	40	212	30	25	131
F	**25-35**	1672	113	575	139	64	276
F	**40-50**	2224	91	1481	354	52	327
F	**55-65**	2283	45	2014	679	29	388
F	**70-90**	1548	29	1355	501	3	383

Πίνακας 2.2

Συχνότητες αυτοκτονίας βάσει ηλικίας, φύλου και μεθόδου.

Ο Πίνακας 2.3 περιλαμβάνει την προσαρμογή (ιεραρχικών) λογαριθμογραμμικών μοντέλων για τα δεδομένα του Πίνακα 2.2. Είναι προφανές από τον πίνακα αυτό ότι κανένα μοντέλο δεν έχει αποδεκτή προσαρμογή στα δεδομένα. Δοθέντος του υπερβολικά μεγάλου

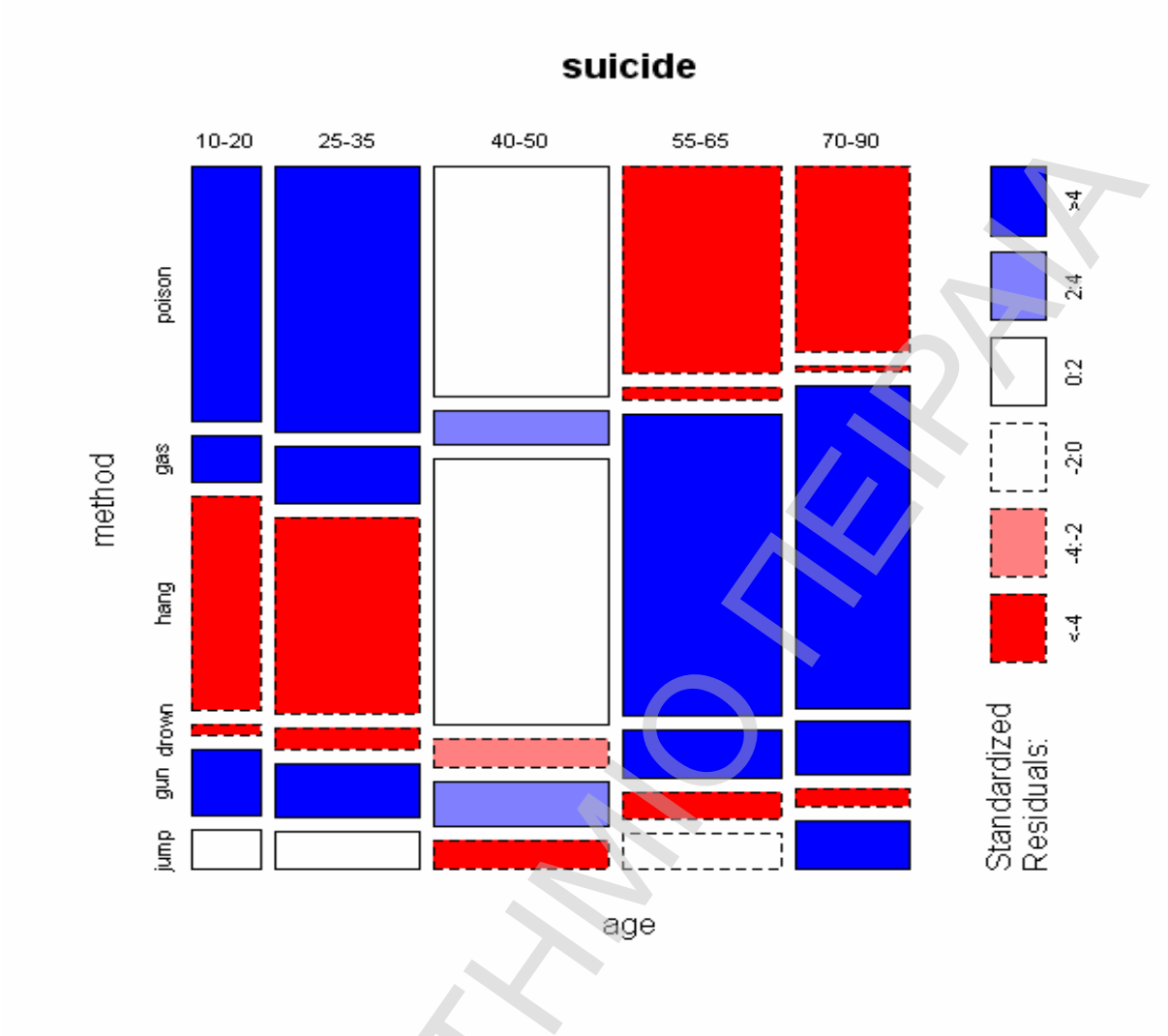
αριθμού του δείγματος ($n = 48.177$), ακόμη και σχετικά μικρές αποκλίσεις από τις αναμενόμενες συχνότητες υπό οποιοδήποτε μοντέλο θα φαίνονται ωστόσο σημαντικές. Πάντως, από τις διαφορές των στατιστικών συναρτήσεων G^2 και X^2 μεταξύ των μοντέλων, διαπιστώνουμε πως όλοι οι παράγοντες τριπλής αλληλεπίδρασης είναι στατιστικά μη σημαντικοί.

Μοντέλο	**β.ε.**	**G^2**	**X^2**
(M, A, S)	49	10119.6	9908.2
(M, AS)	45	8632	8371.3
(A, MS)	44	4719	4387.7
(S, MA)	29	7029.2	6485.5
(MS, AS)	40	3231.5	3030.5
(MA, AS)	25	5541.6	5135
(MA, MS)	24	1628.6	1592.4
(MA, MS, AS)	20	242	237

Πίνακας 2.3

Λογαριθμογραμμικά μοντέλα για τα δεδομένα αυτοκτονίας.

Οι μεταβλητές A, S και M μπορούν να διαταχθούν με διάφορους τρόπους. Αν δε θεωρήσουμε καμία μεταβλητή από τις τρεις μεταβλητές ως αποκριτική, μια απλή υπόθεση για γραφική αναπαράσταση των μεταβλητών θα ήταν η πρώτη και η τρίτη να διαμερίσουν μια διάσταση του μωσαϊκού. Επίσης, είναι χρήσιμο να θεωρήσουμε τέτοια διάταξη των μεταβλητών ώστε το γινόμενο ΙΚ των επιπέδων να μην είναι πολύ μεγάλο και να διατηρήσουμε την ικανότητα ανάλυσης του γραφήματος. Η σειρά A, S, M έχει $5\times6=30$ υποδιαιρέσεις κατακόρυφα και 2 οριζόντια, πράγμα που σημαίνει πως η εικόνα του μωσαϊκού δεν θα δίνει σημαντική πληροφορία. Επίσης, επειδή η σχέση μεταξύ της μεθόδου και της ηλικίας είναι ενδιαφέρουσα, τότε θεωρούμε τη σειρά M, A, S. Ωστόσο, για να διαπιστώσουμε την επίδραση στο μωσαϊκό, αν αλλάξουμε τη σειρά εισαγωγής των μεταβλητών, θα θεωρήσουμε και την περίπτωση S, A, M. Τέλος, μια σειρά των μεταβλητών θα πρέπει να μελετηθεί, η οποία να στηρίζεται στην προσαρμογή των λογαριθμογραμμικών μοντέλων που προαναφέραμε.

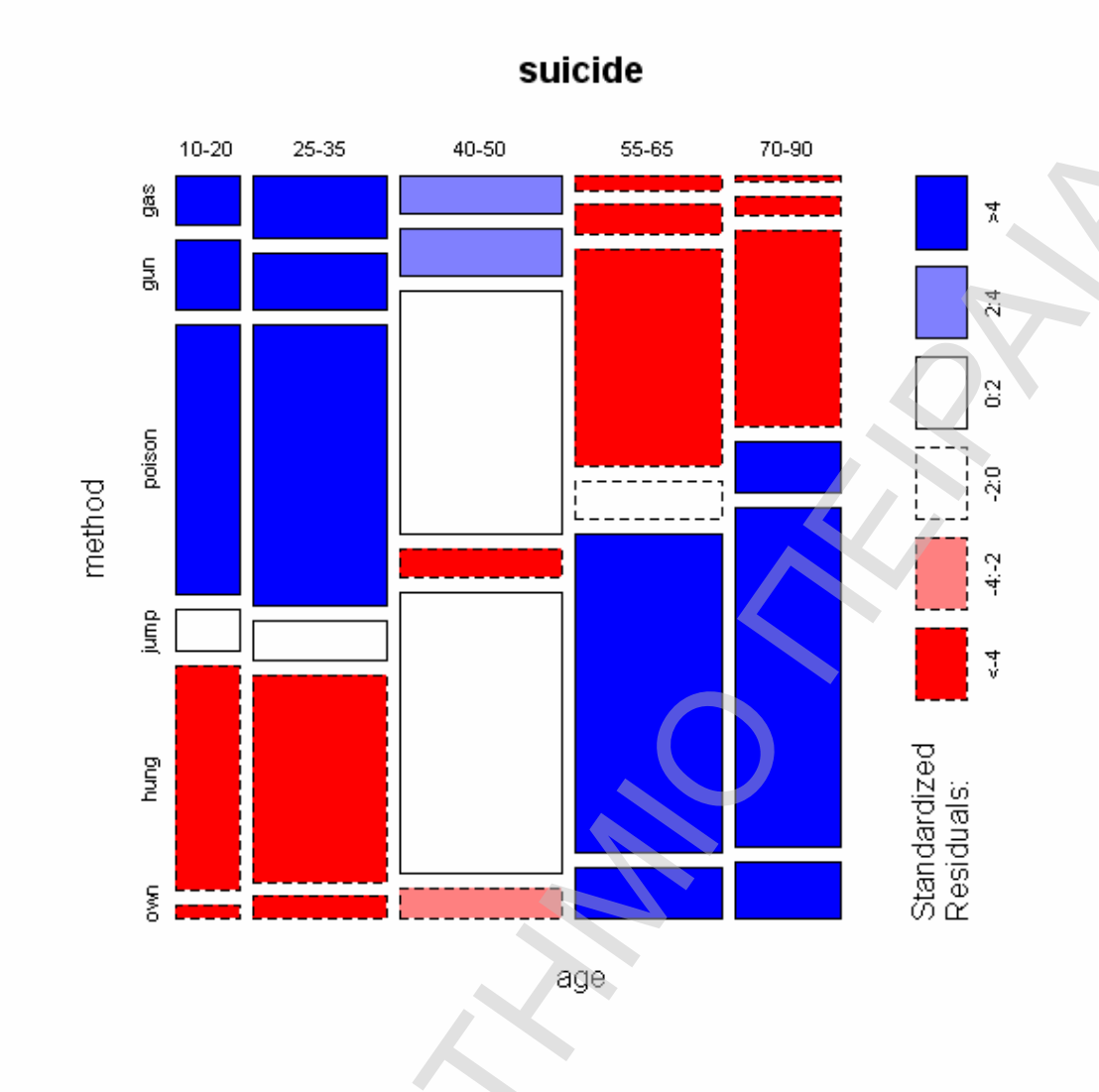


Γράφημα 2.2

Μωσαϊκό για τις μεταβλητές M, A.

Το γράφημα 2.2 αποτελεί το αρχικό μωσαϊκό για τις μεταβλητές M(*method*) και A(*age*) με σειρά εισαγωγής των μεταβλητών M, A, S. Όπως παρατηρούμε από το γράφημα αυτό, οι μέθοδοι POISON, GAS και GUN είναι επικρατέστερες σε νεαρές ηλικίες(10-20, 25-35) και φθίνουν, καθώς αυξάνει η ηλικία, ενώ αντίθετα οι μέθοδοι DROWN και HANG είναι πιο διαδεδομένες σε μεγαλύτερες ηλικίες.

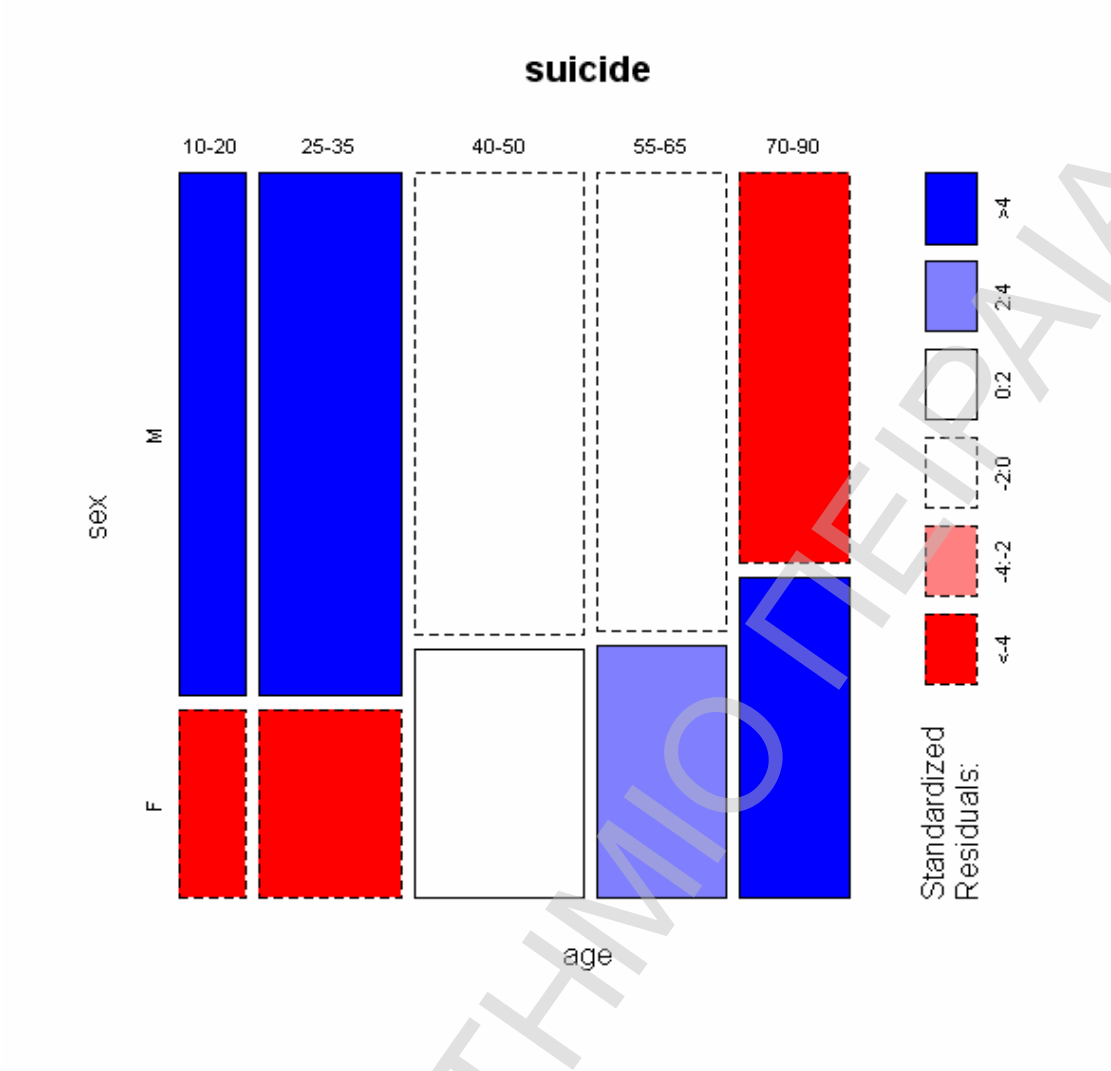
Στη συνέχεια, θα μελετήσουμε το μωσαϊκό του γραφήματος 2.3, όπου εκεί έχουμε αναδιατάξει τις κατηγορίες στην προσπάθειά μας να αναδειχθούν άλλα χρήσιμα συμπεράσματα για τη συσχέτιση των μεταβλητών και τη δομή των δεδομένων που διαθέτουμε. Η διάταξη των μεταβλητών προκύπτει βάσει της ανάλυσης αντιστοιχιών για τις μεταβλητές M και A, την οποία όμως δεν παραθέτουμε εδώ, γιατί θα την μελετήσουμε διεξοδικά και θα τη γενικεύσουμε στο κεφάλαιο 6.



Γράφημα 2.3

Μωσαϊκό αναδιατεταγμένο για τις μεταβλητές Μ, Α.

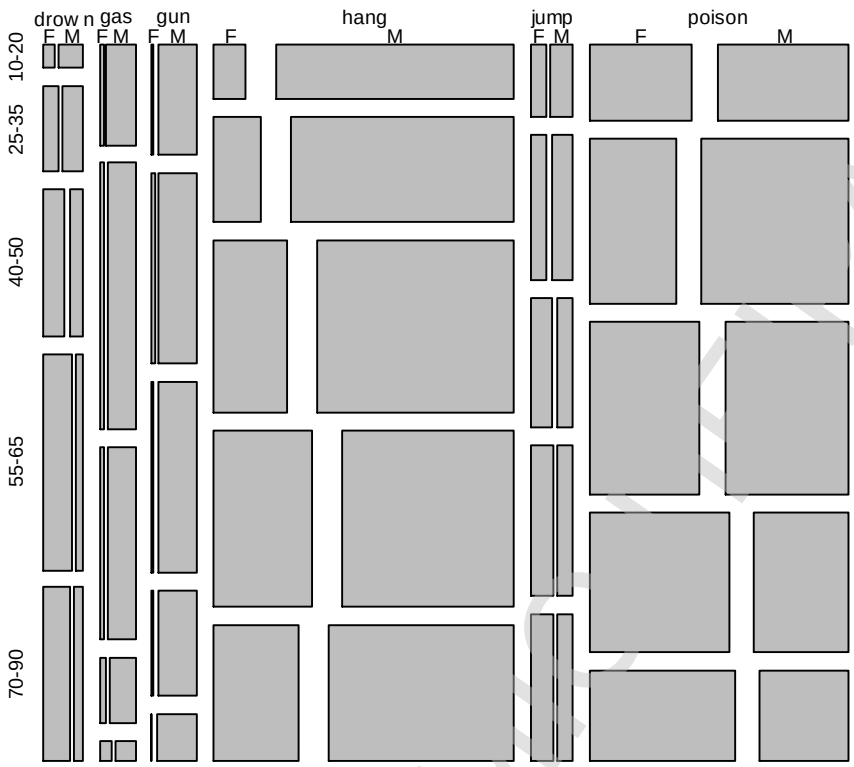
Όπως συμπεραίνουμε από το γράφημα 2.3, η εικόνα πλέον είναι πιο ξεκάθαρη σε σχέση με το προηγούμενο μωσαϊκό του γραφήματος 2.2 και είναι πιο έκδηλη αυτή η τάση που περιγράψαμε καθώς αυξάνει η ηλικία. Ωστόσο, είναι φανερό πως η μέθοδος JUMP διαφοροποιείται σε σχέση με τις άλλες μεθόδους και δεν παρουσιάζει αυτήν την τάση.



Γράφημα 2.4

Μωσαϊκό για τις μεταβλητές A, S, αγνοώντας τη μεταβλητή M.

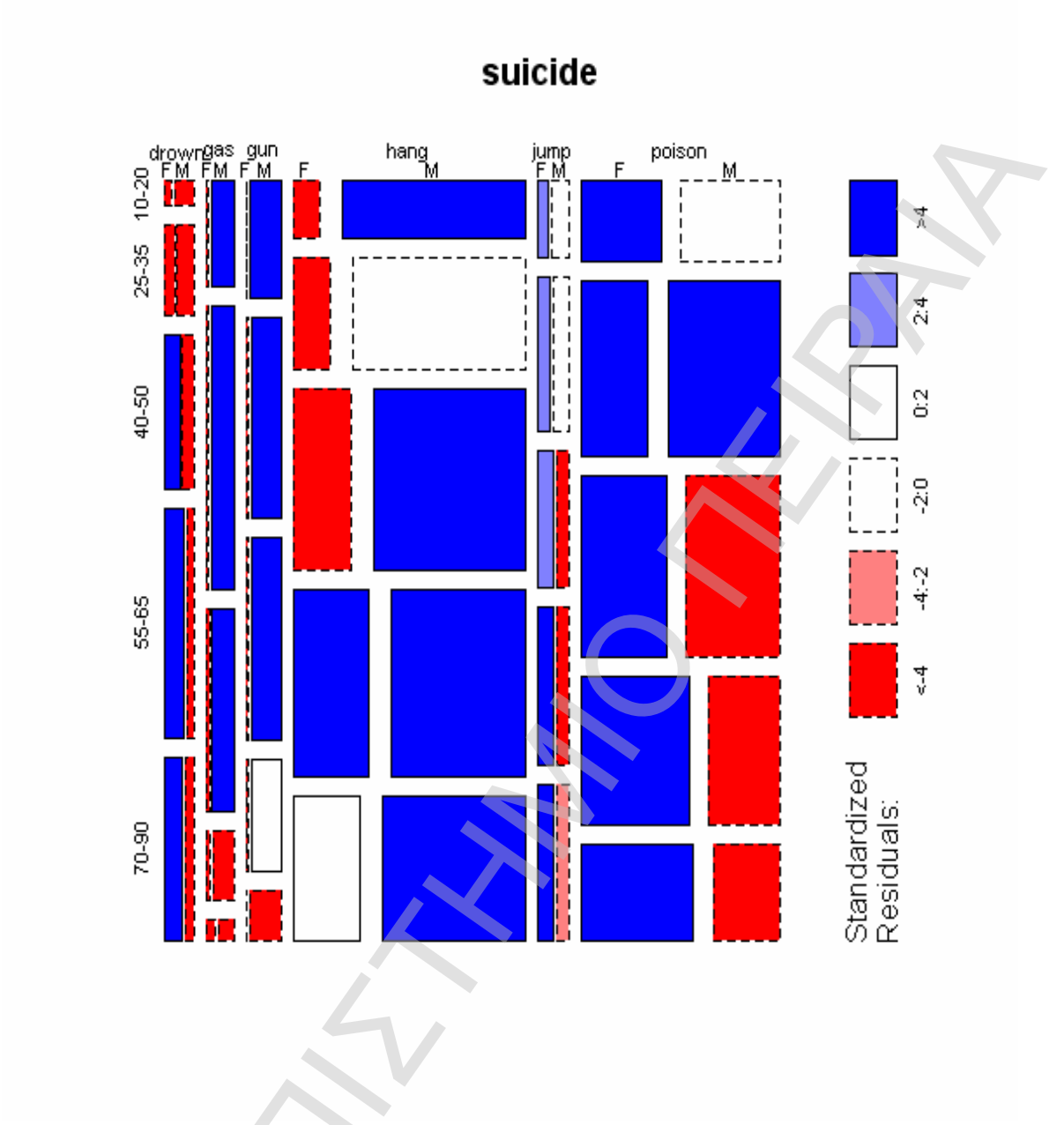
Το γράφημα 2.4 παρουσιάζει την περιθώρια σχέση την μεταβλητών A και S, αγνοώντας την M, δηλαδή διερευνά τη σχέση της ηλικίας και του φύλου των ατόμων που διέπραξαν αυτοκτονία, αγνοώντας τη μέθοδο αυτοκτονίας. Όπως διαπιστώνουμε από το γράφημα αυτό, η αυτοκτονία είναι πιο διαδεδομένη στους άνδρες από ότι στις γυναίκες και, καθώς αυξάνει η ηλικία, η τάση των ανδρών να διαπράξουν αυτοκτονία μειώνεται. Την αντίθετη εικόνα παρουσιάζουν οι γυναίκες.



Γράφημα 2.5

Μωσαϊκό για τις μεταβλητές M, A, S χωρίς χρώμα και σκίαση.

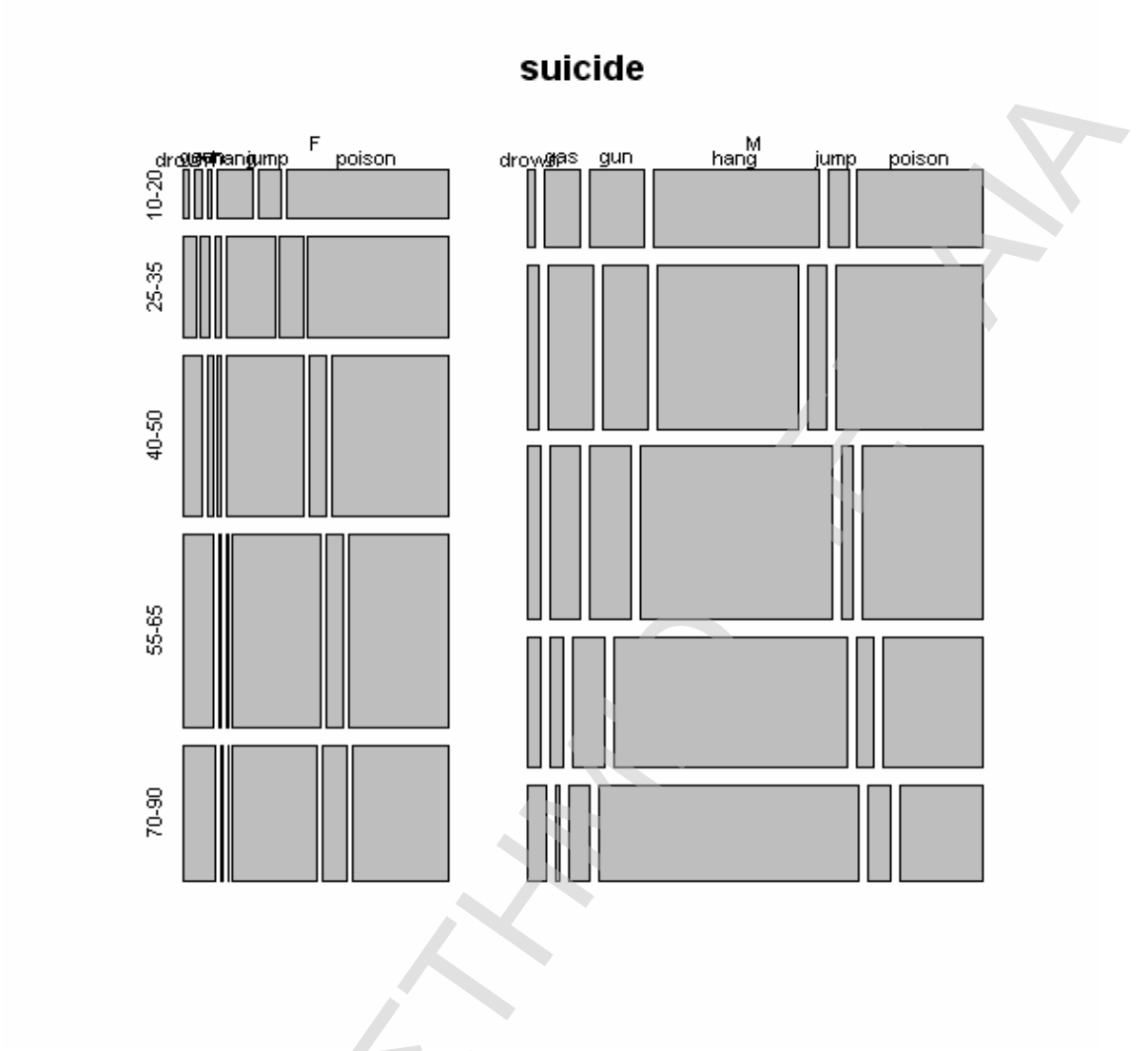
Το γράφημα 2.5 αποτελεί το μωσαϊκό όπου οι μεταβλητές εισήχθησαν με τη σειρά M, A, S, έτσι ώστε να γίνει είναι έκδηλη η επίδραση της μεταβλητής της μεθόδου M σε σχέση με την ηλικία και ως προς το φύλο S και να είναι εύκολη η οπτική σύγκριση των μεθόδων. Όπως εύκολα μπορούμε να διαπιστώσουμε και στα δύο φύλα οι πιο συνηθισμένες μέθοδοι αυτοκτονίας είναι HANG και POISON, γεγονός που θα μελετήσουμε και στη συνέχεια στα μωσαϊκά που ακολουθούν.



Γράφημα 2.6

Μωσαϊκό για τις μεταβλητές M, A, S, όπου φαίνονται και οι αποκλίσεις του μοντέλου ανεξαρτησίας.

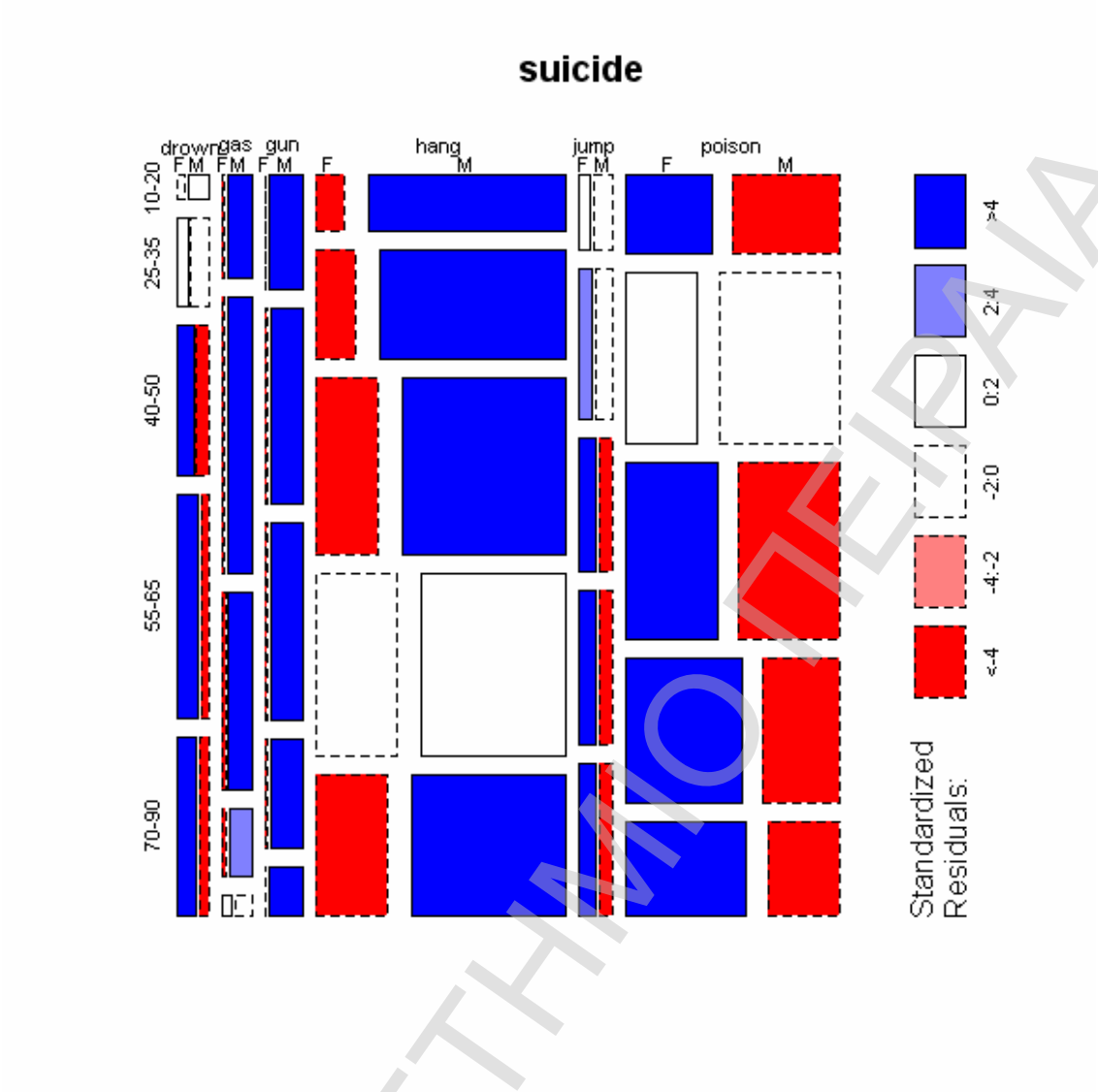
Το γράφημα 2.6 αποτελεί ένα μωσαϊκό και για τις τρεις μεταβλητές M, A, S, αλλά συγχρόνως δείχνει και τις αποκλίσεις από το μοντέλο ανεξαρτησίας. Όπως διαπιστώνουμε από το γράφημα αυτό, οι μέθοδοι POISON και HANG είναι οι πιο διαδεδομένες μέθοδοι και στα δυο φύλα, και κυρίως στους άνδρες, αλλά καθώς αυξάνει η ηλικία η εικόνα αντιστρέφεται. Οι αποκλίσεις από το μοντέλο ανεξαρτησίας φαίνονται με το διαφορετικό χρώμα και τις διακεκομμένες γραμμές.



Γράφημα 2.7

Μωσαϊκό για τις μεταβλητές S, A, M.

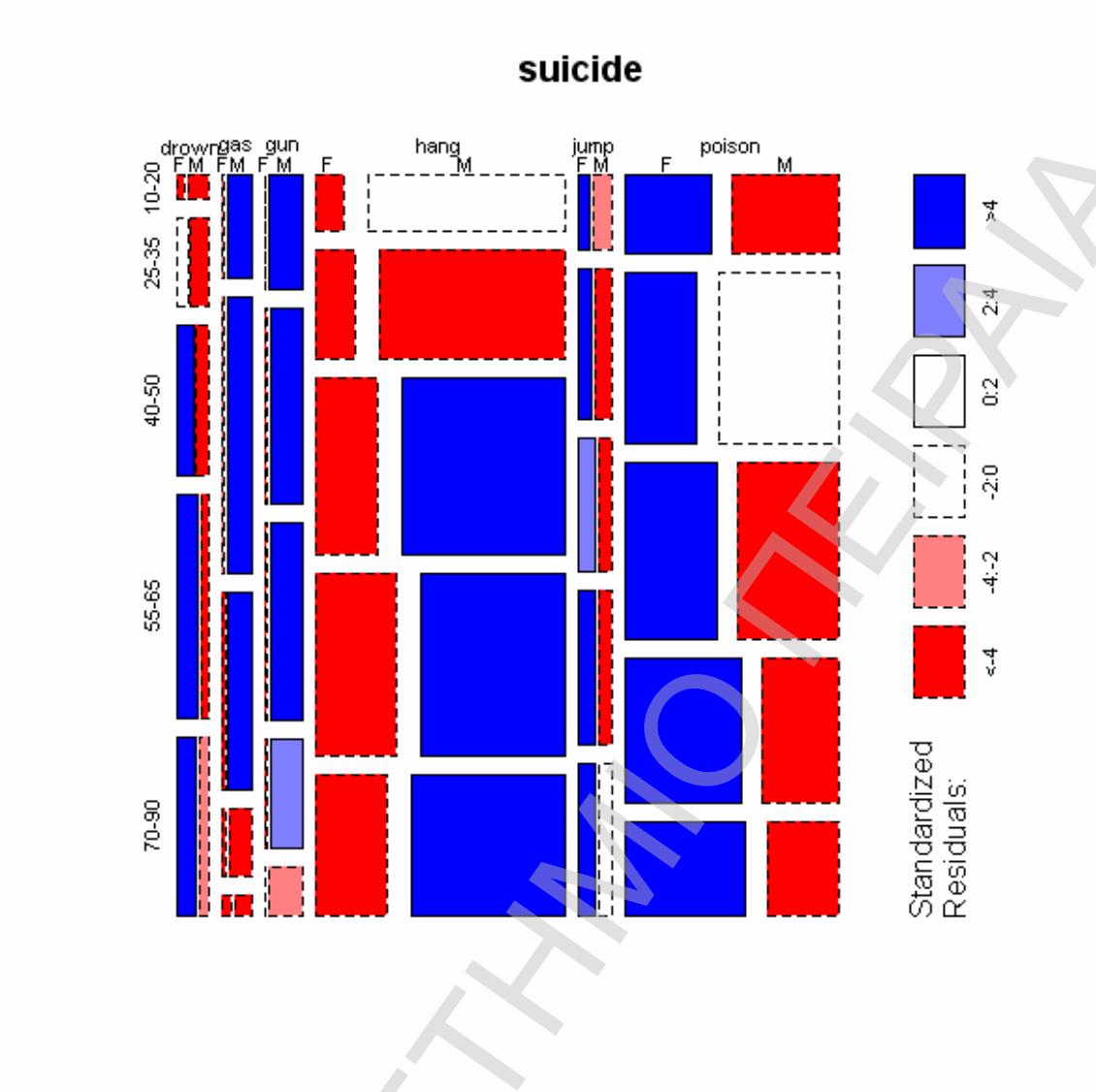
Το γράφημα 2.7 αποτελεί το μωσαϊκό όπου οι μεταβλητές έχουν εισαχθεί με τη σειρά S, A, M, έτσι ώστε να γίνει είναι έκδηλη η επίδραση της μεταβλητής του φύλου S και να είναι εύκολη η οπτική σύγκριση των μεθόδων, αλλά και των ηλικιακών ομάδων για κάθε φύλο. Όπως εύκολα μπορούμε να διαπιστώσουμε για τις γυναίκες, καθώς αυξάνει η ηλικία, αυξάνει και η τάση για αυτοκτονία, εν αντιθέσει με τους άνδρες. Επίσης, και στα δύο φύλα οι πιο συνηθισμένες μέθοδοι αυτοκτονίας είναι οι HANG και POISON και από τις μεθόδους αυτές, η POISON είναι πιο διαδεδομένη για τις γυναίκες σε σχέση με τη HANG, ενώ το αντίστροφο ισχύει για τους άνδρες.



Γράφημα 2.8

Μωσαϊκό για τις μεταβλητές M, A, S, όπου φαίνονται και οι αποκλίσεις του μοντέλου (MA, S).

Το γράφημα 2.8 αποτελεί ένα μωσαϊκό και για τις τρεις μεταβλητές M, A, S, αλλά συγχρόνως δείχνει και τις αποκλίσεις από το υπό συνθήκη (ιεραρχικό) μοντέλο (MA, S). Όπως διαπιστώνουμε από το γράφημα αυτό, οι μέθοδοι GUS και GUN είναι πιο διαδεδομένοι σε μικρότερες ηλικίες και σε άνδρες, ενώ οι μέθοδοι POISON, JUMP και DROWN είναι πιο διαδομένες στις γυναίκες και κυρίως μεγαλύτερης ηλικίας. Οι αποκλίσεις από την πλήρη ανεξαρτησία φαίνονται με το διαφορετικό χρώμα και τις διακεκομμένες γραμμές.



Γράφημα 2.9

Μωσαϊκό για τις μεταβλητές M, A, S, όπου φαίνονται και οι αποκλίσεις του μοντέλου (SA, M).

Το γράφημα 2.9 παρουσιάζει τις αποκλίσεις από την ανεξαρτησία του μοντέλου (SA, M). Οι μέθοδοι έχουν διαταχθεί βάσει της λύσης της ανάλυσης αντιστοιχιών που προαναφέραμε. Πάλι διαπιστώνουμε από το γράφημα αυτό πως οι μέθοδοι GUN και GAS είναι επικρατέστερες σε μικρότερες ηλικίες και κυρίως στους άνδρες και εμφανίζονται λιγότερο καθώς αυξάνει η ηλικία. Για τις γυναίκες, αυτές οι μέθοδοι είναι λιγότερο διαδεδομένες από ότι στην περίπτωση όπου η μέθοδος αυτοκτονίας θα ήταν ανεξάρτητη της ηλικίας και του φύλου. Επίσης, οι μέθοδοι POISON, JUMP και DROWN είναι πιο συνηθισμένες στις γυναίκες.

2.3 Κλασικές παραμετρικές μέθοδοι για κατηγορικά δεδομένα και ανάλυση της σχέσης τους με τα μωσαϊκά

Όπως προαναφέρθηκε στην παράγραφο 2.1, η δομή των μωσαϊκών στηρίζεται στο γεγονός ότι το εμβαδόν του ορθογωνίου είναι ανάλογο της συχνότητας του αντίστοιχου κελιού του πίνακα συνάφειας. Συνεπώς, ένα μωσαϊκό δίνει μια εκτίμηση της κοινής κατανομής όλων των μεταβλητών που περιλαμβάνονται σε αυτό. Το συμπέρασμα αυτό αποτελεί ένα ωραίο σημείο εκκίνησης για άλλες μεθόδους ανάλυσης κατηγορικών δεδομένων και αντίστροφα. Στη συνέχεια, θα μελετήσουμε τη σχέση της ανάλυσης αντιστοιχιών και των λογαριθμογραμμικών μοντέλων με τα μωσαϊκά.

2.3.1 Ανάλυση Αντιστοιχιών

Η ανάλυση αντιστοιχιών είναι μια μαθηματική μέθοδος που προσφέρεται περισσότερο για περιγραφικούς σκοπούς παρά για στατιστική συμπερασματολογία, αφού δεν γίνονται υποθέσεις για τις μεταβλητές. Τα αποτελέσματα της ανάλυσης αντιστοιχιών χρησιμοποιούνται συχνά για την απεικόνιση των κατηγοριών των μεταβλητών ταξινόμησης ως σημεία στο χώρο και όπως είδαμε, συμβάλλουν στην μελέτη της δομής τους και της καλύτερης απεικόνισής τους με τη βοήθεια των μωσαϊκών.

Η εφαρμογή της ανάλυσης αντιστοιχιών είναι άμεση για διδιάστατους πίνακες συνάφειας, αλλά για να ξεπεραστεί ο περιορισμός των διδιάστατων πινάκων συνάφειας, αναπτύχθηκε η πολυδιάστατη (ή πολυμεταβλητή) ανάλυση αντιστοιχιών (βλέπε Nagel, 1996, κ.α.). Όπως προαναφέρθηκε όμως, στη μέθοδο αυτή θα αναφερθούμε εκτενέστερα στο κεφάλαιο 6.

2.3.2 Λογαριθμογραμμικά μοντέλα

Τα λογαριθμογραμμικά μοντέλα (βλέπε Agresti, Fienberg) προκύπτουν άμεσα από τα γραμμικά μοντέλα. Ενώ η ανάλυση αντιστοιχιών απεικονίζει την δομή της αλληλεπίδρασης των μεταβλητών, τα λογαριθμογραμμικά μοντέλα ορίζονται από τη δομή της αλληλεπίδρασης. Η καταλληλότητα του μοντέλου κρίνεται από το στατιστικό X^2 ή G^2 καλής προσαρμογής.

Παρόλο που η μοντελοποίηση των κατηγορικών δεδομένων μέσω των λογαριθμογραμμικών μοντέλων είναι κομψή και βολική, δεν έχει γίνει μια ικανοποιητική προσπάθεια ακόμη ώστε να αναπαρασταθούν κατάλληλα με γραφήματα. Ένα διάγραμμα διασποράς (*scatterplot*) των παρατηρούμενων τιμών έναντι των αναμενόμενων τιμών χρησιμοποιείται συχνά για αναπαράσταση, αλλά δεν ενσωματώνεται καμία δομή στα δεδομένα, αλλά και καμία δομή του μοντέλου. Επίσης, αυτό ισχύει και για τα γραφήματα των

καταλοίπων(*residuals plots*). Στην πράξη, συχνά χρησιμοποιούνται αυτόματοι μέθοδοι επιλογής μοντέλων, αλλά συνήθως δεν μπορούν να αποκαλύψουν την πραγματική πληροφορία ενός συνόλου δεδομένων.

Στην προσπάθεια να αποδοθεί γραφικά ένα λογαριθμογραμμικό μοντέλο, έχουν διατυπωθεί τρεις προσεγγίσεις:

i. να απεικονιστούν οι αναμενόμενες τιμές ενός μοντέλου σε ένα μωσαϊκό.
Τυπικές δομές εξάρτησης για δύο ή τρεις μεταβλητές έχουν συγκεκριμένο σχήμα, οι οποίες μπορεί να κατανοηθούν πολύ εύκολα και δεν εξαρτώνται από τα εν λόγω κάθε φορά δεδομένα.

ii. να προστεθεί η πληροφορία για τα κατάλοιπα του μοντέλου πάνω σε ένα γράφημα των μη επεξεργασμένων δεδομένων.
Αυτή η προσέγγιση πραγματοποιήθηκε από τον Friendly(1994). Παρόμοια προσέγγιση έγινε και από τους Riedwyl και Schuepbach (1994). Ωστόσο, αυτή η προσέγγιση έχει πολλά μειονεκτήματα. Με το να αναπαριστώνται τα κατάλοιπα σε ένα γράφημα των μη επεξεργασμένων δεδομένων μπορεί να αποδειχθεί παραπλανητικό, αφού άδεια ή κελιά με πολύ μικρές συχνότητες δεν είναι ορατά στο γράφημα, το οποίο δεν συμβαίνει σε γράφημα των δεδομένων του μοντέλου. Αυτό οδηγεί στην τρίτη προσέγγιση (Theus&Wilhelm(1996a)).

iii. να προστεθεί η πληροφορία για τα κατάλοιπα του μοντέλου σε ένα γράφημα των δεδομένων του μοντέλου.

Για το σκοπό αυτό, παρατίθεται ο παρακάτω αλγόριθμος που διατυπώθηκε από τον M.Theus(1997), ο οποίος μπορεί να χρησιμοποιηθεί για την απεικόνιση ενός λογαριθμογραμμικού μοντέλου και να ενσωματωθεί σε αυτήν και η πληροφορία για τις τιμές των καταλοίπων:

a. Απεικόνισε την πληροφορία των καταλοίπων του μωσαϊκού του αντίστοιχου μοντέλου.
b. Χρησιμοποίησε χρώμα (κόκκινο ή μπλε) για το πρόσημο (θετικό ή αρνητικό) του σφάλματος: $r_i = \frac{o_i - e_i}{\sqrt{e_i}}$ (απόκλιση), όπου o_i η παρατηρούμενη τιμή και e_i η αναμενόμενη τιμή της συχνότητας του κελιού του πίνακα συνάφειας.
c. Απέδωσε σκορ για όλα τα κατάλοιπα με βάση το μεγαλύτερο κελί, π.χ. υπολόγισε $r_i^* = \frac{r_i}{\max(|r_i|)}$.

d. Απέδωσε σκορ για όλα τα κατάλοιπα με το α-ποσοστημόριο του X^2 στατιστικού. Αυτή η κλιμάκωση ενσωματώνει το αντίστοιχο p-value του μοντέλου στο γράφημα, το οποίο μπορεί να εντοπιστεί από το κελί με το μεγαλύτερο κατάλοιπο. Ένα μοντέλο είναι σημαντικό σε επίπεδο p, αν τουλάχιστον ένα κελί τονίζεται από ένα κατάλοιπο σε περισσότερο από (100-p)%.
e. Χρησιμοποίησε διαφορετικούς τόνους στο χρώμα για να τονιστεί το πόσο μεγάλη ή μικρή είναι η τιμή των καταλοίπων.
f. Ύψωσε την τιμή όλων των καταλοίπων στη δύναμη β:
   - αν $\beta \geq 2$, μόνο κακή προσαρμογή είναι οπτικά σημαντική.
   - αν $\beta \leq 2$, η δομή των καταλοίπων μπορεί να φανεί καθαρά.

Στο σημείο αυτό, πρέπει να τονίσουμε ότι στο παράδειγμα της παραγράφου 2.3, χρησιμοποιήθηκε η τρίτη προσέγγιση.

Κλείνοντας, οφείλουμε να σημειώσουμε ότι η δομή των δεδομένων που υπόκεινται ενός γραφήματος πρέπει αντιμετωπιστεί με τον ίδιο τρόπο όπως ένα παραδοσιακό στατιστικό μοντέλο και η σύνδεση μεταξύ του μοντέλου και της αναπαράστασής του να καθίσταται εύκολη, όπως στα ιστογράμματα και τα διαγράμματα διασποράς (***scatterplots***). Για πιο περίπλοκα γραφήματα, ωστόσο, όπως τα μωσαϊκά, αυτή η τεχνική βοηθά να κατανοήσουμε τη φύση των μεταβλητών που απεικονίζονται σε αυτά. Πίνακες συνάφειας πολλαπλής εισόδου δομούνται ώστε να συλλάβουν την ουσιαστική μετα-πληροφορία που είναι αναγκαία για την απεικόνιση ενός συνόλου κατηγορικών μεταβλητών.

Τελικά, με το να περιγραφεί ένα μοντέλο, καθιστά ικανό τον ερευνητή ώστε να αρπάξουμε τον κοινό σύνδεσμο μεταξύ διαφορετικών μεθόδων απεικόνισης κατηγορικών δεδομένων και με αυτό τον τρόπο αποκτάει τη δυνατότητα να επιλέξει την καλύτερη αναπαράσταση για ένα συγκεκριμένο σκοπό.

2.4 Βασικές ιδιότητες των στατιστικών γραφημάτων

Τα στατιστικά γραφήματα οφείλουν να παρουσιάζουν τις παρακάτω βασικές ιδιότητες:

i. *Ικανότητα για γενίκευση*: ο σχεδιασμός ενός γραφήματος πρέπει να είναι γενικεύσιμος για περισσότερες από τις μεταβλητές για τις οποίες σχεδιάστηκε.
ii. *Συνέπεια*: τα δεδομένα που μετρώνται στην ίδια κλίμακα πρέπει να απεικονίζονται με τον ίδιο τρόπο, π.χ. οι μετρήσεις με εμβαδά, σημεία σε συνεχή κλίμακα με τελείες, κλπ.
iii. *Επεκτασιμότητα*: ο βασικός σχεδιασμός του γραφήματος πρέπει να επιτρέπει την επέκταση του γραφήματος ώστε να χρησιμοποιηθεί με διαφορετικούς τρόπους. Για παράδειγμα, ο

διαφορετικός χρωματισμός των επιμέρους τμημάτων δίνει πληροφορίες και για τα κατάλοιπα ή άλλες πληροφορίες του μοντέλου.

iv. *Αλληλεπίδραση*: η λειτουργικότητα ενός γραφήματος που παράγεται από ένα σύγχρονο στατιστικό πακέτο πρέπει να φθάνει πέραν ενός απλού σχεδιασμού.

Επιπλέον, πρέπει να αναφέρθουν τα εξής για τη σωστή χρήση των γραφημάτων:

- Τα γραφήματα πρέπει να είναι συνδεδεμένα και να επιτρέπουν διαχείριση των επιλεγμένων δεδομένων.
- Ο χρήστης πρέπει να διαθέτει την ικανότητα να αναζητά πληροφόρηση από τα γραφήματα.
- Η παραμετρικοποίηση του γραφήματος πρέπει να είναι εύκολο να αλλάζει δυναμικά.

Τα μωσαϊκά ικανοποιούν όλες τις παραπάνω απαιτήσεις για τα στατιστικά γραφήματα:

i. Ο επαναληπτικός ορισμός επιβεβαιώνει μια γενίκευση σε όσες το δυνατόν μεταβλητές έχει νόημα.
ii. Το μωσαϊκό είναι συνεπές, γιατί όλες οι μετρήσεις αναπαρίστανται από το εμβαδόν των ορθογωνίων.
iii. Ο εντοπισμός των υποομάδων είναι εύκολο να γίνει, αφού μια υποομάδα προστίθεται, με τον ίδιο τρόπο που θα προσθέτονταν μια νέα δυαδική μεταβλητή. Αυτό επιτρέπει και την απεικόνιση των καταλοίπων του μοντέλου.
iv. Η ικανότητα για αλληλεπίδραση μπορεί να επιτευχθεί με τα ίδια εργαλεία που χρησιμοποιούνται στα ραβδογράμματα. Τέλος, παρέχονται ακόμη, προαιρετικά εργαλεία για ευέλικτη επαναδιάταξη των μεταβλητών και των κατηγοριών.

2.5 Μαθηματική θεμελίωση των μωσαϊκών

Οι πίνακες συνάφειας πολλαπλής εισόδου είναι η ιεραρχική μέθοδος ώστε να αναπαρίσταται η πληροφορία από τα δεδομένα. Διαφορετική διάταξη και διαφορετικές κατευθύνσεις των μεταβλητών μπορεί να υποστηρίξει συγκρίσεις μεταξύ διαφορετικών συνδυασμών. Τα μωσαϊκά αποτελούν τη γραφική ισοδυναμία των πινάκων συνάφειας.

Ο H.Hoffman για πρώτη φορά το 2003 παρουσίασε έναν επίσημο ορισμό και μια περιγραφή των μωσαϊκών, στηριζόμενος σε εργαλεία γνωστά από την μαθηματική περιγραφή των πινάκων συνάφειας και διαδικασιών μοντελοποίησης. Η μαθηματική θεμελίωση επιτρέπει τη σύλληψη και την περιγραφή των ιδιοτήτων των μωσαϊκών.

I. Μονοδιάστατες δομές.

Ορισμός 2.1 (Εικόνα(*image, co-domain*) μιας μεταβλητής)

Έστω X μια κατηγορική μεταβλητή με K κατηγορίες. Η εικόνα της μεταβλητής X είναι ένα σύνολο που περιλαμβάνει όλες τις κατηγορίες της X, έστω $c(X)=\{c_1, c_2, \ldots c_K\}$.

Ορισμός 2.2(Δομή μιας μεταβλητής)

Έστω $c(X)=\{c_{(1)}, c_{(2)}, \ldots c_{(K)}\}$ η διάταξη των κατηγοριών της μεταβλητής X. Η μονοδιάστατη δομή της κατηγορικής μεταβλητής X με K κατηγορίες είναι ένα K-διάστατο διάνυσμα από δείκτριες συναρτήσεις, έστω:

$s_X: c(X) \rightarrow \{0, 1\}^K$, $s_X(y)=\{1_{\{c_{(1)}\}}(y), \ldots, 1_{\{c_{(K)}\}}(y)\}$.

Έστω X^i η i-παρατήρηση της X, οπότε $s_X(X^i)=(0\ldots0\ \ 1\ \ 0\ \ldots)$, αν $X^i=c_{(j)}$, $1 \le j \le K$.

↑

j

Συμβολίζουμε με $s_X(X)$ τον πίνακα των βουβών (*dummy*) μεταβλητών της X: $s_X(X)=\sum_{i=1}^{n} e_i s_X(X^i)$, όπου $e_i \in R^n$ είναι το i-οστό διάνυσμα στήλη της κανονικής βάσης του R^n.

Ορισμός 2.3(Απόλυτες και σχετικές συχνότητες)

Έστω X μια κατηγορική μεταβλητή με n παρατηρήσεις και δομή s_X. Οι απόλυτες και σχετικές συχνότητες των κατηγοριών της X μπορούν να γραφούν ως εξής: $\sum_{i=1}^{n} s_X(X^i)$ και $\frac{1}{n}.\sum_{i=1}^{n} s_X(X^i)$.

ΙΙ. Πολυδιάστατες δομές.

Ορισμός 2.4 (Γινόμενο Kronecker)

Έστω A και B nxk και mxr πίνακες. Το γινόμενο $C=A \otimes B$ είναι το γινόμενο Kronecker των A και B, όπου C είναι nmxkr πίνακας.

Στη συνέχεια, το γινόμενο Kronecker χρησιμοποιείται στη μορφή $\upsilon \otimes A$, όπου A είναι πίνακας, αλλά το υ είναι διάνυσμα γραμμή ή στήλη. Στην περίπτωση αυτή, το γινόμενο Kronecker απλοποιείται στη μορφή:

$\upsilon \otimes A=\begin{pmatrix} u_1 A \\ u_2 A \\ . \\ . \\ u_n A \end{pmatrix}$ και $w \otimes A=(w_1 A, w_2 A, \ldots, w_n A)$, με $\upsilon, w \in R^n$. Χρησιμοποιώντας το γινόμενο Kronecker, μπορούμε να ορίσουμε δομές μιας αυθαίρετης διάστασης από τις μονοδιάστατες δομές.

Ορισμός 2.5(Δομή των p μεταβλητών)

Έστω $X_1, X_2, \ldots, X_p$ p κατηγορικές μεταβλητές και $X_{(1)}, X_{(2)}, \ldots, X_{(p)}$ μια διάταξη των μεταβλητών αυτών. Η δομή των $X_1, X_2, \ldots, X_p$ είναι μια έκφραση S της παρακάτω μορφής: $S(y_1, y_2, \ldots, y_p)=h_{1(y1)} \otimes h_{2(y2)} \otimes \ldots \otimes h_{p(yp)}$, όπου $h_j= s_{X_{(J)}}$ ή $s^t{}_{X_{(J)}}, 1 \le j \le p$. Το σύνολο όλων των p-διάστατων δομών των $X_1, X_2, \ldots, X_p$ αποτελείται από όλες τις δυνατές εκφράσεις S όπως παραπάνω.

Ορισμός 2.6(Δομημένος πίνακας συνάφειας)

Ένας δομημένος πίνακας συνάφειας p-οστής εισόδου αποτελείται από p κατηγορικές μεταβλητές $X_1, X_2, \ldots, X_p$ και μια δομή S. Αν οι μεταβλητές έχουν n παρατηρήσεις, ο πίνακας συνάφειας μπορεί να γραφεί ως εξής: $\sum_{i=1}^{n} S(X_{(1)}{}^i, \ldots, X_{(p)}{}^i)$. Αν W είναι μια μεταβλητή στάθμισης, τότε ο αντίστοιχος σταθμισμένος πίνακας συνάφειας είναι: $\sum_{i=1}^{n} W^i S(X_{(1)}{}^i, \ldots, X_{(p)}{}^i)$.

Η δομή των p μεταβλητών $X_1, X_2, \ldots, X_p$ είναι ένας πίνακας με δείκτριες συναρτήσεις, όπου ο αριθμός των γραμμών και των στηλών δίνεται από τον αριθμό των διανυσμάτων γραμμής ή στήλης στην αντίστοιχη έκφραση. Αν η S είναι δομημένη όπως περιγράφηκε παραπάνω και $s_{X_j}, j \in J \subset \{1, \ldots, p\}$ είναι τα διανύσματα στήλη στην αντίστοιχη έκφραση, ο αριθμός των στηλών του S δίνεται ως το γινόμενο των κατηγοριών του: $\prod_{j \in J} \left| c(X_j) \right|$ (Ομοίως ορίζεται και για τον αριθμό των γραμμών).

Ορισμός 2.7 (Επίσημος ορισμός του μωσαϊκού)

Ένα μωσαϊκό των μεταβλητών X_1, X_2, …, X_p είναι μια συνάρτηση, της οποίας η ανεξάρτητη μεταβλητή είναι ένας δομημένος πίνακας συνάφειας πολλαπλής εισόδου και η εξαρτημένη μεταβλητή είναι ένα γράφημα, το οποίο ακολουθεί την εξής διαδικασία δόμησης(με τη μορφή αλγορίθμου):

Διαδικασία 2.8(Δόμηση μωσαϊκού)

Έστω $h_1 \otimes h_2 \otimes \ldots \otimes h_p$, η δομή των μεταβλητών X_1, X_2, …, X_p.

(i)Το σημείο εκκίνησης είναι ένα ορθογώνιο με μήκος w και πλάτος h. Το εμβαδόν του καλείται ως αποθήκη(***bin***). Στην αρχή δεν υπάρχει συνθήκη ώστε οι τιμές των μεταβλητών X_1, X_2, …, X_p να «ρίχνονται» στην αποθήκη. Θέτουμε k=1 και προχωρούμε στο βήμα (ii).

(ii)Αν k=p+1, σταμάτα. Αλλιώς, για την μεταβλητή $X_{(k)}$ υπολόγισε το μέγεθος και την αναλογία ενός νέου συνόλου από αποθήκες από τις διαστάσεις της παλαιότερης αποθήκης: αν h_k είναι ένα διάνυσμα γραμμή, η παλαιότερη αποθήκη διαιρείται οριζόντια, αλλιώς διαιρείται κάθετα. Χωρίς βλάβη της γενικότητας, έστω ότι το h_k είναι ένα διάνυσμα γραμμή. Αφαιρείς από την υπο-δομή $h_1 \otimes h_2 \otimes \ldots \otimes h_K$ του S το διάνυσμα από τις δείκτριες συναρτήσεις, που ικανοποιούν τη συνθήκη από την παλιά αποθήκη. Αυτό οδηγεί σε ένα διάνυσμα γραμμή διάστασης K, αν το $X_{(K)}$ έχει K κατηγορίες. Αφαιρείς από τον πίνακα συνάφειας που δίνεται από τη δομή $h_1 \otimes h_2 \otimes \ldots \otimes h_K$ τα κελιά που αντιστοιχούν στο διάνυσμα των παραπάνω δεικτρίων συναρτήσεων. Έστω ότι είναι τα κελιά $n_{(1)}$, …, $n_{(K)}$. Η θέση p_j (πάνω αριστερή γωνία) και το μέγεθος της αποθήκης που αντιστοιχούν στην j-οστή κατηγορία του $X_{(K)}$ υπολογίζεται ως εξής:

- τιμή y του p_j=τιμή της πάνω αριστερής γωνίας της προηγούμενης αποθήκης, τιμή x του p_j=άθροισμα των μηκών του j-1 προηγούμενων αποθηκών που αντιστοιχούν στις j-1 προηγούμενες κατηγορίες.
- το πλάτος της νέας αποθήκης είναι το πλάτος της παλαιότερης αποθήκης, το πλάτος της νέας αποθήκης υπολογίζεται ως εξής:

  (νέο πλάτος)=(παλαιό πλάτος). $\frac{n_{(j)}}{\sum_{i=1}^{K} n_i}$

(iii)Θέτω k=k+1. Για κάθε κατηγορία $c_{(j)}$ του $X_{(k)}$ περιόρισε την παλιά συνθήκη με την επιπλέον συνθήκη $y_k=c_{(j)}$. Πάρε τη νέα συνθήκη και την αποθήκη που αντιστοιχεί στην κατηγορία $c_{(j)}$ και επέστρεψε στο βήμα (ii).

Στο σημείο αυτό, πρέπει να σημειώσουμε πως είναι προφανές ότι η τάξη στην οποία η μεταβλητή εμφανίζεται στη δομή S είναι κρίσιμη, αν θεωρηθούν παραπάνω από δύο μεταβλητές. Αυτό σημαίνει πως διαφορετικές διατάξεις σε πίνακες τριπλής εισόδου και άνω υποστηρίζουν διαφορετικές συγκρίσεις, όπως προαναφέρθηκε στην παράγραφο 2.1.2.

2.6 Χρήσιμα συμπεράσματα για τα μωσαϊκά

Όπως αναφέρθηκε και στην εισαγωγή, πρώτοι οι Hartigan και Kleiner το 1981 πρότειναν τα μωσαϊκά. Παρόλο που αυτή η γραφική απεικόνιση είναι αρκετά δυναμική και περιεκτική, δεν αποδείχθηκε και ιδιαίτερα δημοφιλής. Αυτό οφείλεται κυρίως στο γεγονός πως ο οπτικός αντίκτυπος των μωσαϊκών εξαρτάται αρκετά από την διάταξη των μεταβλητών. Υπάρχουν αρκετές εφαρμογές και προγράμματα που περιλαμβάνουν την υλοποίηση μωσαϊκών, όπως τα υπολογιστικά πακέτα SAS, S-PLUS, R, κλπ. αλλά δεν μπορούν να ξεπεράσουν τα μειονεκτήματα που προαναφέρθηκαν. Για επεξηγηματικούς λόγους, με τη βοήθεια των μωσαϊκών μπορεί να δοθεί πληροφορία για τα κατάλοιπα, καθιστώντας με αυτόν τον τρόπο δυνατή μια γραφική βήμα προς βήμα (***stepwise***) μοντελοποίηση των κατηγορικών δεδομένων, που φτάνει πολύ μακριά από τις παραδοσιακές μεθόδους.

Αν και έχουν αναπτυχθεί πολλές γραφικές μέθοδοι για την διερεύνηση αλλά και τη μοντελοποίηση συνεχών δεδομένων, κάτι αντίστοιχο δεν συμβαίνει γενικότερα στο πεδίο των κατηγορικών δεδομένων. Θεωρείται ότι υπάρχει έλλειψη σύγχρονων και αποδοτικών μεθόδων γραφικής απεικόνισης των κατηγορικών δεδομένων. Οι παραμετρικοί μέθοδοι χρησιμοποιούνται για ποσοτικά συμπεράσματα, αλλά είναι αδύναμες να εισχωρήσουν στα δεδομένα. Αυτό περιορίζει υπερβολικά την ανάλυση δεδομένων και τα διαγνωστικά του μοντέλου.

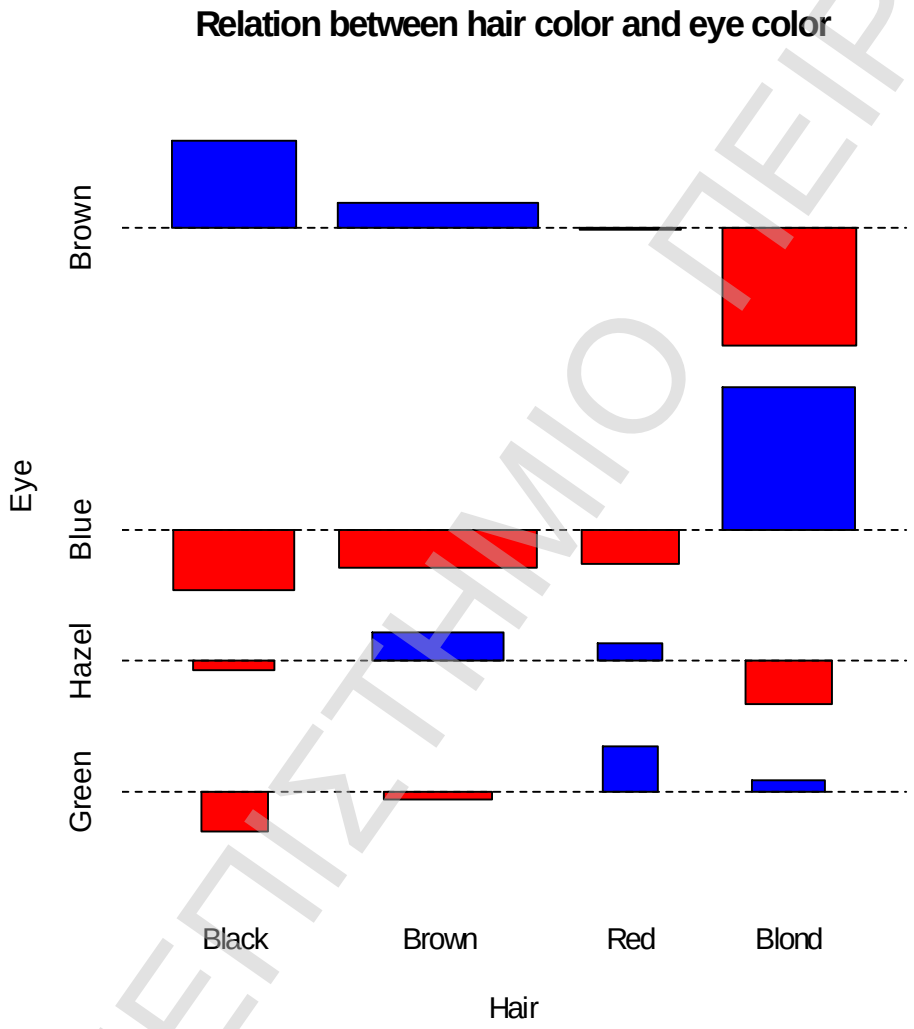
2.7 Γραφικές αναπαραστάσεις συνάφειας

Οι γραφικές αναπαραστάσεις συνάφειας (***association plots***) αποτελούν μια τεχνική απόδοσης γραφικά της υπόθεσης της ανεξαρτησίας σε πίνακες διπλής εισόδου και μπορούν να επεκταθούν και για πίνακες πολλαπλής εισόδου. Όπως προαναφέρθηκε, η στατιστική συνάρτηση X^2 καλής προσαρμογής χρησιμοποιείται ως ένα μέτρο της συνάφειας των μεταβλητών, αλλά τα μωσαϊκά και οι γραφικές αναπαραστάσεις συνάφειας υποστηρίζουν επίσης αυτήν την προσπάθεια ως γραφικές μέθοδοι απεικόνισης.

Οι γραφικές αναπαραστάσεις συνάφειας αποδίδουν γραφικά τα κατάλοιπα Pearson: κάθε κελί αναπαρίσταται με ένα ορθογώνιο του οποίου το ύψος είναι ανάλογο του αντίστοιχου καταλοίπου Pearson και το πλάτος είναι ανάλογο της τετραγωνικής ρίζας των αναμενόμενων

συχνοτήτων, έτσι ώστε το εμβαδόν του ορθογωνίου αυτού να είναι ανάλογο των τιμών των μη επεξεργασμένων καταλοίπων. Το πρόσημο και το μέγεθος των καταλοίπων αυτών αποδίδεται με σκίαση και χρώμα των ορθογωνίων, όπως ακριβώς στα μωσαϊκά.

Το γράφημα 2.10 αποτελεί την γραφική αναπαράσταση συνάφειας για δύο μεταβλητές του παραδείγματος 2.1.1 και μπορεί να γενικευθεί και για περισσότερες από δύο μεταβλητές.



Γράφημα 2.10

Γράφημα συνάφειας για τις μεταβλητές Hair και Eye του παραδείγματος 2.2.1.

Η γραφική αναπαράσταση για τρεις μεταβλητές αποτελεί επέκταση του γραφήματος συνάφειας για δύο μεταβλητές και πρόκειται για έναν πίνακα από γραφικές αναπαραστάσεις συνάφειας του οποίου κάθε κελί περιέχει μια γραφική αναπαράσταση συνάφειας για τις

μεταβλητές της αντίστοιχες γραμμής και στήλης. Αυτού του είδους η απεικόνιση επιτρέπει την γρήγορη και εύκολη μελέτη των γραφικών αναπαραστάσεων συνάφειας για όλους τους δυνατούς συνδυασμούς των μεταβλητών που διαθέτουμε και να εξάγουμε συμπεράσματα για την υπό συνθήκη ανεξαρτησία των μεταβλητών.

2.8 Χρήσιμα συμπεράσματα για τις γραφικές αναπαραστάσεις συνάφειας

Η εργασία πάνω στα γραφικές αναπαραστάσεις συνάφειας για πίνακες πολλαπλής εισόδου πραγματοποιήθηκε κυρίως από τους D.Meyer, A.Zeileis, K.Hornik(2003), αλλά θεωρείται ότι πρέπει να γίνει περισσότερη εργασία πάνω σε αυτά, γιατί με τη συγκεκριμένη μορφή παρουσιάζουν αρκετά μειονεκτήματα, όπως για παράδειγμα, δεν έχουν όλες οι υπό συνθήκη γραφικές αναπαραστάσεις συνάφειας την ίδια κλίμακα μέτρησης, όποτε τα κατάλοιπα από αυτά τα υπογραφήματα δεν είναι άμεσα συγκρίσιμα. Επίσης, το χρώμα επιτρέπει μόνο τον εντοπισμό του μοτίβου της ανεξαρτησίας, ενώ ιδανικά θα έπρεπε το πρόσημο και η τιμή των καταλοίπων να οδηγούσαν στην απόρριψη ή αποδοχή της υπόθεσης της ανεξαρτησίας.

2.9 Κατασκευή των μωσαϊκών και των γραφικών αναπαραστάσεων συνάφειας με τη βοήθεια εντολών της R

Στην παρούσα εργασία, τα μωσαϊκά και οι γραφικές αναπαραστάσεις συνάφειας κατασκευάστηκαν με τη χρήση εντολών της R, οι οποίες είναι διαθέσιμες στο αντίστοιχο πακέτο *vcd*.

Για την κατασκευή των μωσαϊκών, χρησιμοποιήθηκε η εντολή ***mosaicplot***(). Επίσης, για την κατασκευή των γραφημάτων συνάφειας, χρησιμοποιείται η εντολή ***assocplot***(), ενώ για την επέκτασή τους σε γραφήματα συνάφειας για περισσότερες από δύο μεταβλητές χρησιμοποιείται η εντολή ***assoc***(). Όλες οι παραπάνω εντολές εφαρμόζονται σε πίνακες συνάφειας είτε διπλής εισόδου είτε πολλαπλής εισόδου.

ΚΕΦΑΛΑΙΟ 3^{ο}

Μοντέλα Συνάφειας για περισσότερες από δύο Κατηγορικές Μεταβλητές

3.0 Εισαγωγή

Η ανάλυση πινάκων συνάφειας για δύο μεταβλητές μπορεί να πραγματοποιηθεί με τη βοήθεια των λογαριθμογραμμικών μοντέλων, ιεραρχικών και μη ιεραρχικών, αλλά και με τη βοήθεια της ανάλυσης αντιστοιχιών (***correspondence analysis***), που όπως είναι άλλωστε γνωστό, χρησιμεύει για την γεωμετρική κυρίως ερμηνεία της δομής και των σχέσεων των δύο μεταβλητών. Ωστόσο, η ανάλυση πινάκων συνάφειας για τρεις μεταβλητές και άνω, καθίσταται δύσκολη με τη χρήση των λογαριθμογραμμικών μοντέλων (ιεραρχικών και μη). Για την γενίκευση της ανάλυσης αντιστοιχιών σε περισσότερες διαστάσεις, θα αναφερθούμε εκτενέστερα στο κεφάλαιο 6. Στην παρούσα ανάλυση, θα περιοριστούμε σε πίνακες συνάφειας για τρεις μεταβλητές.

Καταρχήν, η δυσκολία στη χρήση λογαριθμογραμμικών μοντέλων για τρεις μεταβλητές έγκειται σε δύο λόγους. Πρώτον, το μοντέλο χωρίς όρους αλληλεπίδρασης οποιασδήποτε τάξης δεν προσαρμόζεται καλά στα δεδομένα μας. Δεύτερον, το μοντέλο της υπό συνθήκη ανεξαρτησίας δεν προσαρμόζεται καλά στα δεδομένα και το μοντέλο χωρίς τον όρο αλληλεπίδρασης τρίτης τάξης που προσαρμόζεται σε αυτά είναι το πιο οικονομικό (***parsimonious***) μοντέλο. Το τελευταίο σχόλιο βέβαια ισχύει και για περισσότερες από τρεις μεταβλητές. Τέλος, οφείλουμε να τονίσουμε ότι το πρόβλημα επιλογής βέλτιστου μοντέλου τελικά είναι επίπονη εργασία, διότι καθώς μεγαλώνει η διάσταση του πίνακα συνάφειας, αυξάνουν και οι δυνατές επιλογές μοντέλων. Για αυτό το λόγο, στην παράγραφο 3.7, προτείνουμε έναν αλγόριθμο που θα μας διευκολύνει προς την κατεύθυνση αυτή.

Ο Becker (1989, 1992) παρουσίασε μια γενική οικογένεια μοντέλων συνάφειας για τρισδιάστατους πίνακες συνάφειας ώστε να είναι δυνατό να αντιμετωπιστούν διάφορες υποθέσεις τριπλής αλληλεπίδρασης μεταξύ των μεταβλητών και να καταστεί δυνατό να οριστούν κατάλληλα μοντέλα που να είναι μεταξύ τέτοιων περιοριστικών μοντέλων που προαναφέρθηκαν. Αν θέσουμε κατάλληλους περιορισμούς στις παραμέτρους των μοντέλων αυτών, μπορούμε να πάρουμε διάφορες περιορισμένες εκδοχές των μοντέλων με τις οποίες έχουν ασχοληθεί οι Agresti&Kezouh(1983), ο Goodman(1986), οι Gilula&Haberman(1988), οι Becker& Clogg(1989), κ.α.

Το μοντέλο συνάφειας RC(*M*) που εισήχθη από τον Goodman(1985, 1986, 1991) αποδείχθηκε πολύ χρήσιμο για τη μοντελοποίηση της σχέσης μεταξύ διακριτών μεταβλητών σε διάφορους τομείς, όπως την κοινωνιολογία, τη βιολογία, την εκπαίδευση, την ιατρική, κλπ. Πρόκειται για ένα λογαριθμοπολλαπλασιαστικό (*log-multiplicative*) μοντέλο όπου μπορεί να θεωρηθεί ως επέκταση του λογαριθμογραμμικού μοντέλου για πίνακες διπλής εισόδου. Στο μοντέλο συνάφειας RC(*M*), η συνάφεια μεταξύ των δυο μεταβλητών αναπαρίσταται με διγραμμικούς όρους που αποτελούνται από τα γινόμενα των ποσοτικοποιήσεων των κατηγοριών (τιμές κλίμακας ή σκορ) για καθεμιά από τις δυο μεταβλητές και ένα μέτρο συνάφειας. Οι τιμές κλίμακας ή σκορ και το μέτρο συνάφειας εκτιμώνται από τα δεδομένα, όπως ακριβώς και οι άλλες παράμετροι του μοντέλου. Πολλαπλά σύνολα σκορ και μέτρων συνάφειας μπορούν να εκτιμηθούν κατά τέτοιο τρόπο ώστε η αλληλεπίδραση να αναπαριστάται με άθροισμα διγραμμικών όρων. Το μοντέλο RC(*M*) είναι παρόμοιο με την ανάλυση αντιστοιχιών (*correspondence analysis*) και τα μοντέλα συσχέτισης(*correlation models*) για κατηγορικά δεδομένα.

Διάφορες στρατηγικές και γενικεύσεις έχουν προταθεί ώστε να επεκταθεί το μοντέλο συνάφειας RC(*M*) σε τρίτης ή υψηλότερης τάξης πίνακες συνάφειας. Σε αυτές τις προτάσεις είτε χρησιμοποιούνται διγραμμικοί όροι, είτε τριγραμμικοί όροι είτε και τα δύο ώστε να αποδοθούν οι συνάφειες σε πίνακες τρίτης τάξης και άνω. Η κεντρική ιδέα στην προσέγγιση του Becker ήταν να θεωρήσει υπό συνθήκη συσχέτιση μεταξύ των δύο μεταβλητών, δοθέντος του επιπέδου της τρίτης μεταβλητής, που μπορεί να διαχωριστούν σε χωρίς συνθήκη δεύτερης και τρίτης τάξης όρους αλληλεπίδρασης. Όλα τα μοντέλα που δεν περιλαμβάνουν όρο τριπλής αλληλεπίδρασης και θεωρήθηκαν από τους Goodman(1979, 1986), Clogg(1982a) και Agresti&Kezouh(1983) αποτελούν ειδικές περιπτώσεις του μοντέλου του Becker.

Οι προτάσεις για μοντέλα που περιλαμβάνουν διγραμμικούς όρους ουσιαστικά προσεγγίζουν το πρόβλημα ανάλυσης πολυδιάστατων πινάκων συνάφειας με το να αναδιατάσσουν τα κελιά ενός πολυδιάστατου πίνακα σε διδιάστατο πίνακα με τέτοιο τρόπο ώστε να μπορούν να αναλυθούν με το μοντέλο συνάφειας RC(*M*). Στο σημείο αυτό, πρέπει να αναφέρουμε ότι υπάρχουν δύο τρόποι αναδιάταξης πολυδιάστατων πινάκων συνάφειας. Ο πρώτος τρόπος για να χρησιμοποιηθεί το RC(*M*) μοντέλο συνάφειας ώστε να αναλυθεί ένας τρισδιάστατος ή μεγαλύτερης τάξης πίνακας είναι να προσαρμοστούν μοντέλα τύπου RC(*M*) σε πίνακες διπλής εισόδου πίνακες για κάθε επίπεδο της τρίτης μεταβλητής ή για συνδυασμούς των άλλων μεταβλητών. Με αυτήν την υπό συνθήκη προσέγγιση, μπορούν να μελετηθούν απλούστερα μοντέλα με το να τεθούν περιορισμοί στις τιμές των σκορ και τις παραμέτρους συνάφειας μέσω των πινάκων.

Ο δεύτερος τρόπος είναι να αντιμετωπιστούν οι πολυδιάστατες μεταβλητές ως απλές μονοδιάστατες μεταβλητές με μεγάλο πλήθος κατηγοριών (*polytomous*). Οι μεταβλητές του πολυδιάστατου πίνακα συνάφειας διαιρούνται σε δυο σύνολα και οι συνδυασμοί των κατηγοριών των μεταβλητών στο ένα σύνολο γίνονται γραμμές και οι συνδυασμοί των κατηγοριών των μεταβλητών στο άλλο σύνολο γίνονται οι στήλες ενός δισδιάστατου πίνακα. Αυτή η διαδικασία χρησιμοποιείται και στην ανάλυση αντιστοιχιών.

Οι διγραμμικοί όροι στις υπό συνθήκη και τις από κοινού προσεγγίσεις αντιπροσωπεύουν τις συνδυασμένες επιδράσεις της συνάφειας των τριών παραγόντων και είτε του ενός ή και των δύο παραγόντων από τις αλληλεπιδράσεις δυο παραγόντων (μια αλληλεπίδραση δυο παραγόντων στην υπό συνθήκη αλληλεπίδραση και δύο αλληλεπιδράσεις δυο παραγόντων στην από κοινού προσέγγιση). Οι διγραμμικοί όροι στα μοντέλα αυτά δεν αναπαριστούν μόνο την αλληλεπίδραση τριών παραγόντων και τα μοντέλα αυτά δεν είναι ιεραρχικά με την έννοια ότι δεν συμμορφώνονται τυπικά με την αρχή της ιεραρχίας, όπως ισχύει στη μοντελοποίηση των περισσότερων κατηγορικών δεδομένων. Ωστόσο, λαμβάνοντας υπόψη την από κοινού προσέγγιση, εάν υπάρχει μια σχετικά πολύπλοκη δομή αλληλεπίδρασης στον πίνακα πολλαπλής εισόδου, π.χ. ισχυρή μερική συνάφεια δύο, αλλά και τριών παραγόντων, η αναπαράσταση των αλληλεπιδράσεων που δίνονται από το διγραμμικό όρο μπορεί να είναι αρκετά περίπλοκη και δύσκολη να ερμηνευθεί. Οι μεταβλητές και στα από κοινού και στα υπό συνθήκη μοντέλα αντιμετωπίζονται μη συμμετρικά και δίνεται έμφαση σε συγκεκριμένες όψεις των δεδομένων έναντι άλλων. Αυτά τα μοντέλα δεν είναι χρήσιμα σε περιπτώσεις που απαιτείται συμμετρική αντιμετώπιση των μεταβλητών.

Τα μοντέλα που προτάθηκαν από τους Choulakian(1988a, 1988b) και Mooijaart(1992) παρέχουν συμμετρική αντιμετώπιση των μεταβλητών. Τα μοντέλα περιλαμβάνουν τριγραμμικούς όρους για να αναπαραστήσουν αλληλεπιδράσεις τριών παραγόντων. Σε αυτά τα μοντέλα, οι τριγραμμικοί όροι έχουν την ίδια μορφή με το μοντέλο CANDECOMP(*Canonical Decomposition Model*), το οποίο είναι ισοδύναμο με το μοντέλο PARAFAC(*Parallel Factors Model*). Τέλος, η C.Anderson(1996) παρουσίασε μοντέλα συνάφειας που είναι γενικεύσεις των μοντέλων συνάφειας RC(*M*) για πίνακες συνάφειας τριπλής εισόδου με όρους αλληλεπίδρασης τριών παραγόντων.

3.1 Συμβολισμοί

Υποθέτουμε ότι έχουμε τρεις κατηγορικές μεταβλητές X, Y, Z με επίπεδα *I, J, K* αντίστοιχα. Συνεπώς, έχουμε έναν I×J×K πίνακα συνάφειας, με μεταβλητή γραμμής την X, μεταβλητή στήλης την Y και μεταβλητή στρώματος την Z. Επίσης, με n_{ijk}, *i=1,…, I, j=1, …, J, k=1, …, K,*

συμβολίζουμε την παρατηρούμενη συχνότητα κάθε κελιού και με m_{ijk} συμβολίζουμε την αναμενόμενη συχνότητα κάθε κελιού κάτω από την υπόθεση ενός μοντέλου. Αν επιθυμούμε να αθροίσουμε ως προς κάποιο δείκτη τότε αυτός αντικαθίσταται από μια τελεία στη θέση των i, j, k, στα σύμβολα n_{ijk}, m_{ijk}.

Επίσης, στις εκφράσεις των λογαριθμογραμμικών μοντέλων που θα ακολουθήσουν, τα σύμβολα l, l_i^X, l_j^Y, l_k^Z είναι αντίστοιχα οι κύριες επιδράσεις γραμμών, στηλών και στρώματος και l_{ij}^{XY}, l_{jk}^{YZ}, l_{ik}^{XZ}, l_{ijk}^{XYZ} είναι αντίστοιχα οι επιδράσεις γραμμών-στηλών, στηλών-στρωμάτων, γραμμών-στρωμάτων και η επίδραση γραμμών-στηλών και στρωμάτων.

Τέλος, αν υποθέσουμε ότι F_{ijk} είναι η αναμενόμενη συχνότητα του (i, j) κελιού του k πίνακα, τότε μπορούμε να ορίσουμε το odds ratio θ_{ijk}. Για παράδειγμα, θα ορίσουμε ως $\theta_{ij(k)}=(F_{ijk}F_{i+1,j+1,k})/(F_{i,j+1,k}F_{i+1,j,k})$. Η παρένθεση στην έκφραση $\theta_{ij(k)}$ σημαίνει «δοθέντος» του επιπέδου της τρίτης μεταβλητής.

3.2 Λογαριθμογραμμικά ιεραρχικά μοντέλα για τρεις μεταβλητές

Η στρατηγική μοντελοποίησης των συμβατικών ιεραρχικών λογαριθμογραμμικών μοντέλων ώστε να μελετηθεί η σχέση μεταξύ των τριών μεταβλητών, συνήθως ξεκινάει με το μοντέλο της πλήρους ανεξαρτησίας και μετά προστίθενται όροι διπλής αλληλεπίδρασης και κατόπιν όροι τριπλής αλληλεπίδρασης αν τα μοντέλα με τις διπλές αλληλεπιδράσεις δεν προσαρμόζονται καλά στα δεδομένα. Επίσης, στην παραπάνω διαδικασία οφείλουμε να σημειώσουμε ότι δεν γίνεται διαχωρισμός μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών, η οποία θα ήταν χρήσιμη σε κάποιες περιπτώσεις, καθώς κάτι τέτοιο δεν είναι απαραίτητο στην λογαριθμική μοντελοποίηση. Οι έλεγχοι καλής προσαρμογής των μοντέλων γίνεται με τη βοήθεια της στατιστικής συνάρτησης X^2 του Pearson ή με τη βοήθεια του λόγου πιθανοφάνειας.

Στη συνέχεια, θα παρουσιάσουμε τα λογαριθμογραμμικά μοντέλα και τα μοντέλα συνάφειας για τρεις μεταβλητές βήμα προς βήμα. Το απλούστερο λογαριθμογραμμικό μοντέλο για τρισδιάστατο πίνακα συνάφειας είναι το μοντέλο της πλήρης ανεξαρτησίας για τρεις μεταβλητές και είναι το εξής:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z \qquad (3.1)$$

με *IJK-I-J-K+2* β.ε. Από το μοντέλο (3.1) αρχίζει συνήθως η μελέτη των ερευνητών (*backward selection*) για τη συνάφεια των μεταβλητών. Αν το μοντέλο (3.1) είναι αποδεκτό, βάσει των ελέγχων καλής προσαρμογής, σημαίνει πως δεν υπάρχει σχέση μεταξύ των μεταβλητών X, Y, Z. Ωστόσο, οι ερευνητές προσθέτουν όρους διπλής αλληλεπίδρασης ώστε

να συλλάβουν την απομάκρυνση από την πλήρη ανεξαρτησία και να εξάγουν συμπεράσματα για τη σχέση μεταξύ των μεταβλητών. Κατά αυτόν τον τρόπο, θα μπορούσε, για παράδειγμα, να προκύψει το μοντέλο της υπό συνθήκης ανεξαρτησίας:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + l_{jk}^{YZ} + l_{ik}^{XZ} \quad (3.2)$$

με *(I-1)(J-1)K* β.ε. Βάσει του μοντέλου (3.2), οι μεταβλητές X και Y είναι υπό συνθήκη ανεξάρτητες δοθείσης της μεταβλητής Z. Ομοίως, ορίζονται και τα άλλα δύο μοντέλα της υπό συνθήκη ανεξαρτησίας δοθείσης της μεταβλητής X και της μεταβλητής Y.

Άλλες φορές, οι ερευνητές προσθέτουν όλους τους όρους διπλής αλληλεπίδρασης στην εξίσωση (3.1), σε περίπτωση έλλειψης καλής προσαρμογής, και το μοντέλο γίνεται:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{jk}^{YZ} + l_{ik}^{XZ} \quad (3.3)$$

το οποίο έχει *(I-1)(J-1)(K-1)* β.ε.

Τέλος, κάποιες φορές είναι απαραίτητο να συμπεριληφθεί στο μοντέλο ο όρος τριπλής αλληλεπίδρασης, όταν κανένα από τα προηγούμενα μοντέλα δεν προσαρμόζεται καλά στα δεδομένα:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{jk}^{YZ} + l_{ik}^{XZ} + l_{ijk}^{XYZ} \quad (3.4)$$

και έχει 0 β.ε., αφού πρόκειται για το πλήρες μοντέλο.

Πολλές φορές, η διαδικασία για την επιλογή του βέλτιστου λογαριθμογραμμικού μοντέλου είναι αντίστροφη, διότι πολλοί ερευνητές ξεκινούν από το πλήρες μοντέλο και με μια «προς τα εμπρός» (*forward*) διαδικασία, παραλείπουν έναν ή περισσότερους όρους.

Με τα κλασικά λογαριθμογραμμικά μοντέλα ασχολήθηκαν εκτενέστατα οι S.Fienberg και A.Agresti, όπου μπορούμε να ανατρέξουμε για περισσότερες λεπτομέρειες.

3.3 Λογαριθμογραμμικά μοντέλα 4ης τάξης και άνω-Μειονεκτήματα τους

Τα λογαριθμογραμμικά μοντέλα για πίνακες συνάφειας τριπλής εισόδου, γενικεύονται με άμεσο τρόπο και για πίνακες τετραπλής εισόδου και άνω. Ωστόσο, καθώς αυξάνει ο αριθμός των διαστάσεων του προβλήματος, αυξάνει και η πολυπλοκότητα της δομής του μοντέλου. Αυτό έχει ως συνέπεια τα εξής: πρώτον, η αύξηση του αριθμού των όρων αλληλεπίδρασης, καθιστά δύσκολη την επιλογή του βέλτιστου μοντέλου. Δεύτερον, αυξάνεται ο αριθμός των κελιών του πίνακα συνάφειας και δημιουργείται πρόβλημα στην ύπαρξη εκτιμητών και την καταλληλότητα ασυμπτωτικών θεωριών.

3.4 Μοντελοποίηση σε πίνακες συνάφειας τριπλής εισόδου με τη βοήθεια των local odds ratio

Ο Goodman(1979) παρουσίασε μια γενική κλάση μοντέλων συνάφειας μεταξύ δύο διακριτών διατάξιμων μεταβλητών και σύντομα γενίκευσε τις μεθόδους του ώστε να εφαρμόζονται σε πολυδιάστατους πίνακες συνάφειας. Λίγο αργότερα, ο Clogg (1982) παρουσίασε κάποια μοντέλα συνάφειας για πολυδιάστατους πίνακες συνάφειας για διατάξιμες μεταβλητές που εξειδικεύονται στις εξής περιπτώσεις: (α)υπό συνθήκη μοντέλα συνάφειας για K πίνακες συνάφειας διπλής εισόδου, (β)μοντέλα μερικής συνάφειας σε πίνακες συνάφειας τριπλής εισόδου και (γ)μοντέλα συμμετρικής συνάφειας για πίνακες συνάφειας τριπλής εισόδου.

Τα μοντέλα αυτά παρουσιάζονται στους πίνακες που ακολουθούν και μπορούν πολύ εύκολα να γενικευθούν σε περισσότερες από τρεις διαστάσεις. Επίσης, τα μοντέλα δίνονται με τη βοήθεια των odds ratio και local odds ratio, θ_{ijk}.

(α)Υπό Συνθήκη Μοντέλα Συνάφειας για K I×J Πίνακα Συνάφειας:

Τα μοντέλα συνάφειας για K I×J πίνακα συνάφειας περιέχονται στον Πίνακα 3.1 και οφείλουμε να σημειώσουμε τα εξής για τα μοντέλα που περιλαμβάνει:

(1)Το μοντέλο ανεξαρτησίας προκύπτει αν $\theta_{ij(k)}=1$.

(2)Το ομοιογενές ομοιόμορφο μοντέλο προκύπτει αν $\theta_{ij(k)}=\theta$.

Το ετερογενούς ομοιόμορφης συνάφειας μοντέλο προκύπτει αν $\theta_{ij(k)}=\theta_{..(k)}$.

(3)Το μοντέλο ομοιογενούς επίδρασης γραμμών προκύπτει αν $\theta_{ij(k)}=\theta_{i...}$

(4)Το μοντέλο ετερογενούς επίδρασης γραμμών προκύπτει αν $\theta_{ij(k)}=\theta_{i.(k)}$.

	Επιδράσεις στο $\theta_{ij(k)}$	**Περιγραφή Μοντέλου**	**Βαθμοί Ελευθερίας:** K(I-1)(J-1) μείον:	**Ειδική Περίπτωση των Μοντέλων**
1	1	Μοντέλο ανεξαρτησίας	0	2a-5d
2a	θ	Ομοιογενές Ομοιόμορφο	1	2b-5d
2b	$\theta_{..(k)}$	Ετερογενές Ομοιόμορφο	K	3b-3c, 4b-4c, 5b-5d
3a	$\theta_{i..}$	Ομοιογένεια Γραμμών	(I-1)	3b-3c, 5a-5d
3b	$\theta_{..(k)}\theta_{i..}$	Απλή Ετερογένεια Γραμμών	K+(I-2)	3c, 5b-5d
3c	$\theta_{i.(k)}$	Ετερογένεια Γραμμών	K+K(I-2)	5b, 5c

4a	$\theta_{.j.}$	Ομοιογένεια Στηλών	(J-1)	4b-4c, 5a-5d
4b	$\theta_{..(k)}\theta_{.j.}$	Απλή Ετερογένεια Στηλών	K+(J-2)	4c, 5b-5d
4c	$\theta_{.j(k)}$	Ετερογένεια Στηλών	K+K(J-2)	5c-5d
5a	$\theta_{i..}\theta_{.j.}$	Ομοιογενής Επίδραση Γραμμών και Στηλών	1+(I-2)+(J-2)	5b-5d
5b	$\theta_{i.(k)}\theta_{.j.}$	Ετερογένεια Γραμμών, Ομοιογένεια Στηλών	K+K(I-2)+(J-2)	5d
5c	$\theta_{i..}\theta_{.j(k)}$	Ομοιογένεια Γραμμών, Ετερογένεια Στηλών	K+(I-2)+K(J-2)	5d
5d	$\theta_{i.(k)}\theta_{.j(k)}$	Ετερογένεια Γραμμών, Ετερογένεια Στηλών	K+K(I-2)+K(J-2)	-

Πίνακας 3.1

Υπό συνθήκη μοντέλα συνάφειας για K IxJ πίνακα συνάφειας.

(β)Μοντέλα Μερικής Συνάφειας[(1)] σε IxJxK Πίνακα Συνάφειας:

	Περιγραφή Μοντέλου	**Βαθμοί Ελευθερίας:** IJK-I-J-K+2 μείον:
1	Μοντέλο Ανεξαρτησίας	0
2	Ομοιόμορφο	
2a	RC	1
2b	RC, RL	2
2c	RC, RL, CL	3
3	Επιδράσεις Γραμμών και Ομοιόμορφες Επιδράσεις	
3a	Επίδραση Γραμμών στο RC	(I-1)+2
3b	Επίδραση Γραμμών στο RL	(I-1)+2
3c	Επίδραση Γραμμών στο RL, RC	2(I-1)+1
4	Επιδράσεις Στηλών και Επιδράσεις Γραμμών και Ομοιόμορφες Επιδράσεις	
4a	Επιδράσεις Στηλών στο RC	3+2(I-2)+(J-2)
4b	Επιδράσεις Στηλών στο CL	3+2(I-2)+(J-2)
4c	Επιδράσεις Στηλών στο RC, CL	3+2(I-2)+2(J-2)

5	Επιδράσεις Στρώματος και Επιδράσεις Γραμμών και Στηλών	
5a	Επιδράσεις Στρώματος στο RC	3+2(I-2)+2(J-2)+(K-2)
5b	Επιδράσεις Στρώματος στο CL	3+2(I-2)+2(J-2)+(K-2)
5c	Επιδράσεις Στρώματος στο RC, CL	3+2(I-2)+2(J-2)+2(K-2)

[(1)]Τα μοντέλα 3α και 3β δεν είναι εμφωλευμένα (*nested*): κανένα δεν είναι ειδική περίπτωση του άλλου, αλλά και τα δύο είναι ειδικές περιπτώσεις του μοντέλου 3c και όλων των μοντέλων κάτω από αυτό. Το ίδιο σχόλιο ισχύει και για τα μοντέλα, 4α, 4β και 5α, 5β.

Πίνακας 3.2

Μοντέλα μερικής συνάφειας[(1)] σε IxJxK πίνακα συνάφειας.

Όσον αφορά στον πίνακα 3.2, πρέπει να σημειώσουμε τα εξής για τα μοντέλα που περιλαμβάνει:

(1)Το μερικής συνάφειας μοντέλο ανεξαρτησίας είναι ισοδύναμο με το μοντέλο ανεξαρτησίας των μεταβλητών γραμμής, στήλης και στρώματος και προκύπτει αν $\{\theta_{ij(\kappa)}=1, \theta_{i(j)k}=1, \theta_{(i)jk}=1\}$.

(2)Το μοντέλο ομοιόμορφα μερικής συνάφειας μοντέλο προκύπτει αν $\{\theta_{ij(\kappa)}=\theta^{RC}, \theta_{i(j)k}=\theta^{RL}, \theta_{(i)jk}=\theta^{CL}\}$.

(3)Το μοντέλο επίδρασης γραμμών στα RC, επίδρασης γραμμών στα RL και ομοιόμορφη επίδραση στο CL προκύπτει αν $\{\theta_{ij(\kappa)}=\theta_{i..}{}^{RC}, \theta_{i(j)k}=\theta_{i..}{}^{RL}, \theta_{(i)jk}=\theta^{CL}\}$.

(4)Το μοντέλο επίδρασης γραμμών στα RC και RL και επίδρασης στηλών στα RL και CL προκύπτει αν $\{\theta_{ij(\kappa)}=\theta_{i..}{}^{RC}\,\theta_{.j.}{}^{RC}, \theta_{i(j)k}=\theta_{i..}{}^{RL}, \theta_{(i)jk}=\theta_{.j.}{}^{CL}\}$.

(γ)Μοντέλα Συμμετρικής Συνάφειας σε IxJxK Πίνακα Συνάφειας:

Το μοντέλο ομοιογενούς επίδρασης γραμμών ή στηλών ή στρώματος προκύπτει αν $\theta_{ij(k)}=\theta_{i..}{}^{R}\,\theta_{.j.}{}^{C}$, $\theta_{i(j)k}=\theta_{i..}{}^{R}\,\theta_{..k}{}^{L}$, $\theta_{(i)jk}=\theta_{.j.}{}^{C}\,\theta_{..k}{}^{L}$ αντίστοιχα.

Το μοντέλο συμμετρικής επίδρασης γραμμής-στήλης προκύπτει αν I=J ή $\theta_{i..}{}^{RC}=\theta_{.i.}{}^{RC}=\theta_{i}{}^{RC}$ και $\theta_{ij(k)}=\theta_{i}{}^{RC}\theta_{j}{}^{RC}$.

Το μοντέλο με όλα τα περιθώρια αθροίσματα συμμετρικά προκύπτει αν $\{\theta_{ij(k)}=\theta_{i}{}^{RC}\,\theta_{j}{}^{RC}, \theta_{i(j)k}=\theta_{i}{}^{RL}\,\theta_{K}{}^{RL}, \theta_{(i)jk}=\theta_{j}{}^{CL}\,\theta_{k}{}^{CL}\}$.

Το μοντέλο της πλήρους συμμετρίας προκύπτει αν απαιτήσουμε $\theta_{i..}{}^{RC}=\theta_{i..}{}^{RL}=\theta_{.i.}{}^{RC}=\ldots=\theta_{..i}{}^{CL}=\theta_i$, δηλαδή $\{\theta_{ij(k)}=\theta_i\,\theta_j, \theta_{i(j)k}=\theta_i\,\theta_k, \theta_{(i)jk}=\theta_j\,\theta_k\}$.

3.5 Λογαριθμο-πολλαπλασιαστικά μοντέλα συνάφειας για τρεις μεταβλητές

Οι ερευνητές συχνά επιθυμούν να εκτιμήσουν διάφορα «ενδιάμεσα» μοντέλα που αναφέρθηκαν στην παράγραφο 3.2 και να είναι οικονομικά (*parsimonious*), αλλά διατηρώντας και την αλληλεπίδραση τρίτης τάξης (μη ιεραρχικά μοντέλα). Επίσης, οι ερευνητές επιθυμούν να εκτιμήσουν «ενδιάμεσα» μοντέλα (ιεραρχικά ή μη ιεραρχικά μοντέλα) αναλύοντας τις αλληλεπιδράσεις δύο παραγόντων, χωρίς να υπάρχει ο όρος τριπλής αλληλεπίδρασης. Σε όλα αυτά τα μοντέλα, θα τεθεί δομή στους όρους αλληλεπίδρασης ώστε να μελετηθεί η δομή και οι σχέσεις μεταξύ των μεταβλητών. Κατά αυτόν τον τρόπο, προκύπτουν τα μοντέλα συνάφειας που θα αναλύσουμε στη συνέχεια.

Στο σημείο αυτό, οφείλουμε να σημειώσουμε ότι ένα βασικό πλεονέκτημα των λογαριθμοπολλαπλασιαστικών μοντέλων είναι ότι οι έλεγχοι καλής προσαρμογής παραμένουν αναλλοίωτοι για κάθε εναλλαγή κατηγορίας γραμμής και στήλης. Για ερευνητές που υποψιάζονται ότι οι μεταβλητές για τις οποίες ενδιαφέρονται είναι διατάξιμες από τη φύση τους, αλλά δεν είναι σίγουροι όμως για την ακριβή τους διάταξη, οι εκτιμήσεις για τα σκορ γραμμής και στήλης στα μοντέλα συνάφειας, παρέχουν μια εμπειρική διάταξη των μεταβλητών. Επίσης, τα μοντέλα αυτά δεν περιορίζονται μόνο για κατηγορικές μεταβλητές, αλλά και για διακριτές ή ποιοτικές μεταβλητές.

3.5.1 Ανάλυση μοντέλων συνάφειας χωρίς τον όρο τριπλής αλληλεπίδρασης

Το παρακάτω μοντέλο συνάφειας είναι εναλλακτικό του μοντέλου (3.3):

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + f_1^{XY} m_{i1} n_{j1} + f_1^{YZ} m_{1i}^* h_{k1} + f_1^{XZ} n_{j1}^* h_{k1}^* \qquad (3.5)$$

με περιορισμούς:

$$\sum_{i=1}^{I} m_{i1} = \sum_{i=1}^{I} m_{i1}^* = \sum_{j=1}^{J} n_{j1} = \sum_{j=1}^{J} n_{j1}^* = \sum_{k=1}^{K} h_{k1} = \sum_{k=1}^{K} h_{k1}^* = 0 \text{ και}$$

$$\sum_{i=1}^{I} m_{i1}^2 = \sum_{i=1}^{I} m_{i1}^{*2} = \sum_{j=1}^{J} n_{j1}^2 = \sum_{j=1}^{J} n_{j1}^{*2} = \sum_{k=1}^{K} h_{k1}^2 = \sum_{k=1}^{K} h_{k1}^{*2} = 1,$$

όπου m_{i1}, m_{i1}^* είναι τα εκτιμηθέντα σκορ γραμμής για την XY και XZ μερική συνάφεια, n_{i1}, n_{i1}^*, είναι τα εκτιμηθέντα σκορ στήλης για την XY και YZ μερική συνάφεια και h_{i1}, h_{i1}^* είναι τα εκτιμηθέντα σκορ στρώματος για την XY και XZ μερική συνάφεια. Επίσης, $f_1^{XY}, f_1^{YZ}, f_1^{XZ}$ είναι οι παράμετροι συνάφειας (*intrinsic association parameter*) για τις XY, YZ, XZ μερικές συνάφειες αντίστοιχα.

Στο μοντέλο (3.5) πρέπει να τεθούν περιορισμοί και κεντροποίησης(*centering*) και κλίμακας (*scaling*) ώστε να προσδιοριστεί.

Το μοντέλο (3.5) έχει *IJK-3I-3J-3K+11* β.ε. και μπορεί να συμβολιστεί ως RC(1)+CL(1)+RL(1) μοντέλο συνάφειας, αφού χρησιμοποιείται μόνο μια διάσταση για να διασπάσει τους όρους διπλής αλληλεπίδρασης. Στο μοντέλο (3.5), αν θέσουμε τους εξής περιορισμούς: $\mu^{*}_{i1}=\mu_{i1}$, $\nu^{*}_{i1}=\nu_{j1}$ και $\eta^{*}_{i1}=\eta_{j1}$, έχουμε το εξής μοντέλο:

$$\log(m_{ijk}) = l + l^{X}_{i} + l^{Y}_{j} + l^{Z}_{k} + f^{XY}_{1} m_{i1} n_{j1} + f^{YZ}_{1} m_{1i} h_{k1} + f^{XZ}_{1} n_{j1} h_{k1} \qquad (3.6)$$

και έχει *IJK-2I-2J-2K+5* β.ε., συμβολίζεται ως RC(1)+RL(1)+CL(1) και απαιτεί αντίστοιχους περιορισμούς για τον προσδιορισμό των σκορ γραμμών, στηλών και στρωμάτων, όπως στο μοντέλο (3.5).

Η διαφορά της στατιστικής συνάρτησης του ελέγχου πιθανοφάνειας μεταξύ των μοντέλων (3.5) και (3.6) αποδίδει ένα X^2 στατιστικό με *I+J+K-6* β.ε. και μπορεί να χρησιμοποιηθεί ώστε να διαπιστωθεί εάν οι περιορισμοί στα σκορ είναι πραγματικά ανάλογοι των δεδομένων.

Όταν κανένα από τα μοντέλα (3.5) και (3.6) δεν προσαρμόζεται καλά στα δεδομένα, τότε το επόμενο βήμα είναι η αύξηση της διαστασιμότητας για καθένα από τους όρους διπλής αλληλεπίδρασης ώστε να διαπιστωθεί η περίπλοκη σχέση μεταξύ των μεταβλητών. Η πιο γενική μορφή του μοντέλου αυτού είναι το μοντέλο RC(M_1)+RC(M_2)+RC(M_3):

$$\log(m_{ijk}) = l + l^{X}_{i} + l^{Y}_{j} + l^{Z}_{k} + l^{XY}_{ij} + l^{YZ}_{jk} + l^{XZ}_{ik} + \sum_{r=1}^{M_1} f_r m_{ir} n_{jr} +$$

$$+ \sum_{s=1}^{M_2} f_s m^{*}_{is} h_{ks} + \sum_{t=1}^{M_3} f_t n^{*}_{jt} h^{*}_{kt} \qquad (3.7)$$

με $0 \le M_1 \le \min(I-1, J-1)$, $0 \le M_2 \le \min(I-1, K-1)$ και $0 \le M_3 \le (J-1, K-1)$.

Χωρίς βλάβη της γενικότητας, κάθε σύνολο των παραμέτρων συνάφειας μπορεί να διαταχθεί από το μεγαλύτερο προς το μικρότερο. Το μοντέλο (3.7) αποτελεί τη γενίκευση του RC(*M*) μοντέλου συνάφειας του Goodman(1986, 1991) και επεκτάθηκε από τον Becker(1989, 1992).

Αν ισχύει $M^{*}_1 = \min(I-1, J-1), M^{*}_2 = \min(I-1, K-1), M^{*}_3 = \min(J-1, K-1)$, τότε το μοντέλο (3.7) είναι το πλήρες μοντέλο για κάθε όρο διπλής αλληλεπίδρασης και ισοδυναμεί με το μοντέλο (3.3). Οι απαιτούμενοι περιορισμοί για τον προσδιορισμό των παραμέτρων είναι:

$$\sum_{i=1}^{I} m_{im} = \sum_{i=1}^{I} m^{*}_{im} = \sum_{j=1}^{J} n_{jm} = \sum_{j=1}^{J} n^{*}_{jm} = \sum_{k=1}^{K} h_{km} = \sum_{k=1}^{K} h^{*}_{km} = 0 \text{ και}$$

$$\sum_{i=1}^{I} m_{im} m_{im'} = \sum_{i=1}^{I} m^{*}_{im} m^{*}_{im'} = \sum_{j=1}^{J} n_{jm} n_{jm'} = \sum_{j=1}^{J} n^{*}_{jm} n^{*}_{jm'} = \sum_{k=1}^{K} h_{km} h_{km'} = \sum_{k=1}^{K} h^{*}_{km} h^{*}_{km'} = d_{mm'},$$

όπου $\delta_{mm'}$ είναι η συνάρτηση δέλτα του Kronecker με $\delta_{mm'}$={1, αν m=m΄, 0 αλλιώς}. Το μοντέλο (3.7) έχει *IJK-I-J-K+2-M₁(I+J-M₁-2)-M₂(I+K-M₂-2)-M₃(J+K-M₃-2)* β.ε.

Αν στο μοντέλο (3.7) θέσουμε $M_1=M_2=M_3=M$, τότε οι β.ε. για το μοντέλο RC(*M*)+RL(*M*)+CL(*M*) απλοποιούνται σε *IJK-I-J-K+2-M(2I+2J+2K-3M-6)*.

3.5.2 Ανάλυση μοντέλων συνάφειας με όρο τριπλής αλληλεπίδρασης

Εάν το μοντέλο με την τριπλή αλληλεπίδραση είναι απαραίτητο για να περιγράψει τη σχέση μεταξύ των μεταβλητών X, Y, Z, οι ερευνητές θα πρέπει να αποφασίσουν αν θα αναλύσουν όλους ή μερικούς από τους όρους l_{ij}^{XY}, l_{jk}^{YZ}, l_{ik}^{XZ} με τον l_{ijk}^{XYZ}. Τρεις αναλύσεις είναι πιθανές:

(i)να αναλυθεί μόνο η παράμετρος της τριπλής αλληλεπίδρασης,

(ii)να αναλύσουμε την παράμετρο τριπλής αλληλεπίδρασης και μερικές αλλά όχι όλες τις παραμέτρους διπλής αλληλεπίδρασης και

(iii)να αναλύσουμε την παράμετρο τριπλής αλληλεπίδρασης και όλες τις παραμέτρους διπλής αλληλεπίδρασης. Μεταξύ των τριών επιλογών, το δεύτερο σενάριο έλαβε περισσότερο προσοχή κυρίως για τις εφαρμογές του σε κοινωνικές επιστήμες.

(i)ΑΝΑΛΥΟΥΜΕ ΤΑ l_{ij}^{XY}, l_{ijk}^{XYZ} (ΥΠΟ ΣΥΝΘΗΚΗ ΜΟΝΤΕΛΑ ΣΥΝΑΦΕΙΑΣ)

Το μοντέλο RC(*M*)-L είναι μια γενίκευση του μοντέλου RC(*M*) και προτάθηκε ώστε να μελετηθούν διαφορές μεταξύ των ομάδων για κάθε στρώμα:

$$\log(m_{ijk}) = l + l_i^{X} + l_j^{Y} + l_k^{Z} + l_{ik}^{XZ} + l_{jk}^{YZ} + \sum_{r=1}^{M} f_{rk} m_{irk} n_{jrk} . \qquad (3.9)$$

Το μοντέλο (3.9) έχει *(I-M-1)(J-M-1)K* β.ε. και χαρακτηρίζεται ως ετερογενές RC(*M*)-L μοντέλο με περιορισμούς:

$$\sum_{i=1}^{I} m_{irm} = \sum_{j=1}^{J} n_{jrm} = 0 \text{ και } \sum_{i=1}^{I} m_{irm} m_{ir'm} = \sum_{j=1}^{J} n_{jrm} n_{jr'm} = d_{mm'}.$$

Σε αντίθεση με το ετερογενές ομόλογό του μοντέλο, το ομοιογενές RC(*M*)-L μοντέλο δεν απαιτεί την ύπαρξη τριπλής αλληλεπίδρασης και θέτει περιορισμούς ισότητας των παραμέτρων φ_{mk}, μ_{imk}, ν_{jmk} σε κάθε στρώμα, όποτε το μοντέλο (3.9) γίνεται:

$$\log(m_{ijk}) = l + l_i^{X} + l_j^{Y} + l_k^{Z} + l_{ik}^{XZ} + l_{jk}^{YZ} + \sum_{r=1}^{M} f_r m_{ir} n_{jr} \qquad (3.10)$$

Η διαφορά μεταξύ δυο μοντέλων μπορεί να παρέχει πληροφορία σχετικά με τις αλλαγές στα σκορ και στην παράμετρο συνάφειας σε όλες τις διαστάσεις.

Ένα συγκεκριμένο μοντέλο που εφαρμόζεται συχνά στις κοινωνικές επιστήμες, προκύπτει αν θέσουμε περιορισμούς ομοιογένειας μόνο στα σκορ γραμμών και στηλών και είναι η επέκταση του απλού ετερογενούς RC(1) με υψηλότερες διαστάσεις:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + l_{ik}^{XZ} + l_{jk}^{YZ} + \sum_{r=1}^{M} f_{rk} m_{ir} n_{jr} \qquad (3.11)$$

το οποίο έχει *(I-1)(J-1)K-(I+J+K-4)M* β.ε. με περιορισμούς:

$$\sum_{i=1}^{I} m_{ir} = \sum_{j=1}^{J} n_{jr} = 0 \text{ και } \sum_{i=1}^{I} m_{ir}^2 = \sum_{j=1}^{J} n_{jr}^2 = 1.$$

Αν ξαναγράψουμε το μοντέλο (3.11), θέτοντας όπου $\varphi_{rk}=\varphi_r\eta_{\kappa r}$, έχουμε το μοντέλο:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + l_{ik}^{XZ} + l_{jk}^{YZ} + \sum_{r=1}^{M} f_r m_{ir} n_{jr} h_{kr} \qquad (3.12)$$

το οποίο είναι παρόμοιο με το μοντέλο PARAFAC/CANDECOMP, το οποίο χρησιμοποιείται κυρίως για μετρικά δεδομένα στην ψυχομετρία, αλλά εδώ χρησιμοποιείται ώστε να αναλυθούν λογαριθμογραμμικοί παράμετροι. Είναι γνωστό ότι υπό κανονικές συνθήκες, οι λύσεις που προκύπτουν από την ανάλυση PARAFAC/CANDECOMP είναι μοναδικές και δεν απαιτούν περιορισμούς περιστροφής. Το μοντέλο (3.12) συμβολίζεται (XY+XYZ) PARAFAC/CANDECOMP RCL(M) όπου (XY+XYZ) τονίζει τους όρους που διασπώνται και το RCL τονίζει την λογαριθμο-τριγραμμική (*log-trilinear*) σχέση μεταξύ των μεταβλητών γραμμής, στήλης και στρώματος.

Μια άλλη διαδεδομένη ανάλυση στην ψυχομετρία που επίσης ενσωματώνει λογαριθμο-τριγραμμικούς όρους είναι στο Tucker's 3-mode(Tucker 1964, 1966, Kroonenberg, 1983) μοντέλο, το οποίο μπορεί να συμβολιστεί ως (XY+XYZ) Tucker 3-mode RCL(M_1, M_2, M_3):

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{ik}^{XZ} + \sum_{r=1}^{M_1}\sum_{s=1}^{M_2}\sum_{t=1}^{M_3} f_{rst} m_{ir} n_{js} h_{kt} \qquad (3.13)$$

με *(I-1)(J-1)K-R(I-R-1)-S(J-S-1)-T(K-T)-RST* β.ε.

(ii) ΑΝΑΛΥΟΥΜΕ ΤΟ l_{ijk}^{XYZ}

Οι Siciliano και Mooijaart(1997) ασχολήθηκαν με τη γενική οικογένεια μοντέλων που περιλαμβάνουν τριγραμμικούς όρους για να αναπαραστήσουν αλληλεπιδράσεις τριών παραγόντων και παρουσίασαν κάποιες περιορισμένες εκδοχές με μια εναλλακτική τυποποίηση. Η γενική ιδέα ήταν η θεώρηση διγραμμικών αναλύσεων (*bilinear*

decompositions) για τους όρους δεύτερης τάξης και τριγραμμικών αναλύσεων (*trilinear decomposition*) για τους όρους τρίτης τάξης. Συγκεκριμένα, χρησιμοποίησαν το μοντέλο RARAFAC/CONDECOMP μοντέλο για να ορίσουν την αλληλεπίδραση τρίτης τάξης. Το μοντέλο αυτό είναι της μορφής:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{jk}^{YZ} + l_{ik}^{XZ} + \sum_{r=1}^{M} f_{r(k)} m_{ir(k)} n_{jr(k)} \quad (3.14)$$

όπου οι παράμετροι $\mu_{ir(\kappa)}$ και $\nu_{jr(k)}$ είναι τα σκορς γραμμών και στηλών αντίστοιχα του υπό συνθήκη πίνακα IxJ δοθέντος του επιπέδου k της τρίτης μεταβλητής και η παράμετρος $\varphi_{r(k)}$ είναι η παράμετρος συνάφειας (*intrinsic*) μεταξύ των μεταβλητών X και Y δοθέντος του επιπέδου της μεταβλητής Z.

Αν θεωρήσουμε την ειδική περίπτωση του μοντέλου (3.14), όπου υποθέτουμε ότι οι παράμετροι $\mu_{ir(\kappa)}$ και $\nu_{jr(k)}$ είναι ομοιογενείς για τους k πίνακες, άρα $\mu_{ir(\kappa)}=\mu_{ir}$ και $\nu_{jr(k)}=\nu_{jr}$ και ξαναγράφοντας τον όρο $\varphi_{r(k)}$ ως έναν πολλαπλασιαστικό παράγοντα $\varphi_{r(k)}=\varphi_r\eta_{\kappa r}$, τότε το μοντέλο (3.14) γίνεται:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{jk}^{YZ} + l_{ik}^{XZ} + \sum_{r=1}^{M} f_r m_{ir} n_{jr} h_{kr} \quad (3.15)$$

και είναι το γνωστό μοντέλο PARAFAC. Το μοντέλο (3.15) έχει *(I-1)(J-1)(K-1)-M(I+J+K-5)* β.ε., συμβολίζεται ως XYZ-PARAFAC/CANDECOMP RCL(*M*) και μπορεί να θεωρηθεί ως μια ειδική επέκταση του μοντέλου RC τύπου παλινδρόμησης επίδραση στρώματος που προτάθηκαν από τους Goodman και Hout(1998).

Η C.Anderson (1996) πρότεινε μοντέλα λογαριθμοπολλαπλασιαστικά που μπορεί να θεωρηθούν ως απλουστεύσεις των πλήρων λογαριθμογραμμικών μοντέλων για πίνακες τριπλής εισόδου όπου οι όροι αλληλεπίδρασης τριών παραγόντων ή κάποιοι συνδυασμοί όρων διπλής και τριπλής αλληλεπίδρασης διασπώνται και προσεγγίζονται από το Tucker's 3-mode μοντέλο κύριων συνιστωσών (Tucker 1964, 1966, Kroonenberg, 1983), το οποίο έχει πιο απλές και πιο εύκολες στην ερμηνεία αποδόσεις. Εν αντιθέσει, οι Mooijart και Choulakian ανέλυσαν μόνο τον παράγοντα τριπλής αλληλεπίδρασης με την βοήθεια του μοντέλου PARAFAC/CANDECOMP.

Το Tucker's 3-mode μοντέλο κύριων συνιστωσών μοντέλο είναι της μορφής:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{jk}^{YZ} + l_{ik}^{XZ} + \sum_{r=1}^{M_1}\sum_{s=1}^{M_2}\sum_{t=1}^{M_3} f_{rst} m_{ir} n_{js} h_{kt} \quad (3.16)$$

με $l, l_i^X, l_j^Y, l_k^Z, l_{ij}^{XY}, l_{jk}^{YZ}, l_{ik}^{XZ}$ είναι οι κύριες επιδράσεις και οι όροι αλληλεπίδρασης δύο παραγόντων, $\mu_{ir(\kappa)}$, $\nu_{jr(k)}$ είναι οι τιμές κλίμακας(σκορ) για τις κατηγορίες i, j, k των μεταβλητών

X, Y, Z στις συνιστώσες r, s, t αντίστοιχα και $\varphi_{r(k)}$ είναι η παράμετρος συνάφειας (***intrinsic association***). Το μοντέλο(3.16) έχει ***(I-1)(J-1)(K-1)-R(I-R-1)-S(J-S-1)-T(K-T-1)-RST*** β.ε., συμβολίζεται ως XYZ-Tucker-3-mode RCL(R, S, T) και μπορεί να θεωρηθεί ως μια ειδική επέκταση του μοντέλου RC τύπου παλινδρόμησης επίδραση στρώματος με πολύπλοκες σχέσεις αλληλεπίδρασης. Επίσης, $0 \le M_1 \le (I-1)$, $0 \le M_2 \le (J-1)$ και $0 \le M_3 \le (K-1)$ και όταν ισχύει $M_1=I-1$, $M_2=J-1$, $M_3=K-1$, ο τύπος (3.16) αντιστοιχεί στο πλήρες λογαριθμογραμμικό μοντέλο.

Ενώ το μοντέλο CANDECOMP έχει απλούστερη μορφή από το μοντέλο Tucker's 3-mode, δεν σημαίνει απαραίτητα πως το μοντέλο CANDECOMP θα οδηγήσει σε πιο οικονομική απεικόνιση των δεδομένων. Με βάση την δομή των δεδομένων, έχοντας διαφορετικό αριθμό συνιστωσών για διαφορετικές μεταβλητές μπορεί να οδηγηθούμε σε λιγότερες παραμέτρους και πιο οικονομική αναπαράσταση των δεδομένων.

(iii) ΑΝΑΛΥΟΥΜΕ ΤΑ $l_{ij}^{XY}, l_{jk}^{YZ}, l_{ik}^{XZ}, l_{ijk}^{XYZ}$

Τέλος, θα μπορούσαμε να αναλύσουμε και τους τέσσερις όρους χρησιμοποιώντας οποιαδήποτε μέθοδο ανάλυσης. Τέτοιου είδους αναλύσεις είναι χρήσιμες όταν καμιά από τις μεταβλητές δεν μπορεί να θεωρηθεί ξεκάθαρα ως αποκριτική ή/και επεξηγηματική και ο ερευνητής ενδιαφέρεται για την κατανόηση της περίπλοκης σχέση αλληλεπίδρασης μεταξύ και των τριών μεταβλητών. Από την στιγμή που δεν περιλαμβάνεται κανένας όρος αλληλεπίδρασης δεύτερης τάξης στο μοντέλο, περιορισμοί centering γραμμής, στήλης, στρώματος στις παραμέτρους δεν είναι απαραίτητοι. Το μοντέλο μπορεί να γραφεί ως εξής:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + \sum_{r=1}^{M} f_r m_{ir} n_{jr} h_{kr} \qquad (3.17)$$

με περιορισμούς:

$$\sum_{i=1}^{I} m_{ir}^2 = \sum_{j=1}^{J} n_{jr}^2 = \sum_{k=1}^{K} h_{kr}^2 = 1.$$

Το μοντέλο (3.17) έχει ***IJK-(I+J+K-2)(M+1)***β.ε.

Ομοίως, το μοντέλο (XY+XZ+YZ+XYZ) Tucker-3-mode RCL(R, S, T) μπορεί να γραφεί ως εξής:

$$\log(m_{ijk}) = l + l_i^X + l_j^Y + l_k^Z + \sum_{r=1}^{M_1}\sum_{s=1}^{M_2}\sum_{t=1}^{M_3} f_{rst} m_{ir} n_{js} h_{kt} \qquad (3.18)$$

με τους εξής περιορισμούς:

$$\sum_{i=1}^{I} m_{ir}^2 = \sum_{j=1}^{J} n_{js}^2 = \sum_{k=1}^{K} h_{kt}^2 = 1,\ \sum_{i=1}^{I} m_{ir} m_{ir'} = \sum_{j=1}^{J} n_{js} n_{js'} = \sum_{k=1}^{K} h_{kt} h_{kt'} = 0,$$

$$\sum_{s=1}^{M_2}\sum_{t=1}^{M_3} f_{rst} f_{r'st} = \sum_{r=1}^{M_1}\sum_{t=1}^{M_3} f_{rst} f_{rs't} = \sum_{r=1}^{M_1}\sum_{s=1}^{M_2} f_{rst} f_{rst'} = 0, r \neq r', s \neq s', t \neq t'$$

και *(I-1)(J-1)K-R(I-R)-S(J-S)-T(K-T)-RST* β.ε.

Εκτός από τα παραπάνω μοντέλα που αναφέραμε, είναι δυνατό να συνδυαστούν στο μοντέλο λογαριθμο-διγραμμικοί όροι και λογαριθμό-τριγραμμικοί όροι. Επίσης, μπορεί να τεθούν συνεπή σκορ στα σκορ της γραμμής, στήλης και στρώματος. Αν και τα μοντέλα με μεικτούς λογαριθμο-διγραμμικούς όρους και λογαριθμό-τριγραμμικούς όρους έχουν το δέλεαρ τους, είναι αρκετά δύσκολο να επιτευχθεί σύγκλιση σε αυτά του είδους την μοντελοποίηση και κάποια επιμέρους μοντέλα παρουσιάζουν δυσκολίες στη διαδικασία εκτίμησης. Μοντέλα αυτού του είδους προτάθηκαν από τους Siciliano και Mooijart, αλλά δεν παρέχουν κάποια εμπειρική επεξήγηση. Μέχρι να λυθεί ικανοποιητικά το πρόβλημα της εκτίμησης, το δέλεαρ της τυποποίησης με μεικτούς λογαριθμο-διγραμμικής και λογαριθμό-τριγραμμικοί είναι περιορισμένο.

Συνολικά, η παραπάνω συζήτηση για διαφορετικές αναλύσεις με ή χωρίς όρο τριπλής αλληλεπίδρασης συστηματοποιεί στρατηγικές για ερευνητές ώστε να μοντελοποιήσουν και να κατανοήσουν τις περίπλοκες σχέσεις μεταξύ τριών μεταβλητών X, Y, Z. Η απόφαση ποιοι όροι θα αναλυθούν τελικά πρέπει να βασίζεται στο θεωρητικό και ουσιαστικό ενδιαφέρον στις μεταβλητές που σχετίζονται, ειδικά όταν κάποιες από αυτές θεωρηθούν ως επεξηγηματικές και άλλες ως αποκριτικές. Η επιλογή ανάμεσα στην διάσπαση PARAFAC/CANDECOMP και την Tucker 3-mode πρέπει να στηρίζεται στην καλή προσαρμογή του μοντέλου, την οικονομία (*parsimony*) του μοντέλου και την ευκολία στην ερμηνεία. Όταν και οι δυο αναλύσεις δίνουν παρόμοια αποτελέσματα, η μέθοδος PARAFAC/CANDECOMP προτιμάται επειδή υπερτερεί στο τελευταίο, δηλαδή στην ευκολία στην ερμηνεία του μοντέλου.

3.6 Εκτιμήσεις

Από τη στιγμή που τα πολυδιάστατα και πολυγραμμικά μοντέλα έγιναν λογαριθμο-πολλαπλασιαστικά στη φύση τους, οι εκτιμητές μέγιστης πιθανοφάνειας δεν είναι διαθέσιμοι στα περισσότερα στατιστικά πακέτα. Εν αντιθέσει, υπολογίζονται με τη βοήθεια επαναληπτικών μεθόδων. Τα προγράμματα μακροεντολών χρησιμοποιούν συνήθως τη μονοδιάστατη αριθμητική μέθοδο του Newton(βλέπε Goodman 1979, Becker 1990). Γενικά εκτιμήσεις των παραμέτρων: σκορ γραμμών, στηλών, στρωμάτων και παραμέτρου συνάφειας εκτιμώνται και ανανεώνονται κατά τη διάρκεια κάθε επανάληψης του προγράμματος, ενώ

κάποιες άλλες άγνωστες παράμετροι αντιμετωπίζονται σαν καθορισμένα σκορ. Στο τέλος κάθε επανάληψης, τα στατιστικά του μοντέλου συγκρίνονται με αυτά από το προηγούμενο βήμα και το πρόγραμμα σταματά βάσει ενός προκαθορισμένου κριτηρίου. Περιορισμοί προσδιορισμού των παραμέτρων μπορούν να τεθούν σε κάθε επανάληψη ή στο τέλος της σύγκλισης του επαναληπτικού αλγορίθμου. Επίσης, προσεγγιστικά ασυμπτωτικά τυπικά σφάλματα των εκτιμήσεων των παραμέτρων μπορούν να ληφθούν με την μέθοδο Jackknife (βλέπε Clogg&Shihadeh 1994).

Τέλος, σαν γενικός κανόνας, οι βαθμοί ελευθερίας του μοντέλου ισούνται με το συνολικό αριθμό των κελιών μείον τον αριθμό των παραμέτρων, το οποίο εξαρτάται από τον αριθμό των περιορισμών προσδιορισμού που είχαν τεθεί.

3.7 Αλγόριθμος επιλογής βέλτιστου μοντέλου

Η ανάλυση πινάκων συνάφειας τριπλής εισόδου με τη βοήθεια των μοντέλων συνάφειας είναι αρκετά επίπονη και χρονοβόρα διαδικασία, όσον αφορά στην επιλογή του βέλτιστου μοντέλου συνάφειας με την καλύτερη προσαρμογή στα δεδομένα μας, αλλά και την καλύτερη ερμηνεία των σχέσεων μεταξύ των μεταβλητών του προβλήματος. Η παραπάνω διαδικασία γίνεται δυσκολότερη, καθώς αυξάνει το πλήθος των μεταβλητών. Για αυτό τον λόγο, προτείνουμε τον παρακάτω αλγόριθμο για την βέλτιστη επιλογή μοντέλου συνάφειας:

Βήμα 1

Επέλεξε το βέλτιστο λογαριθμογραμμικό μοντέλο, δηλαδή το λογαριθμογραμμικό μοντέλο με την καλύτερη προσαρμογή βάσει του στατιστικού X^2 του Pearson ή του λόγου πιθανοφάνειας G^2 και εντόπισε τον στατιστικά σημαντικότερο παράγοντα επίδρασης.

Βήμα 2

Κατέληξε σε δομή συνάφειας για αυτό τον παράγοντα, εξετάζοντας πρώτα όλες τις δυνατές περιπτώσεις και αντικατέστησε τον εν λόγω παράγοντα με τη δομή αυτή.

Βήμα 3

Εντόπισε τον επόμενο στατιστικά σημαντικότερο παράγοντα και επέστρεψε στο βήμα 2. Αλλιώς, τερμάτισε αν όλοι οι παράγοντες επίδρασης έχουν ελεγχθεί.

Τον παραπάνω αλγόριθμο, θα τον εφαρμόσουμε στο παράδειγμα της παραγράφου 3.8.

3.8 Παράδειγμα

Πριν ξεκινήσουμε την ανάλυση των δεδομένων του παραδείγματος, οφείλουμε να ορίσουμε τους παρακάτω συμβολισμούς προς διευκόλυνση του αναγνώστη. Συνεπώς, συμβολίζουμε με:

- R το μοντέλο επίδρασης γραμμών (*row effect*)
- C το μοντέλο επίδρασης στηλών (*column effect*)
- L το μοντέλο επίδρασης στρώματος (*layer effect*)
- U το μοντέλο ομοιόμορφης επίδρασης (*uniform effect*)
- R^{XY} το μοντέλο επίδρασης γραμμών στην αλληλεπίδραση XY
- C^{XY} το μοντέλο επίδρασης στηλών στην αλληλεπίδραση XY
- U^{XY} το μοντέλο ομοιόμορφης επίδρασης στην αλληλεπίδραση XY
- $R^{XY}+C^{XY}$ το μοντέλο επίδρασης γραμμών στην αλληλεπίδραση XY και επίδρασης στηλών στην αλληλεπίδραση XY
- $R^{XY}+C^{XY}+U^{XY}$ το μοντέλο επίδρασης γραμμών στην αλληλεπίδραση XY, επίδρασης στηλών στην αλληλεπίδραση XY και ομοιόμορφης επίδρασης στην αλληλεπίδραση XY, κ.ο.κ.

Τα δεδομένα του Πίνακα 3.3 αφορούν σε 1329 άνδρες του Framingham της Μασαχουσέτης, ηλικίας 40-59, βάσει του επιπέδου της χοληστερίνης (μεταβλητή X), της συστολικής πίεσης (μεταβλητή Y) και της παρουσίας ή όχι στεφανιαίας καρδιακής ασθένειας (μεταβλητή Z) κατά τη διάρκεια μιας εξαετούς περιόδου. Τα δεδομένα αυτά περιέχονται σε βιβλίο του Fienberg(1980), αλλά αρχικώς αναφέρθηκαν από τον Cornfield (1962). Επίσης, αναλύθηκαν και από τους Agresti&Kezouh(1983). Βάσει του αλγορίθμου της παραγράφου 3.7, έχουμε:

<u>Βήμα 1</u>

Όπως διαπιστώνουμε από τον Πίνακα 3.4, το λογαριθμογραμμικό μοντέλο XY+YZ+XZ προσαρμόζεται ικανοποιητικά στα δεδομένα μας και θα το επιλέξουμε για να συνεχίσουμε την ανάλυσή μας. Στη συνέχεια, θα προσπαθήσουμε να αντικαταστήσουμε τους όρους διπλής αλληλεπίδρασης με μια δομή συνάφειας. Ο στατιστικά σημαντικότερος παράγοντας αλληλεπίδρασης είναι ο XZ (p-value=0) και θα προσπαθήσουμε να του αποδώσουμε μια δομή συνάφειας.

		Συστολική πίεση σε mm Hg (Y)			
Στεφανιαία καρδιακή ασθένεια(Z)	**Χοληστερίνη σε mg/100cc (X)**	**<127**	**127-146**	**147-166**	**167+**
Παρούσα	**<200**	2	3	3	4
	200-219	3	2	0	3
	220-259	8	11	6	6
	>260	7	12	11	11
Απούσα	**<200**	117	121	47	22
	200-219	85	98	43	20
	220-259	119	209	68	43
	>260	67	99	46	33

Πίνακας 3.3

Ο πίνακας συνάφειας τριπλής εισόδου των μεταβλητών X, Y, Z.

Μοντέλο	**Έλεγχος καλής προσαρμογής**	**Τιμή**	**β.ε.**	**p-value**
XY+YZ+XZ	X^2	8,076	9	0,526
	G^2	6,564	9	0,682

Πίνακας 3.4

Ο έλεγχος καλής προσαρμογής του λογαριθμογραμμικού μοντέλου XY+YZ+XZ.

Βήμα 2

Αφού εξετάσαμε τις δομές συνάφειας που περιέχονται στον Πίνακα 3.5, καταλήγουμε στη δομή R^{XZ} επίδρασης γραμμής (βάσει των ελέγχων καλής προσαρμογής) στην αλληλεπίδραση XZ και αντικαθιστούμε τον εν λόγω παράγοντα με τη δομή αυτή.

Μοντέλο	Έλεγχος καλής προσαρμογής	Τιμή	β.ε.	p-value
R^{XZ}	X^2	8,076	9	0,526
	G^2	6,564	9	0,682
L^{XZ}	X^2	11,632	11	0,392
	G^2	10,008	11	0,530
U^{XZ}	X^2	11,632	11	0,392
	G^2	10,008	11	0,530

Πίνακας 3.5

Οι έλεγχοι καλής προσαρμογής των μοντέλων συνάφειας του βήματος 2.

Βήμα 3

Ο επόμενος στατιστικά σημαντικότερος παράγοντας είναι ο YZ (p-value=0,0003) και επαναλαμβάνοντας το βήμα 2, καταλήγουμε (βάσει των ελέγχων καλής προσαρμογής) στη δομή συνάφειας C^{YZ} επίδρασης στηλών στην αλληλεπίδραση YZ, όπως διαπιστώνουμε και από τον Πίνακα 3.6.

Ο επόμενος στατιστικά σημαντικότερος παράγοντας(p-value=0,0206) είναι ο XY και επαναλαμβάνοντας το βήμα 2, καταλήγουμε (βάσει των ελέγχων καλής προσαρμογής) στη δομή συνάφειας U^{XY} ομοιόμορφης επίδρασης στην αλληλεπίδραση XY, όπως διαπιστώνουμε και από τον Πίνακα 3.7.

Στο σημείο αυτό οφείλουμε να σημειώσουμε ότι το μοντέλο XY+YZ+XZ του Πίνακα 3.4 συμπίπτει με το μοντέλο R^{XZ} του Πίνακα 3.5 και με το μοντέλο $R^{XZ}+C^{YZ}$ του Πίνακα 3.6, γιατί η μεταβλητή Z έχει μόνο δύο επίπεδα.

Ο αλγόριθμος τερματίζει, αφού αποδώσαμε δομή συνάφειας σε όλους τους παράγοντες διπλής αλληλεπίδρασης και επιλέγουμε τελικά το μοντέλο συνάφειας $R^{XZ}+ C^{YZ}+U^{XY}$ επίδρασης γραμμών στην αλληλεπίδραση XZ, επίδρασης στηλών στην αλληλεπίδραση YZ και ομοιόμορφης επίδρασης στην αλληλεπίδραση XY (Πίνακας 3.7).

Μοντέλο	Έλεγχος καλής προσαρμογής	Τιμή	β.ε.	p-value
$R^{XZ}+C^{YZ}$	X^2	8,076	9	0,526
	G^2	6,564	9	0,682
$R^{XZ}+L^{YZ}$	X^2	11,097	11	0,435
	G^2	10,217	11	0,511
$R^{XZ}+U^{YZ}$	X^2	11,097	11	0,435
	G^2	10,217	11	0,511

Πίνακας 3.6

Οι έλεγχοι καλής προσαρμογής των μοντέλων συνάφειας του βήματος 3.

Μοντέλο	Έλεγχος καλής προσαρμογής	Τιμή	β.ε.	p-value
$R^{XZ}+ C^{YZ}+R^{XY}$	X^2	15,568	15	0,411
	G^2	14,284	15	0,504
$R^{XZ}+ C^{YZ}+C^{XY}$	X^2	13,970	15	0,528
	G^2	12,538	15	0,638
$R^{XZ}+ C^{YZ}+U^{XY}$	X^2	15,802	17	0,538
	G^2	14,513	17	0,630

Πίνακας 3.7

Οι έλεγχοι καλής προσαρμογής των μοντέλων συνάφειας του βήματος 3.

3.9 Ανακεφαλαίωση

Σύγχρονες στατιστικές μέθοδοι για κατηγορικές μεταβλητές σε πίνακες συνάφειας έχουν αναπτυχθεί σημαντικά από την εισαγωγή των τεχνικών των λογαριθμικών μοντέλων αρκετές δεκαετίες πριν. Νέα και δυναμικά στατιστικά μοντέλα που αναλύουν τα log-odds-ratio κατέστησαν ικανούς τους ερευνητές κοινωνικών επιστημών να επιτύχουν μια καλύτερη κατανόηση της συστηματικής σχέσης μεταξύ των μεταβλητών συμπεριλαμβανομένων μοτίβων και επιπέδων της συνάφειας. Συγκεκριμένα, παραμετρικά μοντέλα που τονίζουν την ιδιότητα

της διάταξης των κατηγορικών μεταβλητών συχνά έχουν ως αποτέλεσμα το δυναμικό και οικονομικό (***parsimonious***) χαρακτηρισμό των σχέσεων τους, ενώ τα απαραμετρικά μοντέλα που δεν κάνουν τέτοιες υποθέσεις για ποιοτικές/διακριτές μεταβλητές προσφέρουν συχνά λιγότερο οικονομική αλλά ωστόσο ανταγωνιστική ερμηνεία. Μεταξύ διαφόρων παραμετρικών μοντέλων που αναπτύχθηκαν έως τώρα, αυτό που έλαβε ευρεία προσοχή είναι το λογαριθμο-πολλαπλασιαστικό μοντέλο επιδράσεων γραμμής και στήλης, το γνωστό μοντέλο συνάφειας RC που αναπτύχθηκε από τον Leo A.Goodman(1979, 1985, 1986) και επεκτάθηκε από άλλους, όπως η Anderson(1996), ο Becker(1989, 1992), οι Becker&Clogg(1989), ο Clogg(1982), οι Goodman&Hout(1998), ο Wong(1995) και ο Xie(1992).

Διάφοροι τύποι RC πολυδιάστατων πολυγραμμικών μοντέλων συνάφειας με διγραμμικούς και τριγραμμικούς όρους σε πίνακες πολλαπλής εισόδου έχουν αναπτυχθεί ώστε να περιγράψουν το σύνθετο μοτίβο συνάφειας όσον αφορά στους τομείς της στατιστικής και της ψυχομετρίας και συνεργάζονται μεταξύ τους στο γενικότερο πλαίσιο πολυγραμμικότητας. Αποτελούν χρήσιμα εργαλεία για κοινωνιολόγους και ερευνητές κοινωνικών επιστημών. Στη συνέχεια, παρουσιάζονται τα μοντέλα με διγραμμικούς και τριγραμμικούς όρους και αυτή η γενικότερη πολυγραμμική προσέγγιση μπορεί να επεκταθεί σε πίνακες συνάφειας τετραπλής εισόδου και υψηλότερης τάξης. Τέλος, δόθηκε ένας αλγόριθμος επιλογής βέλτιστου μοντέλου ώστε να διευκολυνθεί αυτή η χρονοβόρα και επίπονη διαδικασία λόγω της αύξησης της διαστασιμότητας του προβλήματος.

ΚΕΦΑΛΑΙΟ 4ο

Γραφήματα Συνάφειας και Γραφικά Μοντέλα

4.0 Εισαγωγή

Τα λογαριθμογραμμικά μοντέλα για πίνακες συνάφειας τετραπλής εισόδου και άνω είναι δύσκολα ως προς τη φυσική τους ερμηνεία και για αυτό αναζητούμε τρόπους ώστε να απεικονίσουμε γραφικά τις μεταβλητές του προβλήματος, να μελετήσουμε τη σχέση μεταξύ τους και ίσως να μειώσουμε τελικά τη διαστασιμότητα του προβλήματος. Για αυτό το σκοπό, θα ορίσουμε τα γραφήματα συνάφειας (*association graphs*), τα οποία συμβάλλουν πολύ στην εύκολη κατανόηση τέτοιων περίπλοκων δομών και στη μελέτη των συναφειών που διέπουν τις μεταβλητές. Η γραφική αναπαράσταση των συναφειών στα λογαριθμογραμμικά μοντέλα υποδεικνύει τα ζεύγη των υπό συνθήκη ανεξάρτητων μεταβλητών. Ο Darroch et all(1980) χρησιμοποίησαν τη θεωρία των μαθηματικών γραφημάτων για να αναπαραστήσουν συγκεκριμένα μοντέλα, τα οποία καλούνται γραφικά μοντέλα (*graphical models*) και έχουν μια δομή υπό συνθήκη ανεξαρτησίας.

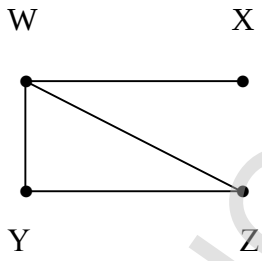
Ένα σημαντικό πρόβλημα ωστόσο που προκύπτει στα γραφήματα αυτά, τα οποία περιλαμβάνουν παραπάνω από τρεις μεταβλητές, είναι το γεγονός ότι διαφορετικά μοντέλα που περιλαμβάνουν διαφορετικούς όρους αλληλεπίδρασης οδηγούν στο ίδιο γράφημα και άρα στην ίδια ερμηνεία όσον αφορά στην υπό συνθήκη και περιθώρια ανεξαρτησία. Με την έννοια αυτή, διάφορα μοντέλα στην πραγματικότητα είναι ισοδύναμα με το πλήρες μοντέλο. Αυτό σημαίνει πως για τα μοντέλα αυτά, δεν υπάρχει πιο απλουστευμένη αναπαράσταση των συναφειών στον πίνακα με όρους υπό συνθήκης και περιθώριας ανεξαρτησίας.

4.1 Γράφημα συνάφειας

Ας υποθέσουμε ότι έχουμε τέσσερις μεταβλητές X, Y, Z, W. Θα προσπαθήσουμε να ορίσουμε και να κατασκευάσουμε γραφήματα συνάφειας (βλέπε Simonoff, 2003), αλλά και να δώσουμε μια σύντομη αναφορά στα γραφικά μοντέλα.

4.1.1 Ορισμός των γραφημάτων συνάφειας

Το γράφημα συνάφειας είναι ένα σύνολο από κορυφές (*nodes*), όπου κάθε κορυφή αναπαριστά μια μεταβλητή. Μια ακμή (*edge*) που συνδέει δυο μεταβλητές αναπαριστά μια υπό συνθήκη συνάφεια (*association*) μεταξύ τους. Επίσης, δύο μεταβλητές είναι ανεξάρτητες εάν δεν συνδέονται καθόλου στο γράφημα. Ορίζουμε ως κλίκα (*clique*) σε ένα γράφημα το μέγιστο υποσύνολο κορυφών που ενώνονται όλες μεταξύ τους. Δύο μεταβλητές X και Y είναι διαχωρίσιμες (*separated*) από ένα σύνολο μεταβλητών αν όλα τα μονοπάτια που συνδέουν τις μεταβλητές X και Y διχοτομούν αυτό το υποσύνολο. Στη συνέχεια, παραθέτουμε το γράφημα συνάφειας για το μοντέλο (WX, WY, WZ, YZ):



Γράφημα 4.1

Το γράφημα συνάφειας που αντιστοιχεί στο μοντέλο (WX, WY, WZ, YZ).

Οι ακμές μεταξύ δύο κορυφών δηλώνουν υπό συνθήκη συνάφεια και οι ακμές που λείπουν δηλώνουν υπό συνθήκη ανεξαρτησία των μεταβλητών αυτών. Δηλαδή, οι μεταβλητές X και Y, αλλά και X και Z είναι υπό συνθήκη ανεξάρτητες δοθέντων των υπολοίπων μεταβλητών. Η W διαχωρίζει τις X και Y, αφού οποιοδήποτε μονοπάτι που συνδέει τις X και Y, πηγαίνει διαμέσου τις W. Το υποσύνολο {W, Z} επίσης διαχωρίζει τις μεταβλητές X και Y. Ένα θεμελιώδες πόρισμα αναφέρει ότι δύο μεταβλητές είναι υπό συνθήκη ανεξάρτητες όταν οποιοδήποτε υποσύνολο μεταβλητών τις διαχωρίζει. Συνεπώς, οι X και Y είναι υπό συνθήκη ανεξάρτητες όχι μόνο δοθέντων των W και Z, αλλά και μόνο της W. Ομοίως, οι X και Z είναι υπό συνθήκη ανεξάρτητες δοθείσης της W μόνο.

Τέλος, όπως προαναφέρθηκε στην εισαγωγή, δύο λογαριθμογραμμικά μοντέλα με τις ίδιες κατά ζεύγη συνάφειες μπορεί να έχουν το ίδιο γράφημα. Το μοντέλο με το ίδιο γράφημα με το (WX, WY, WZ, YZ) είναι και το (WX, WYZ), στο οποίο έχει προστεθεί και ο όρος τριπλής αλληλεπίδρασης WYZ.

Όλα τα (ιεραρχικά) μοντέλα τεσσάρων μεταβλητών δίνονται στον Πίνακα 4.1 και τα αντίστοιχα γραφήματα συνάφειας για τα μοντέλα αυτά δίνονται στο γράφημα 4.2 και που

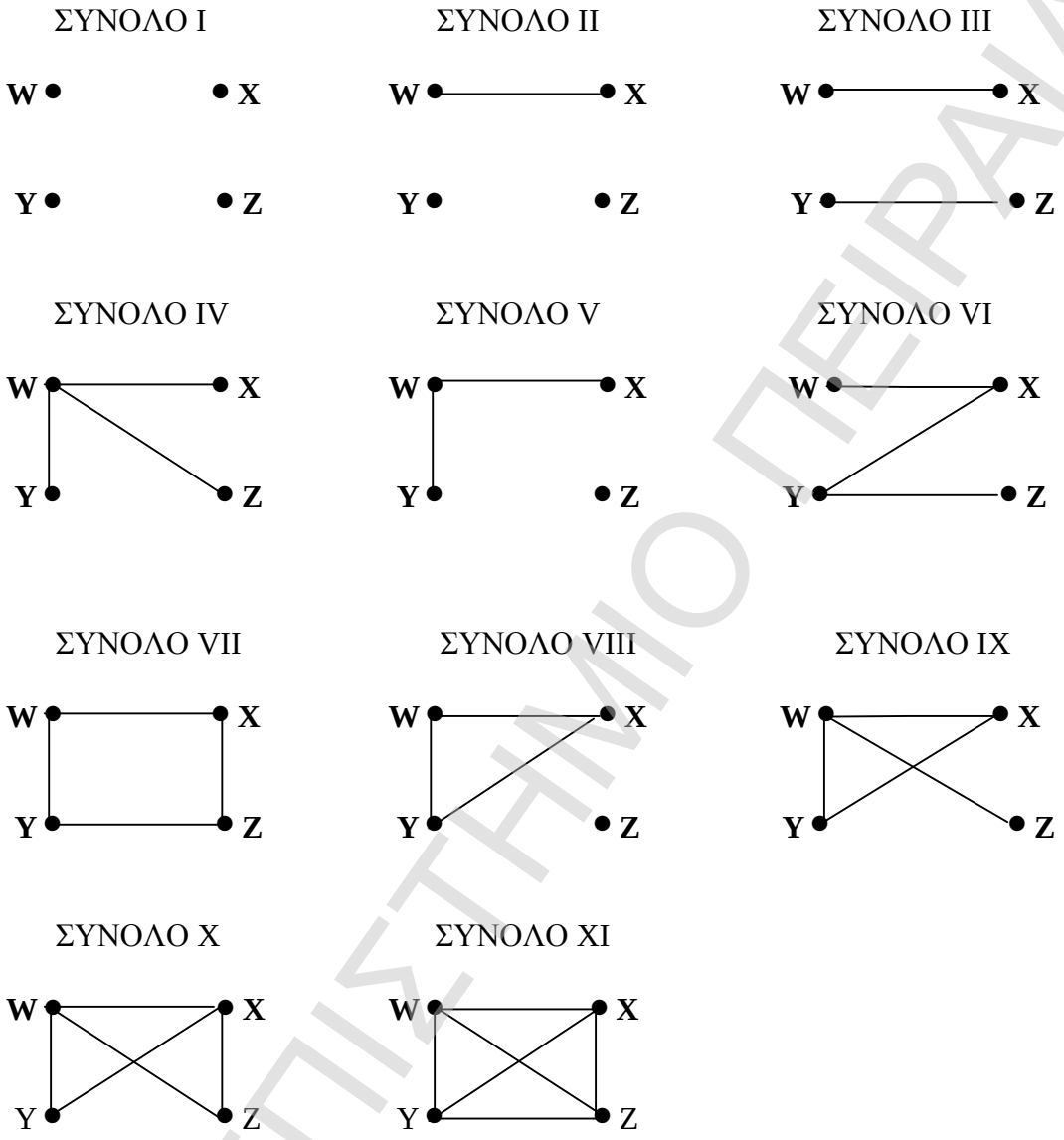
ακολουθεί. Όταν περισσότερα από ένα μοντέλα οδηγούν στο ίδιο γράφημα, υπάρχει ένα πιο απλό μοντέλο και ένα πιο σύνθετο για το γράφημα αυτό. Τα πιο απλά μοντέλα είναι τελευταία στη σειρά στη λίστα του Πίνακα 4.1 και αναπαριστούν όλες τις ακμές του γραφήματος με διπλές αλληλεπιδράσεις. Το πιο περίπλοκο μοντέλο από τα παραπάνω μοντέλα καλείται γραφικό μοντέλο και θα οριστεί στη συνέχεια.

ΣΧΗΜΑ	**ΜΟΝΤΕΛΟ**	**ΠΕΡΙΓΡΑΦΗ**
I	(W, X, Y, Z)	W ⊥ X ⊥ Y ⊥ Z
II	(WX, Y, Z)	W, X ⊥ Y, Z&Y ⊥ Z
III	(WX, YZ)	W, X ⊥ Y, Z
IV	(WX, WY, WZ)	X ⊥ Y ⊥ Z/W
V	(WX, WY, Z)	Z ⊥ W, X, Y & X ⊥ Y/W
VI	(WX, XY, YZ)	Z ⊥ W, X/Y& W ⊥ Y, Z/X
VII	(WX, WY, XZ, YZ)	W ⊥ Z/X, Y& X ⊥ Y/W, Z
VIII	(WXY, Z), (WX, WY, XY, Z)	Z ⊥ W, X, Y
IX	(WXY, WZ), (WX, WY, WZ, XY)	Y ⊥ Z/W, X ⊥ Z/W
X	(WXY, WXZ), (WXY, WZ, XZ), (WX, WY, WZ, XY, XZ)	Y ⊥ Z/W, X
XI	(WXYZ), (WXY, WXZ, WYZ, XYZ), (WXY, WXZ, WYZ), (WXY, WXZ, YZ), (WXY, WZ, XZ, YZ), (WX, WY, WZ, XZ, YZ)	

Πίνακας 4.1

Ιεραρχικά μοντέλα[(1)] με την αντίστοιχη περιγραφή για καθένα από γραφήματα συνάφειας του γραφήματος 4.2.

[(1)]Στον Πίνακα 4.1, το σύμβολο $\perp$ δηλώνει την ανεξαρτησία των μεταβλητών και το σύμβολο / σημαίνει «δοθείσης».



Γράφημα 4.2

Τα γραφήματα συνάφειας για τα αντίστοιχα (ιεραρχικά) μοντέλα του Πίνακα 4.1.

4.1.2 Χρησιμότητα των γραφημάτων συνάφειας

Πίνακες συνάφειας τριπλής ή μεγαλύτερης εισόδου δεν είναι πολύ εύκολο να αναπαρασταθούν γραφικά και για αυτόν το λόγο είναι καλή ιδέα να προσπαθήσουμε να

απεικονίσουμε τα δεδομένα τέτοιων πινάκων συνάφειας σε ένα χώρο μικρότερης διάστασης. Υπάρχουν δύο τρόποι για να μειωθεί η διαστασιμότητα του προβλήματος.

Πρώτον, αν στο μοντέλο δυο μεταβλητές, για παράδειγμα η X και W, συνδέονται (*associated*) μεταξύ τους, τότε ο διδιάστατος παράγοντας XW μπορεί να αντιμετωπιστεί σαν ένας παράγοντας με ij επίπεδα, αν i είναι τα επίπεδα της X και j τα επίπεδα της W. Για παράδειγμα, στο μοντέλο (WXY, WXZ), οι μεταβλητές W και X αντιμετωπίζονται σαν ζεύγος με την έννοια ότι μια μεταβλητή έχει μια πιθανή συνάφεια με την W αν και μόνο αν έχει μια συνάφεια με την X. Το διάγραμμα (σχήμα X, γράφημα 4.2) του μοντέλου (WXY, WXZ) δείχνει ότι η μεταβλητή Y συνδέεται και με την W και με την X και η Z συνδέεται και με την W και με την X. Αυτό σημαίνει πως ο τετραπλός πίνακας εισόδου μπορεί να θεωρηθεί ως τρισδιάστατος με παράγοντες τους W×X, Y και Z. Η ερμηνεία του μοντέλου αυτού είναι πως ότι η Y και Z είναι υπό συνθήκη ανεξάρτητες, δοθέντος του κοινού επιπέδου της W×X.

Επίσης, το μοντέλο (WXY, WZ) μπορεί να αντιμετωπιστεί τρισδιάστατα με παρόμοιο τρόπο. Οι μεταβλητές X και Y είναι από κοινού μοντελοποιημένες σε αυτό, αφού συνδέονται μεταξύ τους, η W συνδέεται με καθεμία από τις X, Y, Z και η Z συνδέεται μόνο με την W. Αυτό σημαίνει πω ο πίνακας τετραπλής εισόδου μπορεί να θεωρηθεί ως πίνακας τριπλής εισόδου με παράγοντες X×Y, W, Z, όπου η Z είναι υπό συνθήκη ανεξάρτητη της X×Y δοθέντος της W.

Ο δεύτερος τρόπος για να θεωρήσουμε τον τετρασδιάστατο πίνακα χρησιμοποιώντας λιγότερες διαστάσεις είναι να συμπτυχθεί ο πίνακας αυτός σε δύο ή περισσότερους περιθώριους πίνακες. Φυσικά, πρέπει να δοθεί προσοχή ώστε να συμπτυχθεί μόνο σε περιθώριους πίνακες που έχουν την ίδια δομή συνάφειας μεταξύ των μεταβλητών όπως στον πλήρη πίνακα. Αν υποτεθεί ότι οι μεταβλητές σε έναν τετρασδιάστατο πίνακα ή μεγαλύτερης διάστασης μπορούν να διαχωριστούν σε τρία αμοιβαίως αποκλειόμενα υποσύνολα A, B, C, έτσι ώστε το B να διαχωρίζει τα A και C. Τότε, αν ο πίνακας συμπτύσσει τις μεταβλητές στο C, επιδράσεις που σχετίζονται με τις μεταβλητές στο A και με συνάφειες μεταξύ των μεταβλητών στα A και B μένουν αμετάβλητες.

Για παράδειγμα, στα μοντέλα που αντιστοιχούν στο σύνολο (WXY, WXZ), ο παράγοντας WX βρίσκεται «μεταξύ» των Y και Z στο γράφημα συνάφειας και άρα ο πίνακας μπορεί να συμπτυχθεί είτε στον Y (ώστε να μελετηθεί η συνάφεια WX×Z) ή τον Z (ώστε να μελετηθεί η συνάφεια WX×Y). Στο μοντέλο (WXY, WZ), ο παράγοντας W βρίσκεται μεταξύ των XY και Z, έτσι ώστε ο πίνακας να μπορεί να συμπτυχθεί ως προς τους XY ή Z. Το μοντέλο (WX, XY, YZ) έχει τέτοιο γράφημα συνάφειας (σχήμα VI, Γράφημα 4.2) που υποδεικνύει εύλογη σύμπτυξη ως προς τους W, WX, WXY, Z, YZ ή XYZ. Τα διαγράμματα

για τα μοντέλα (WX, WY, WZ) και (WX, WY, Z) (σχήμα IV και V, Γράφημα 4.2) υποδηλώνουν σύμπτυξη των X, Y ή Z, ενώ το μοντέλο (WX, YZ) επιτρέπει σύμπτυξη των WX ή YZ.

Συνεπώς, η βέλτιστη λύση ώστε να αναλυθεί οποιοσδήποτε πίνακας συνάφειας τεσσάρων μεταβλητών και άνω είναι να γίνει προσπάθεια ώστε να βρεθούν μοντέλα που να υποδεικνύουν κατανοητές και κατάλληλες συνθήκες ανεξαρτησίας, και επιτρέπουν την σύμπτυξη με τέτοιους τρόπους, έτσι ώστε ο τύπος του μοντέλου τελικά να απλουστευθεί με σωστό και εύχρηστο τρόπο. Ωστόσο, αποτελεί πρόκληση να κατανοηθεί η συνάφεια τριών αλληλεπιδράσεων χωρίς κάποια δομή απλοποίησης, αλλά είναι αρκετά δύσκολο να συλληφθεί η φυσική ερμηνεία για μία τετραπλή (και άνω) αλληλεπίδραση. Τα μοντέλα και οι αρχές που τα διέπουν γενικεύονται σε περισσότερες από τέσσερις διαστάσεις, αλλά με τίμημα να αυξηθεί η πολυπλοκότητά τους.

4.1.3 Παράδειγμα

Η γέννηση νεκρού εμβρύου (***stillbirth***) θεωρείται ως η απεβίωση ενός εμβρύου μετά την $20^{η}$ εβδομάδα κυοφορίας. Ένα παιδί που γεννιέται ζωντανό μετά την $20^{η}$ εβδομάδα και πριν την $37^{η}$ εβδομάδα, θεωρείται σαν πρόωρη γέννηση. Είναι γνωστό πως διάφοροι περιβαλλοντικοί λόγοι μπορεί να αυξήσουν τις πιθανότητες για γέννηση νεκρού εμβρύου, όπως η ηλικία της μητέρας (άνω των 35 ετών), η διατροφή, το κάπνισμα, το αλκοόλ, χρήση ναρκωτικών, μεταδιδόμενες ή κληρονομικές ασθένειες, κλπ. Ο Coory (1998) αναφέρει τα αποτελέσματα από την εξέταση των γεννήσεων νεκρών εμβρύων στην δεύτερη μεγαλύτερη σε έκταση πολιτεία της Αυστραλίας, την Queensland, για τα έτη 1987-1992, τα οποία παρουσιάζονται στον Πίνακα 4.2. Αυτά τα δεδομένα περιλαμβάνουν κάθε γέννηση στην Queensland για τα έτη αυτά. Εκτός από την κατάσταση των γεννήσεων (μεταβλητή B): νεκρά έμβρυα (***Stillbirth***) ή νεογνά (***Livebirth***), ο Πίνακας 4.2 περιέχει και την διάρκεια κυοφορίας του εμβρύου σε εβδομάδες (μεταβλητή A), την φυλή των γονέων: λευκοί ή αυτόχθονες (μεταβλητή R) και το φύλο του εμβρύου (μεταβλητή G). Τέλος, γεννήσεις με διάρκεια κυοφορίας άνω των 41 εβδομάδων δεν αναφέρονται, γιατί αναφέρθηκαν μόνο δύο.

Η πολιτεία Queensland έχει τον μεγαλύτερο αριθμό σε γεννήσεις από αυτόχθονες κατοίκους σε όλη την Αυστραλία, περίπου το 4% των γεννήσεων στην πολιτεία, οι οποίοι όμως έχουν το πιο χαμηλό οικονομικό επίπεδο από όλους τους κατοίκους της Queensland, όποτε και είναι λογικό να συγκεντρώνουν μεγαλύτερους κινδύνους σε θέματα υγείας, άρα και στις γεννήσεις. Το συγκεκριμένο παράδειγμα έχει αναλυθεί και στο βιβλίο του J.S.Simonoff, *Analysing Categorical Data,* 2003.

		Φυλή (R)			
		Αυτόχθονες		Λευκοί	
Φύλο Εμβρύου(G)	**Διάρκεια Κυοφορίας σε εβδομ. (A)**	**Νεκρά Έμβρυα**	**Νεογνά**	**Νεκρά Έμβρυα**	**Νεογνά**
Αρσενικό	**≤24**	22	16	171	121
	25-28	21	42	109	358
	29-32	12	73	95	944
	33-36	4	387	112	5155
	37-41	7	3934	169	102
Θηλυκό	**≤24**	17	16	167	107
	25-28	13	19	100	314
	29-32	10	76	78	727
	33-36	10	451	92	4224
	37-41	13	3729	209	96

Πίνακας 4.2

Τα δεδομένα του παραδείγματος για τις μεταβλητές A, B, R, G.

Ο Πίνακας 4.3 περιλαμβάνει τα αποτελέσματα από τα (ιεραρχικά) λογαριθμικά μοντέλα που προσαρμόστηκαν στα δεδομένα του Πίνακα 4.2 για τις μεταβλητές A, B, R, G. Όπως διαπιστώνουμε, θα πρέπει να επιλεχθεί να προσαρμοστεί ένα μοντέλο τουλάχιστον μιας αλληλεπίδρασης τρίτης τάξης. Το μοντέλο (ABR, ARG, BG) είναι αποδεκτό αφού έχει p-value=0.779, αλλά ανήκει στα μοντέλα με γράφημα XI του Πίνακα 4.1, που σημαίνει ότι δεν υπάρχει κάποιου είδους συνάφεια και δεν μπορεί να γίνει σύμπτυξη (*collapsibility*) οποιουδήποτε περιθωρίου.

Το μοντέλο (ABR, ARG) (Πίνακας 4.3) είναι στατιστικά αποδεκτό και ανήκει στα μοντέλα με γράφημα X του Πίνακα 4.1, οπότε μπορεί να ερμηνευθεί κάποιου είδους συνάφειας. Η επιλογή του μοντέλου αυτού έγινε διότι σε όλα τα μοντέλα υψηλότερης τάξης, όταν περιέχεται η αλληλεπίδραση ARG είναι αποδεκτά. Το γράφημα συνάφειας που αντιστοιχεί στο μοντέλο αυτό είναι το σχήμα X (Πίνακας 4.1) του γραφήματος 4.2, αν θέσουμε W=A, X=R, Y=G, Z=B. Από αυτό το γράφημα συνάφειας διαπιστώνουμε πως θα

μπορούσαμε να μελετήσουμε την επίδραση ARG, αν συμπτύξουμε τα δεδομένα μας ως προς τη μεταβλητή B, αφού παρατηρούμε ότι το φύλο(G) και η κατάσταση γέννησης του εμβρύου(B) είναι υπό συνθήκη ανεξάρτητες, δοθέντος της διάρκειας κυοφορίας (A) και της φυλής (R).

ΜΟΝΤΕΛΟ	**X^2**	**Β.ε**	**p-value**	**AIC**[(1)]
(A, B, R, G)	22372.076	32	0.000	3.314670
(AB, AR, AG, BR, BG, RG)	4487.789	17	0.000	6.125337
(ABR, ABG, BRG, ARG)	3.148	4	0.533	10.991060
(ABR, ARG, ABG)	3.442	5	0.632	10.907410
(ABR, ARG, BRG)	5.235	8	0.732	10.554400
(ABG, ARG, BRG)	4288.793	8	0.000	0.159810
(ABR, ABG, $BRG)	21.844	8	0.005	9.309684
(ABR, ARG, BG)	5.598	9	0.779	10.491150
(ABR, ARG)	7.687	10	0.659	10.218040
(ARG, AB, GB, BR)	4295. 579	13	0.000	6.154267
(ARG, AB, BR)	4310.059	14	0.000	6.151887
(ARG, AB, BG)	4973.591	14	0.000	5.322212
(ARG, BG, BR)	6969.149	17	0.000	4.903698
(ABR, AG, BG, RG)	33.522	13	0.001	8.941216
(ABR, AG, RG)	41.344	14	0.000	8.761187
(ABR, BG, RG)	36.277	17	0.004	8.872895
(ABR, AG, BG)	42.093	14	0.000	8.748367

[(1)]Μικρότερη τιμή στο AIC αντιστοιχεί σε βέλτιστο μοντέλο.

Πίνακας 4.3

Έλεγχος καλής προσαρμογής X^2 και κριτήριο AIC για τα (ιεραρχικά) λογαριθμικά μοντέλα που προσαρμόστηκαν στα δεδομένα του Πίνακα 4.2 για τις μεταβλητές A, B, R, G.

Ο Πίνακας 4.4 περιέχει τα δεδομένα του Πίνακα 4.2, αφού συμπτύξουμε ως προς τη μεταβλητή Β, και δε δίνει καμία πληροφορία για την κατάσταση του εμβρύου όταν γεννηθεί, αλλά εξακολουθεί και δείχνει πως σχετίζονται οι άλλες τρεις μεταβλητές μεταξύ τους (A, R, G).

Διάρκεια Κυοφορίας σε εβδομ. (A)	**Φύλο Εμβρύου (G)**	**Φυλή (R)**	
		Αυτόχθονες	**Λευκοί**
≤24	**Αρσενικό**	38	292
	Θηλυκό	33	274
25-28	**Αρσενικό**	63	167
	Θηλυκό	32	414
29-32	**Αρσενικό**	85	1039
	Θηλυκό	86	805
33-36	**Αρσενικό**	391	5267
	Θηλυκό	461	4316
37-41	**Αρσενικό**	3941	102
	Θηλυκό	3742	97

Πίνακας 4.4

Τα δεδομένα του Πίνακα 4.2 για τις μεταβλητές A, R, G,
συμπτυγμένα ως προς τη μεταβλητή Β.

Στο σημείο αυτό, οφείλουμε να σημειώσουμε ότι η σύμπτυξη ως προς την αποκριτική μεταβλητή Β, από μαθηματική σκοπιά είναι ορθή, αλλά στατιστικά δεν ενδείκνυται. Ωστόσο, προχωρούμε στην σύμπτυξη αυτή, απλά για να διαπιστώσουμε με ποιον τρόπο σχετίζονται οι άλλες τρεις μεταβλητές μεταξύ τους. Από την μελέτη του Πίνακα 4.4 διαπιστώνουμε πως για γεννήσεις μεταξύ $37^{ης}$ έως $41^{ης}$ εβδομάδας, δεν υπάρχει συνάφεια μεταξύ της φυλής και του φύλου του εμβρύου (αφού $\theta \approx 1$). Επίσης, ο συνολικός αριθμός των γεννήσεων αυξάνει καθώς αυξάνει και η διάρκεια κυοφορίας, πράγμα άλλωστε αναμενόμενο. Ανεξαρτήτως φύλου, υπάρχει υψηλή συνάφεια μεταξύ της φυλής και της διάρκειας κυοφορίας, με πιο αυξημένες τις γεννήσεις από αυτόχθονες, καθώς αυξάνει η διάρκεια κυοφορίας.

4.2 Γραφικά μοντέλα με τη βοήθεια των γραφημάτων συνάφειας

Στη συνέχεια, θα δώσουμε μια σύντομη αναφορά στα γραφικά μοντέλα με τη βοήθεια των γραφημάτων συνάφειας, που ορίστηκαν στην παράγραφο 4.1.

4.2.1 Χρησιμότητα των γραφικών μοντέλων

Στατιστικές εφαρμογές σε πεδία όπως η βιοπληροφορική, η ανάκτηση πληροφορίας, η επεξεργασία λόγου, η επεξεργασία εικόνας και οι επικοινωνίες, συχνά απαιτούν μοντέλα με πάρα πολλές μεταβλητές, οι οποίες συνδέονται μεταξύ τους με περίπλοκους τρόπους. Τα γραφικά μοντέλα παρέχουν μια γενική μεθοδολογία για να προσεγγιστούν τέτοια προβλήματα και πολλά από αυτά τα μοντέλα που αναπτύχθηκαν από στατιστικούς ερευνητές αποτελούν παραδείγματα για την τυποποίηση των γραφικών μοντέλων.

Επίσης, το πεδίο της στατιστικής και της επιστήμης των υπολογιστών έχει ακολουθήσει γενικά διαφορετικά μονοπάτια τις τελευταίες δεκαετίες, αλλά κάθε πεδίο παρέχει χρήσιμες υπηρεσίες στο άλλο. Οι στατιστικοί ολοένα και περισσότερο ενδιαφέρονται για προοπτικές σε υπολογιστικές εφαρμογές, θεωρητικές και πρακτικές, όσον αφορά σε μοντέλα και οι επιστήμονες των υπολογιστών ενδιαφέρονται για συστήματα που επικοινωνούν με το εξωτερικό περιβάλλον και ερμηνεύουν αβέβαια δεδομένα με όρους πιθανοτικών μοντέλων. Η περιοχή όπου είναι έκδηλα όλα τα παραπάνω είναι τα γραφικά μοντέλα.

4.2.2 Ορισμός γραφικού μοντέλου μέσω γραφήματος συνάφειας

Ένα γραφικό μοντέλο αποτελεί μια οικογένεια από κατανομές πιθανότητας που ορίζονται με όρους ενός γραφήματος συνάφειας. Οι κορυφές σε ένα γράφημα αναπαριστούν τις τυχαίες μεταβλητές και οι από κοινού πυκνότητες πιθανότητας ορίζονται ως το γινόμενο των συναρτήσεων που ορίζονται στα υποσύνολα των κορυφών που συνδέονται μεταξύ τους.

Υπάρχουν δύο τύποι γραφικών μοντέλων, τα κατευθυνόμενα (***directed***) και τα μη κατευθυνόμενα (***undirected***) γραφικά μοντέλα, που στηρίζονται αντίστοιχα στα κατευθυνόμενα ακυκλικά γραφήματα και στα μη κατευθυνόμενα γραφήματα.

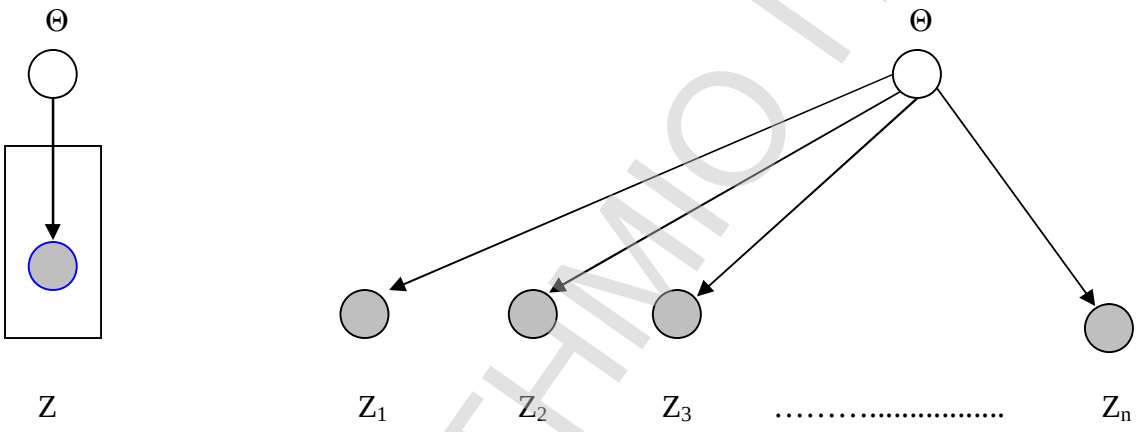
Ορισμός 4.1 (Κατευθυνόμενα γραφήματα)

Έστω G(V, E) ένα κατευθυνόμενο ακυκλικό γράφημα, όπου V είναι το σύνολο των κορυφών (***nodes***) του γραφήματος και E είναι το σύνολο των ακμών (***edges***) του γραφήματος. Έστω $\{X_\upsilon : \upsilon \in V\}$ είναι το σύνολο των τυχαίων μεταβλητών που ορίζονται από τις κορυφές του γραφήματος. Για κάθε κορυφή $\upsilon \in V$, έστω π_υ το υποσύνολο των δεικτών των προγόνων (***parents***) της. Επιτρέπουμε το σύνολο των δεικτών να εμφανίζεται οποτεδήποτε εμφανίζεται

ένας μονός δείκτης, συνεπώς, X_{p_u} δηλώνει το διάνυσμα των τυχαίων μεταβλητών που κατατάσσεται ως πίνακας (***indexed***) από τους προγόνους του υ. Δοθέντος του συνόλου των πυρήνων $\{k(x_\upsilon/x_{p_u}):\upsilon\in V\}$ που αθροίζουν στο 1 (στη διακριτή περίπτωση) ή το ολοκλήρωμά τους είναι 1 (στη συνεχή περίπτωση), ορίζουμε την από κοινού κατανομή πιθανότητας ως

$$p(X_V)=\prod_{u\in V} k(x_\upsilon/x_{\pi_\upsilon}) \qquad (4.1).$$

Είναι εύκολο να επιβεβαιώσουμε ότι αυτή η από κοινού κατανομή πιθανότητας (4.1) έχει $\{k(x_\upsilon/x_{p_u})\}$ως παράγοντες της, που σημαίνει $k(x_\upsilon/x_{p_u})=p(x_\upsilon/x_{p_u})$. Στο σημείο αυτό αξίζει να σημειώσουμε ότι δεν γίνεται διάκριση μεταξύ των δεδομένων και των παραμέτρων και είναι φυσικό να συμπεριλάβουμε και παραμέτρους μεταξύ των κορυφών του γραφήματος.



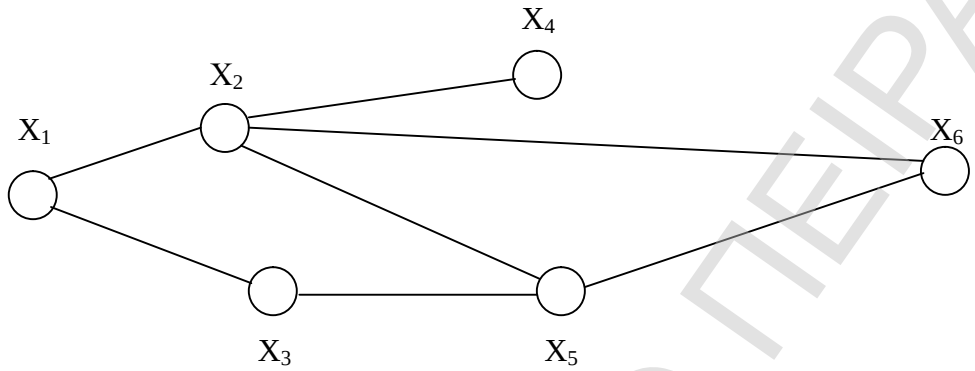
Γράφημα 4.3

Κατευθυνόμενο γράφημα συνάφειας για τις μεταβλητές Θ, Z_1, Z_2, Z_3,…, Z_n.

Το πιάτο(***plate***) του γραφήματος (4.3) είναι ένα χρήσιμο τέχνασμα για να συλλάβουμε την επανάληψη στα γραφικά μοντέλα, συμπεριλαμβανομένων παραγοντικών και εμφωλευμένων (***nested***) δομών που συναντάμε κυρίως στους πειραματικούς σχεδιασμούς. Τα κατευθυνόμενα γραφικά μοντέλα είναι συνηθισμένα ως αναπαραστάσεις των ιεραρχικών μπεϋζιανών μοντέλων. Το γράφημα αποτελεί μια επικαλούμενη οπτική αναπαράσταση της από κοινού κατανομής πιθανότητας, αλλά παρέχει και άλλες σημαντικές πληροφορίες ακόμη. Όποιοι και αν είναι οι συναρτησιακοί τύποι των πυρήνων $p(x_\upsilon/x_{p_u})$, η παραγοντοποίηση της σχέσης (4.1), συνεπάγεται ένα σύνολο από υποθέσεις υπό συνθήκη ανεξαρτησίας μεταξύ των μεταβλητών X_υ και ολόκληρο το σύνολο των υπό συνθήκη ανεξαρτησίας συνθηκών μπορεί να επιτευχθεί με έναν αλγόριθμο που βασίζεται στο αντίστοιχο γράφημα.

Ορισμός 4.2 (Μη κατευθυνόμενα γραφήματα)

Δοθέντος ενός μη κατευθυνόμενου γραφήματος G(V, E), έστω $\{X_v : v \in V\}$ το σύνολο των τυχαίων μεταβλητών που ορίζονται από τις κορυφές του γραφήματος και C το σύνολο των κλικών του γραφήματος. Σε σχέση με κάθε κλίκα $c \in C$, ορίζεται ως $\psi_c(x_c)$ η μη αρνητική δυνητική συνάρτηση.



Γράφημα 4.4

Μη κατευθυνόμενο γράφημα συνάφειας για τις μεταβλητές X_1, X_2, X_3, X_4, X_5, X_6.

Ορίζουμε την από κοινού πιθανότητα $p(x_V)$ με το να θεωρήσουμε το γινόμενο από αυτές τις συναρτήσεις και να κανονικοποιήσουμε:

$$p(x_V) = \frac{1}{Z} \prod_{c \in C} y_c(x_c) \qquad (4.2),$$

όπου Z είναι ο παράγοντας κανονικοποιήσης που προκύπτει αν αθροίσουμε ή ολοκληρώσουμε το γινόμενο ως προς x_V.

Γραφήματα συνάφειας όπως τα γραφήματα 4.3 και 4.4, συνήθως χρησιμοποιούνται στην διαστημική στατιστική, στην επεξεργασία λόγου, στα δίκτυα επικοινωνίας, αλλά αν συμπεριληφθούν και παράμετροι στις κορυφές του γραφήματος, τα μη κατευθυνόμενα γραφήματα αποτελούν εργαλεία για τα μπεϋζιανά μοντέλα. Γενικά, τα κατευθυνόμενα και μη κατευθυνόμενα γραφήματα κάνουν διαφορετικές υποθέσεις για την υπό συνθήκη ανεξαρτησία. Συνεπώς, υπάρχουν οικογένειες κατανομών πιθανότητας που μπορούν να περιγραφούν μόνο από ένα κατευθυνόμενο γράφημα και όχι από ένα μη κατευθυνόμενο και αντίστροφα.

4.2.3 Συμπέρασμα για τα γραφικά μοντέλα

Με το να χρησιμοποιηθεί η θεωρία των γραφημάτων στην αναπαράσταση μοντέλων, η τυποποίησή τους παρέχει γενικούς αλγορίθμους για υπολογισμό των περιθώριων και υπό συνθήκη πιθανοτήτων που μπορεί να ενδιαφέρουν τον ερευνητή. Επίσης, η τυποποίηση

παρέχει έλεγχο για την πολυπλοκότητα των υπολογισμών που σχετίζεται με τέτοιες διαδικασίες.

Ωστόσο, η τυποποίηση των γραφικών μοντέλων αγνοεί τον διαχωρισμό μεταξύ της κλασικής και μπεϋζιανής στατιστικής. Επίσης, παρέχει εργαλεία για το χειρισμό κατανομών από κοινού πιθανοτήτων και ειδικότερα καθιστά εύκολη την αναπαράσταση ιεραρχικών μοντέλων λανθανουσών (*latent*) μεταβλητών, τα οποία θα μελετήσουμε στο κεφάλαιο 5. Με το να αντιμετωπίζεται η μπεϋζιανή στατιστική ως συστηματική εφαρμογή της θεωρίας πιθανοτήτων στη στατιστική και τα γραφικά μοντέλα ως συστηματική εφαρμογή των αλγορίθμων που στηρίζονται στη θεωρία γραφημάτων και στη θεωρία πιθανοτήτων, δεν αποτελεί έκπληξη ότι πολλοί συγγραφείς θεωρούν τα γραφικά μοντέλα ως μια γενική μπεϋζιανή μηχανή αναφοράς. Το πιο χαρακτηριστικό ίσως στην προσέγγιση των γραφικών μοντέλων είναι η φυσικότητά τους να τυποποιούν πιθανοτικά μοντέλα περίπλοκων φαινομένων σε εφαρμοσμένα πεδία, όπου διατηρούν τον έλεγχο για το υπολογιστικό κόστος που σχετίζεται με τα μοντέλα αυτά.

ΚΕΦΑΛΑΙΟ 5^{Ο}

Μοντέλα Συνάφειας Λανθανουσών Μεταβλητών

5.0 Εισαγωγή

Τα λογαριθμο-πολλαπλασιαστικά μοντέλα συνάφειας μπορούν να χρησιμοποιηθούν ως (γραφικά) μοντέλα λανθανουσών μεταβλητών (*latent variables models*) για διακριτές μεταβλητές. Τα μοντέλα αυτά βασίζονται σε γραφικά μοντέλα για διακριτές και συνεχείς μεταβλητές, οι οποίες ακολουθούν μια υπό συνθήκη γκαουσσιανή κατανομή και έχουν αρκετά πλεονεκτήματα, όπως σχηματικές ή γραφικές απεικονίσεις της σχέσης των παρατηρούμενων και μη παρατηρούμενων μεταβλητών. Αντίστροφα, τα λογαριθμο-πολλαπλασιαστικά μοντέλα μπορούν να προκύψουν από τα γραφήματα. Επίσης, οι εκτιμήσεις των μέσων, των διακυμάνσεων και των συνδιακυμάνσεων των λανθανουσών μεταβλητών, δοθέντος των τιμών των παρατηρούμενων μεταβλητών, είναι συνάρτηση των παραμέτρων των λογαριθμο-πολλαπλασιαστικών μοντέλων. Τα μοντέλα λανθάνουσας μεταβλητής προτάθηκαν αρχικά από τους Lauritzen and Vermunt(1989), Vermunt and Lauritzen(1990).

5.1 Γραφικά μοντέλα λανθανουσών μεταβλητών

Γενικά, τα μοντέλα θα πρέπει να διευκολύνουν τον ερευνητή να μελετά τις υποκείμενες δομικές σχέσεις μεταξύ των μη παρατηρούμενων μεταβλητών. Ιδανικά, οι ερευνητές θα πρέπει να είναι ικανοί να μετασχηματίζουν τις υποθέσεις τους σχετικά με τις σχέσεις μεταξύ των παρατηρούμενων και μη παρατηρούμενων μεταβλητών στα στατιστικά μοντέλα, τα οποία αντιστρόφως να μπορούν εύκολα να προσαρμοστούν στα παρατηρούμενα δεδομένα.

Στα λογαριθμο-πολλαπλασιαστικά μοντέλα, τα οποία αποτελούν επεκτάσεις των λογαριθμογραμμικών μοντέλων, οι εξαρτήσεις μεταξύ των διακριτών μεταβλητών αναπαρίστανται με πολλαπλασιαστικούς όρους. Τα μοντέλα αυτά συμβάλλουν με αποτελεσματικό τρόπο όσον αφορά στη λήψη απόφασης σχετικά με το είδος των συναφειών των κατηγορικών δεδομένων. Ωστόσο, είναι λιγότερο χρήσιμα για την περιγραφή της φύσης των παρατηρούμενων συναφειών για περισσότερες από μία μεταβλητές. Όταν οι συνάφειες προκύπτουν από τη σύμπτυξη μη παρατηρούμενων ή όχι απευθείας μετρούμενων συνεχών μεταβλητών, η περιγραφή και η ερμηνεία των συναφειών μπορεί να διευκολυνθεί πολύ, αν στα

μοντέλα αυτά οι παρατηρούμενες συνάφειες αναπαρασταθούν με όρους μη παρατηρούμενων ή αλλιώς λανθανουσών μεταβλητών.

Ειδικές περιπτώσεις των μοντέλων αυτών είναι τα γνωστά μας μοντέλα για κατηγορικά δεδομένα, όπως τα μοντέλα γραμμικής αλληλεπίδρασης, τα ordinal by nominal μοντέλα συσχέτισης, τα μοντέλα ομοιόμορφης συσχέτισης για διατάξιμες κατηγορικές μεταβλητές, τα μοντέλα συσχέτισης RC(M) για δύο μεταβλητές και οι γενικεύσεις τους για τρεις ή περισσότερες μεταβλητές.

Μια απλή περίπτωση των λογαριθμογραμμικών μοντέλων λανθανουσών μεταβλητών συζητήθηκε από τους Lauritzen και Vermunt(1989, 1990), οι οποίοι παρείχαν ερμηνεία του μοντέλου συνάφειας RC του Goodman(1979) μιας λανθάνουσας συνεχούς μεταβλητής για δύο παρατηρούμενες μεταβλητές. Ο Whittaker (1989) ασχολήθηκε με την περίπτωση των πολλαπλών ασυσχέτιστων λανθανουσών μεταβλητών για δύο ή τρεις παρατηρούμενες μεταβλητές. Τέλος, οι Anderson&Vermunt(2000) και η Anderson(2002), ασχολήθηκαν με λογαριθμοπολλαπλασιαστικά μοντέλα συνάφειας ως μοντέλα λανθανουσών μεταβλητών για ποιοτικές ή/και κατηγορικές μεταβλητές.

5.2 Μοντέλα με μια λανθάνουσα μεταβλητή ανά δείκτη

Στα λογαριθμογραμμικά μοντέλα λανθάνουσας μεταβλητής, κάθε παρατηρούμενη μεταβλητή είναι ένας δείκτης (συνδέεται άμεσα) μιας και μόνο λανθάνουσας συνεχούς μεταβλητής. Στα μοντέλα αυτά, όπως προαναφέρθηκε στην εισαγωγή, η από κοινού κατανομή των διακριτών και συνεχών μεταβλητών θεωρείται ότι είναι υπό συνθήκη γκαουσσιανή κατανομή, όπου η περιθωριακή κατανομή των διακριτών μεταβλητών είναι πολυωνυμική και η υπό συνθήκη κατανομή των συνεχών μεταβλητών δοθείσης της κατανομής των διακριτών μεταβλητών είναι πολυμεταβλητή κανονική κατανομή, όπου ο μέσος και ο πίνακας συνδιακυμάνσεων μπορεί να διαφέρει από επίπεδο σε επίπεδο της διακριτής μεταβλητής. Προς το παρόν, υποθέτουμε ότι ο πίνακας συνδιακυμάνσεων ταυτίζεται σε κάθε επίπεδο της διακριτής μεταβλητής.

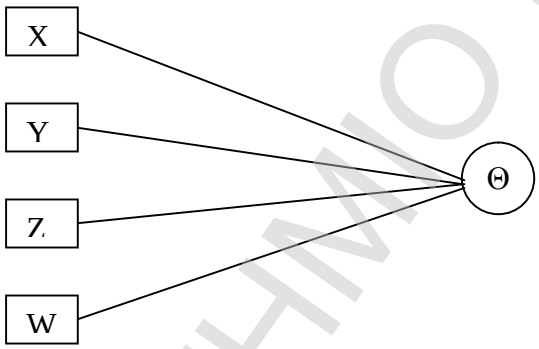
Ας υποθέσουμε ότι έχουμε τέσσερις διακριτές μεταβλητές X, Y, Z, W με αντίστοιχες I, J, K, L κατηγορίες. Επίσης, με $n_{ijkl}, i = 1,...,I,\ j = 1,...,J, k = 1,...,K,\ l = 1,...,L$ συμβολίζουμε την παρατηρούμενη συχνότητα κάθε κελιού και με m_{ijkl} συμβολίζουμε την αναμενόμενη συχνότητα κάθε κελιού κάτω από την υπόθεση ενός μοντέλου.

Έστω το μοντέλο:

$$\log m_{ijkl} = l + l_i^X + l_j^Y + l_k^Z + l_l^W + s^2 m_i n_j + s^2 m_i h_k + s^2 m_i z_l + s^2 n_j h_k + s^2 n_j z_l + s^2 h_k z_l \quad (5.1).$$

Η εξίσωση (5.1) περιλαμβάνει πολλαπλασιαστικούς όρους με την ίδια παράμετρο συνάφειας σε κάθε όρο (σ^2) και ένα σύνολο από σκορ: m_i, n_j, h_k, z_l αντίστοιχα για κάθε μεταβλητή, όπου εμφανίζεται στους διαφορετικούς πολλαπλασιαστικούς όρους. Πρόκειται για λογαριθμοπολλαπλασιαστικά μοντέλα συνάφειας με διμεταβλητές αλληλεπιδράσεις μεταξύ όλων των ζευγών των διακριτών μεταβλητών. Η καλύτερη προσαρμογή που μπορεί να επιτευχθεί με τη χρήση της εξίσωσης (5.1) δίνεται από όλους τους όρους διπλής αλληλεπίδρασης των λογαριθμογραμμικών μοντέλων. Αν όλα τα μοντέλα αυτά προσαρμόζονται σε ένα σύνολο δεδομένων, τότε έχουμε λόγο να χρησιμοποιήσουμε και το μοντέλο (5.1).

Το γράφημα που αντιστοιχεί στο μοντέλο (5.1) είναι το γράφημα 5.1, όπου οι ακμές δηλώνουν την υπό συνθήκη ανεξαρτησία μεταξύ των μεταβλητών.



Γράφημα 5.1

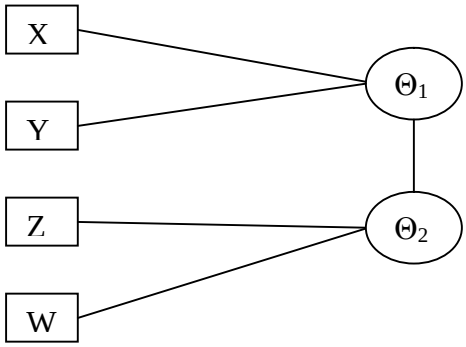
Μοντέλο ενός δείκτη μεταξύ μιας λανθάνουσας μεταβλητής, έστω Θ,
και τεσσάρων παρατηρούμενων μεταβλητών X, Y, Z, W.

Το μοντέλο μιας λανθάνουσας μεταβλητής είναι αρκετά απλό μοντέλο. Σε πολλά σύνολα δεδομένων, οι παρατηρούμενες μεταβλητές μπορεί να είναι δείκτες διαφορετικών λανθανουσών μεταβλητών. Στη συνέχεια, ακολουθεί η γενίκευση του μοντέλου (5.1):

$$\log m_{ijkl} = l + l_i^X + l_j^Y + l_k^Z + l_l^W + s_{11} m_i n_j + s_{12} m_i h_k + s_{12} m_i z_l + s_{12} n_j h_k + s_{12} n_j z_l + s_{22} h_k z_l \quad (5.2)$$

Το μοντέλο (5.2) περιέχει τους πολλαπλασιαστικούς όρους για όλες τις συνάφειες μεταξύ δυο μεταβλητών και υπάρχει ένα μονό σύνολο για τα σκορ της κάθε μεταβλητής. Στο μοντέλο αυτό υπάρχουν τρεις διαφορετικοί όροι συνάφειας σ_{11}, σ_{22}, σ_{12}. Όταν οι διακριτές μεταβλητές μέσα σε ένα σύνολο σχετίζονται διότι όλες είναι δείκτες της ίδιας λανθάνουσας μεταβλητής, τότε η παράμετρος συνάφειας είναι η διακύμανση. Όταν οι διακριτές μεταβλητές από δυο διαφορετικά σύνολα σχετίζονται επειδή οι αντίστοιχες λανθάνουσες μεταβλητές

σχετίζονται, η παράμετρος συνάφειας είναι η συνδιακύμανση μεταξύ των μεταβλητών αυτών. Το γράφημα (5.2) αντιστοιχεί στο μοντέλο (5.2).



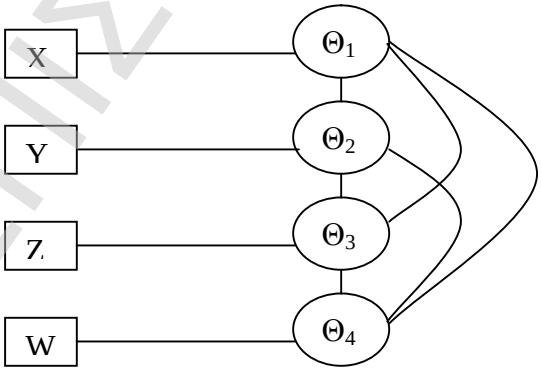
Γράφημα 5.2

Μοντέλο ενός δείκτη μεταξύ δυο συσχετισμένων λανθανουσών μεταβλητών, έστω Θ_1, Θ_2 και τεσσάρων παρατηρούμενων μεταβλητών X, Y, Z, W.

Το πιο περίπλοκο μοντέλο ενός παράγοντα ανά δείκτη είναι το μοντέλο όπου κάθε διακριτή μεταβλητή είναι δείκτης διαφορετικών λανθανουσών μεταβλητών και όλες οι λανθάνουσες μεταβλητές είναι συσχετισμένες μεταξύ τους. Το μοντέλο αυτό για τέσσερις μεταβλητές δίνεται από τον τύπο (5.3).

$$\log m_{ijkl} = l + l_i^X + l_j^Y + l_k^Z + l_l^W + s_{12}m_i n_j + s_{13}m_i h_k + s_{14}m_i z_l + s_{23}n_j h_k + s_{24}n_j z_l + s_{34}h_k z_l \quad (5.3)$$

Το αντίστοιχο γράφημα για το μοντέλο (5.3) είναι το γράφημα 5.3.



Γράφημα 5.3

Το πιο σύνθετο μοντέλο ενός δείκτη

(μεταξύ τεσσάρων λανθάνουσων μεταβλητών, έστω Θ_1, Θ_2, Θ_3, Θ_4 και τεσσάρων παρατηρούμενων μεταβλητών X, Y, Z, W).

Τα παραπάνω μοντέλα, αν τα σκορ είναι γνωστά, είναι λογαριθμογραμμικά μοντέλα με linear-by-linear όρους αλληλεπίδρασης για κάθε ζεύγος των παρατηρούμενων μεταβλητών. Αν τα σκορ είναι γνωστά για κάποιες μεταβλητές ενώ όχι για κάποιες άλλες, τότε τα μοντέλα περιέχουν κάποιους όρους αλληλεπίδρασης ordinal-by-nominal.

Αν δεν υπάρχει μερική συνάφεια μεταξύ ενός ζεύγους διακριτών μεταβλητών, τότε η αντίστοιχη διακύμανση είναι ίση με μηδέν και άρα ο όρος αλληλεπίδρασης του ζεύγους των μεταβλητών ισούται με μηδέν. Αν υπάρχει μερική συνάφεια για ένα ζεύγος διακριτών μεταβλητών, τότε υπάρχει περίπτωση να λάβουμε πιο απλά μοντέλα με το να θέτουμε συγκεκριμένους περιορισμούς στις παραμέτρους του μοντέλου, π.χ. περιορισμούς ισότητας συνδιακύμανσης των πολλαπλασιαστικών όρων.

Όλα τα μοντέλα ενός δείκτη ανά λανθάνουσα μεταβλητή περιλαμβάνουν διμεταβλητές αλληλεπιδράσεις μεταξύ των ζευγών των μεταβλητών και συνεπώς, αν ένα λογαριθμογραμμικό μοντέλο έχει την καλύτερη προσαρμογή για την λανθάνουσα μεταβλητή, τότε ένα λογαριθμοπολλαπλασιαστικό μοντέλο υπάρχει. Τα λογαριθμογραμμικά μοντέλα είναι χρήσιμα στο να υποδείξουν εάν ένα συγκεκριμένο μοντέλο λανθάνουσας μεταβλητής είναι κατάλληλο.

5.3 Περιορισμοί προσδιορισμού-κλίμακας για μοντέλα με μια λανθάνουσα μεταβλητή ανά δείκτη

Για όλα τα λογαριθμοπολλαπλασιαστικά μοντέλα, πρέπει να τεθούν περιορισμοί θέσης ώστε να προσδιοριστούν οι όροι περιθώριας επίδρασης π.χ. οι l, l_i^X, l_j^Y, l_k^Z, l_l^W, αλλά και για τα σκορ. Αυτοί οι περιορισμοί μπορεί να προκύψουν αν τεθεί μια τιμή ίση με το μηδέν ή αν τεθεί το άθροισμα των τιμών ίσο με μηδέν.

Ένας επιπλέον περιορισμός απαιτείται για κάθε λανθάνουσα μεταβλητή. Ενώ η διακύμανση για κάθε λανθάνουσα μεταβλητή μπορεί να τεθεί ίση με μία σταθερά, π.χ. $\sigma_{mm}=1$ για όλα τα m, είναι σημαντικό να τεθούν περιορισμοί για τα σκορ των κατηγοριών για μια διακριτή μεταβλητή όπου σχετίζεται άμεσα με τη λανθάνουσα μεταβλητή (περιορισμοί κλίμακας). Αυτοί οι περιορισμοί μπορεί να μεταφραστούν ως περιορισμοί ισότητας στην ολική «δύναμη» της σχέσης μεταξύ των διακριτών μεταβλητών και των λανθανουσών μεταβλητών.

5.4 Μοντέλα με πολλαπλές λανθάνουσες μεταβλητές ανά δείκτη

Οι παρατηρούμενες μεταβλητές μπορούν να συνδεθούν με περισσότερες από μια λανθάνουσες μεταβλητές. Με το να προστεθεί αυτή η περίπλοκη δομή στα μοντέλα, δεν απαιτείται και η

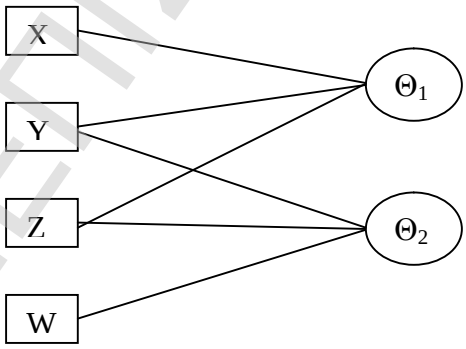
χρήση ενός πιο σύνθετου μοντέλου, αλλά ωστόσο πιο σύνθετης παραμετροποίησης. Στα μοντέλα αυτά, αντίθετα με τα μοντέλα ενός δείκτη, όπου υπάρχει ένα σύνολο από σκορ της διακριτής μεταβλητής, στα μοντέλα με πολλαπλούς δείκτες, μια διακριτή μεταβλητή μπορεί να έχει ένα πολλαπλό σύνολο από σκορ.

Η μεγαλύτερη δυσκολία στη χρήση λογαριθμο-πολλαπλασιαστικών μοντέλων ως μοντέλα πολλαπλών δεικτών έγκειται στο να τεθούν οι απαιτούμενοι περιορισμοί ώστε να καθοριστούν κατά μοναδικό τρόπο οι παράμετροι του λογαριθμο-πολλαπλασιαστικού μοντέλου. Όλοι οι περιορισμοί που αναφέρθηκαν στην παράγραφο 5.3 είναι απαραίτητα και αν τεθεί οποιοσδήποτε άλλος επιπλέον περιορισμός εξαρτάται από την περιπλοκότητα του μοντέλου.

Στη συνέχεια, διακρίνουμε δύο περιπτώσεις τέτοιων μοντέλων, ανάλογα με το αν οι λανθάνουσες μεταβλητές είναι ασυσχέτιστες ή όχι.

5.4.1 Μοντέλα με ασυσχέτιστες λανθάνουσες μεταβλητές

Αν τα μοντέλα που προαναφέραμε στην παράγραφο 5.2 δεν προσαρμόζονται καλά στα δεδομένα μας, τότε υπάρχει μια πιθανότητα για επιπλέον συνάφεια κατά ζεύγη των μεταβλητών και δεν αρκεί μια λανθάνουσα μεταβλητή, αλλά απαιτούνται επιπλέον λανθάνουσες μεταβλητές για κάθε ζεύγος διακριτών μεταβλητών. Για παράδειγμα, ας υποθέσουμε ότι οι μεταβλητές X, Y, Z, W σχετίζονται άμεσα με την λανθάνουσα μεταβλητή A_1 και πρέπει να σχετιστούν και με μία άλλη λανθάνουσα μεταβλητή A_2, η οποία όμως είναι ασυσχέτιστη με την A_1. Τέτοια δομή απεικονίζεται στο γράφημα 5.4.



Γράφημα 5.4

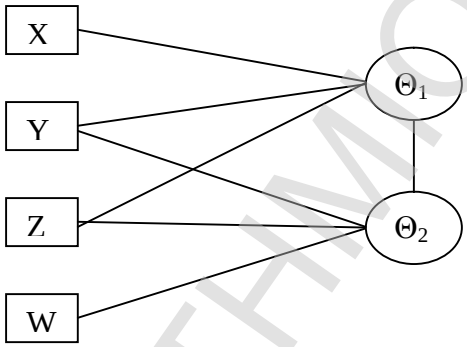
Το μοντέλο πολλαπλών δεικτών μεταξύ δύο ασυσχέτιστων λανθάνουσων μεταβλητών, έστω A_1, A_2,

και τεσσάρων παρατηρούμενων μεταβλητών X, Y, Z, W.

Εκτός από τους περιορισμούς προσδιορισμού που απαιτούνται, για τον προσδιορισμό των σκορ για κάθε διακριτή μεταβλητή που συνδέεται με μια λανθάνουσα μεταβλητή απαιτούνται και περιορισμοί κλίμακας. Επίσης, επιπλέον συνάφεια μπορεί να προκύψει εξαιτίας των πολλαπλών ασυσχέτιστων λανθανουσών μεταβλητών στο μοντέλο και αφού αυτές οι μεταβλητές είναι ασυσχέτιστες, ο πίνακας διασπορών-συνδιασπορών είναι διαγώνιος.

5.4.2 Μοντέλα με συσχετισμένες λανθάνουσες μεταβλητές

Πολλές φορές στις κοινωνικές επιστήμες, οι λανθάνουσες μεταβλητές είναι συσχετισμένες. Συνεπώς, υποθέτουμε τώρα ότι έχουμε την περίπτωση όπου καθεμία από τις I παρατηρούμενες μεταβλητές συνδέονται άμεσα με καθεμία από τις M (M>1) λανθάνουσες μεταβλητές και οι λανθάνουσες μεταβλητές είναι συσχετισμένες. Τέτοια είδους δομή απεικονίζεται στο γράφημα 5.5.



Γράφημα 5.5

Το μοντέλο πολλαπλών δεικτών μεταξύ δύο συσχετισμένων λανθανουσών μεταβλητών, έστω A_1, A_2, και τεσσάρων παρατηρούμενων μεταβλητών X, Y, Z, W.

Όταν οι λανθάνουσες μεταβλητές είναι συσχετισμένες, είναι εξαιρετικά δύσκολο να καθοριστούν περιορισμοί προσδιορισμού των παραμέτρων και τους προσαρμόζουμε εμπειρικά. Συνήθως, ακολουθείται η εξής διαδικασία: δοθέντων του συνόλου των συνθηκών που απαιτούνται για τον προσδιορισμό των παραμέτρων μοντέλου, προσαρμόζουμε το μοντέλο με τις λιγότερες συνθήκες. Στη συνέχεια, θέτουμε τους συνήθεις περιορισμούς που απαιτούνται για τον προσδιορισμό των παραμέτρων του μοντέλου και προσθέτουμε επιπλέον περιορισμούς σε αυτό. Αν κάποιος από τους επιπλέον περιορισμούς χρειάζεται μόνο για προσδιορισμό, τότε τα στατιστικά καλής προσαρμογής του μοντέλου, θα είναι τα ίδια και χωρίς αυτούς τους επιπλέον περιορισμούς.

Αν μια συνθήκη για το μοντέλο γίνει περιορισμός, τότε το μοντέλο δε θα προσαρμόζεται καλά στα δεδομένα μας. Συνεπώς, αφού καθοριστεί ότι μια συνθήκη δεν είναι επιπλέον περιορισμός, τότε το μοντέλο με αυτές τις συνθήκες πρέπει να προσαρμόζεται αρκετές φορές στα δεδομένα μας με τυχαίες αρχικές τιμές. Αν οι συνθήκες ήταν αρκετές για προσδιορισμό, τότε κάθε φορά οι εκτιμήσεις τω παραμέτρων πρέπει να είναι οι ίδιες. Αλλιώς, οι εκτιμήσεις αυτές θα διαφέρουν και θα χρειαστούν επιπλέον συνθήκες στις παραμέτρους ώστε να προσδιοριστούν αυτές κατά μοναδικό τρόπο.

5.5 Ομοιογενή και ετερογενή μοντέλα λανθανουσών μεταβλητών

Στο σημείο αυτό οφείλουμε να σημειώσουμε ότι στα παραπάνω μοντέλα που μελετήσαμε, ο πίνακας διακυμάνσεων-συνδιακυμάνσεων Σ δε διέφερε μεταξύ των κελιών του πίνακα των διακριτών μεταβλητών. Τέτοια μοντέλα ονομάζονται ομοιογενή μοντέλα και ο πίνακας Σ ενσωματώνεται στη σταθερά λ. Στα ετερογενή μοντέλα, ο πίνακας διακυμάνσεων-συνδιακυμάνσεων Σ διαφέρει διαμέσου των κελιών του πίνακα συνάφειας και αφομοιώνεται από άλλους όρους του λογαριθμο-πολλαπλασιαστικού μοντέλου ή διαφορετικά απαιτούνται επιπλέον παράμετροι.

Για μοντέλα λανθανουσών μεταβλητών, μιας ή περισσότερων, ομοιογενή ή ετερογενή, υπάρχει πάντοτε ένα λογαριθμογραμμικό μοντέλο που παρέχει την καλύτερη προσαρμογή για ένα λογαριθμο-πολλαπλασιαστικό μοντέλο.

5.6 Κατασκευή μοντέλων λανθανουσών μεταβλητών από τα γραφήματα συνάφειας

Η εξαγωγή λογαριθμο-πολλαπλασιαστικών μοντέλων από γραφήματα είναι στην ουσία το ίδιο για ομοιογενή ή ετερογενή μοντέλα. Για όλα τα μοντέλα, οι όροι περιθώριας επίδρασης περιλαμβάνονται πάντα για κάθε διακριτή μεταβλητή, όπως και μια σταθερά ώστε να εξασφαλιστεί ότι οι προσαρμοσμένες τιμές αθροίζουν στο παρατηρηθέν σύνολο. Στα γραφήματα συνάφειας, οι ακμές που συνδέουν τις παρατηρηθείσες και παρατηρούμενες μεταβλητές χαρακτηρίζονται από τα αντίστοιχα σκορ. Οι όροι αλληλεπίδρασης στα λογαριθμο-πολλαπλασιαστικά μοντέλα, ισούται με το μισό των γινομένων των ζευγών των σκορ και των συνδιακυμάνσεων μεταξύ των λανθανουσών μεταβλητών από όλα τα κατευθυνόμενα μονοπάτια μεταξύ όλων των παρατηρηθέντων μεταβλητών. Υπάρχουν δύο είδη μονοπατιών (*paths*) στο γράφημα: μονοπάτια όπου από μια διακριτή μεταβλητή συνδέεται πάλι με τον εαυτό της και μονοπάτια όπου οδηγούν από τη μια διακριτή μεταβλητή στην άλλη. Και οι δυο αυτοί τύποι των μονοπατιών μπορεί να περιέχουν περισσότερες από μια

λανθάνουσες μεταβλητές ή ένα ζεύγος λανθανουσών μεταβλητών. Για παράδειγμα, στο γράφημα 5.3, το μονοπάτι X→A_1→Y οδηγεί στον όρο $s_{12}m\eta_j$ που υπάρχει στο μοντέλο (5.3).

5.7 Συμπέρασμα-Πλεονεκτήματα των μοντέλων λανθανουσών μεταβλητών

Όπως είναι ήδη γνωστό, οι μέθοδοι για ανάλυση πολυμεταβλητών κατηγορικών δεδομένων περιλαμβάνουν λογαριθμογραμμικά μοντέλα, πολυμεταβλητά μοντέλα λογιστικής παλινδρόμησης (McGullagh&Nelder, 1989), αλλά και πολλαπλή ανάλυση αντιστοιχιών ή ανάλυση ομοιογένειας που θα αναφερθούμε στη συνέχεια. Αν και τα μοντέλα αυτά είναι πολύ χρήσιμα σε αρκετές περιπτώσεις, δεν προσφέρονται για τη μελέτη δομών, όπως αιτίας και αποτελέσματος, που είναι πολύ διαδεδομένες κυρίως στον τομέα της κλινικής ψυχολογίας. Τότε, το επόμενο βήμα είναι η χρήση γραφημάτων, λογαριθμογραμμικών μοντέλων, μοντέλων λανθανουσών μεταβλητών και λογαριθμο-πολλαπλασιαστικών μοντέλων συνάφειας όλα σε ένα κοινό πλαίσιο.

Τα γραφικά μοντέλα λανθανουσών μεταβλητών είναι λογαριθμο-πολλαπλασιαστικά μοντέλα συνάφειας, γενικεύσεις των μοντέλων συνάφειας RC(***M***) του Goodman από πίνακες διπλής εισόδου σε πίνακες πολλαπλής εισόδου. Τα μοντέλα που προκύπτουν έχουν σχηματικές ή γραφικές απεικονίσεις που δείχνουν τη σχέση μεταξύ όλων των μεταβλητών, και των παρατηρούμενων και των λανθανουσών.

Τα γραφικά μοντέλα λανθανουσών μεταβλητών έχουν τα εξής δύο βασικά πλεονεκτήματα:

1. τα γραφικά μοντέλα παρέχουν φυσικές ερμηνείες σχετικά με τις σχέσεις αιτίας αποτελέσματος, π.χ. ερέθισμα-αντίδραση Από αυτήν την άποψη, τα γραφικά μοντέλα παρέχουν ένα πλαίσιο για διερεύνηση ενός μεγάλου εύρους από υποθέσεις σχετικά με τέτοιου είδους σχέσεις. Είναι σημαντικό να σημειώσουμε ότι οι προκύπτουσες μαθηματικές δομές γενικεύονται εύκολα σε άλλου είδους μελέτες, κυρίως στον τομέα της κλινικής ψυχολογίας.
2. τα γραφικά μοντέλα παρέχουν ένα ενοποιημένο πλαίσιο για τη μελέτη πολυμεταβλητών συναφειών και περιλαμβάνουν τα λογαριθμογραμμικά και τα λογαριθμο-πολλαπλασιαστικά μοντέλα σαν εδικές περιπτώσεις. Συνεπώς, τα γραφικά μοντέλα για διακριτά δεδομένα παρέχουν μια οικονομική (***parsimonious***) προσέγγιση για τη μελέτη των συναφειών σε πολυμεταβλητά δεδομένα.

ΚΕΦΑΛΑΙΟ 6^{ο}

Ανάλυση Ομοιογένειας ή Πολλαπλή Ανάλυση Αντιστοιχιών

6.0 Εισαγωγή

Οι μέθοδοι ανάλυσης δεδομένων που θα μελετήσουμε στην παρούσα εργασία, παρουσιάστηκαν σε διαφορετικές χώρες και για αυτό το λόγο, υπάρχει μια μεγάλη ποικιλία από ονομασίες για τις μεθόδους αυτές που τελικά αποδεικνύεται ότι είναι στην ουσία μια και μόνο μέθοδος. Γενικά, υπάρχουν οι αμερικανικές μέθοδοι: Optimal Scaling, Optimal Scoring, Appropriate Scoring, η καναδική μέθοδος Dual Scaling, η δανέζικη Ανάλυση Ομοιογένειας (*Homogeneity Analysis*), η γαλλική Πολλαπλή Ανάλυση Αντιστοιχιών (*Multiple Correspondence Analysis*), η ισραηλινή Scalogram Analysis και η ιαπωνέζικη Μέθοδος Ποσοτικοποίησης (*Quantification Method*). Οι διαφορετικές ονομασίες για τη μέθοδο αυτή οφείλονται κυρίως στο γεγονός ότι τα δεδομένα που αναλύθηκαν μπορούσαν να εκφραστούν με διαφορετικούς τρόπους, αλλά στην ουσία ισοδύναμους.

Στην παρούσα εργασία, θα υιοθετήσουμε την ονομασία ανάλυση ομοιογένειας (*Homogeneity Analysis-HA*) ή πολλαπλή ανάλυση αντιστοιχιών (*Multiple Correspondence Analysis-MCA*), ανάλογα από ποια οπτική γωνία τη μελετάμε. Η ανάλυση ομοιογένειας μπορεί να αντιμετωπιστεί ως ένας τρόπος ανάλυσης ενός αντικειμένου με ένα μεταβλητό πίνακα με κατηγορικές μεταβλητές ή ως ένα αντικείμενο με έναν πίνακα στοιχείο με πολυμεταβλητά δεδομένα ή ως ένας πίνακας πολλαπλής εισόδου. Σε όλες τις περιπτώσεις, η μέθοδος αυτή κλιμακοποιεί τα αντικείμενα και την κατηγορία (αντίστοιχα τα στοιχεία και τα επίπεδα καθεμιάς εισόδου του πίνακα). Η κλιμάκωση είναι πολυδιάστατη, αφού διαφορετικές τιμές κλίμακας (σκορ) λαμβάνονται για κάθε υποκείμενο και κατηγορία.

6.1 Ιστορική αναδρομή

Η ανάλυση ομοιογένειας είναι πολύ διαδεδομένη στη γαλλική βιβλιογραφία και έχει φθάσει σε ένα πολύ υψηλό επίπεδο ανάπτυξης και χρήσης (Benzecri, 1973, 1977, Lebart, 1975, Lebart, Morineau&Tabard, 1977). Στην αγγλική βιβλιογραφία, η μέθοδος αυτή ήταν λιγότερο δημοφιλής, αλλά έλαβε ολοένα και αυξανόμενη προσοχή πρόσφατα (Nishisato, 1980, Gifi, 1981, Greeanacre, 1984, Lebart, Morineau&Warwick, 1984). Επίσης, ανακαλύφθηκε αρκετές

φορές από ανεξάρτητους μελετητές, οι οποίοι χρησιμοποίησαν κάθε φορά διαφορετικές υποθέσεις, αλλά κατέληξαν τελικά στην ίδια μέθοδο (Richardson&Kuder, 1933, Hirsfeld, 1935, Horst, 1935, Fisher, 1940, Guttman,1941, Burt, 1950, Hayashi, 1950). Στην πραγματικότητα, στην αγγλική βιβλιογραφία, έχουν προταθεί και μελετηθεί τουλάχιστον τέσσερις διαφορετικές μέθοδοι ανάλυσης δεδομένων, οι οποίες όλες οδηγούν στις ίδιες εξισώσεις της ανάλυσης ομοιογένειας. Αξίζει να γίνει μια σύντομη ιστορική αναφορά αυτών των μεθόδων:

- **Η μέθοδος των reciprocal averaging**

Ο Horst(1935) αναφερόμενος σε μια παλαιότερη εργασία των Richardson&Kuder(1933), προτείνει τη μέθοδο των reciprocal averagings. Οι Horst&Richardson χρησιμοποιούν αυτήν την τεχνική στη δεκαετία του 1930 στην εταιρεία Prokter&Gamble. Ο Fisher (1940) πρότεινε την ίδια μέθοδο προφανώς ανεξάρτητα από τον Horst και τον Richardson. Ο Mosier (1946) χρησιμοποίησε αυτή τη μέθοδο σε υπολογιστικές μηχανές της εποχής. Τα πρώτα υπολογιστικά προγράμματα είναι πιθανότατα αυτά που γράφτηκαν από το Beker(1960) και το Lingoes(1964). Αυτή η προσέγγιση περιγράφεται λεπτομερώς (Lebart&Fenelon, 1971, Hill, 1973, Levine, 1979, Mardia, Kent&Bibby, 1979, Nishisato&Sheu, 1980) και είναι η πιο απλή προσέγγιση στην ανάλυση αντιστοιχιών.

- **Η προσέγγιση της ανάλυσης διασποράς**

Αυτή είναι η πιο συνηθισμένη προσέγγιση στην αγγλική βιβλιογραφία. Ο Guttman (1941, 1950, 1953, 1959) έδωσε μια αρκετά ολοκληρωμένη διατύπωση του προβλήματος και το έλυσε. Αυτή η προσέγγιση έχει επίσης μελετηθεί από τους: Mosteller(1949), Hayashi(1950, 1952, 1954), Torgerson(1958), Bock(1960), Lingoes(1963), Shiba(1965), McDonald(1968), Nishisato(1972, 1973, 1976, 1978, 1979, 1980), Saito(1973), de Leeuw(1973), Nishisato&Leong(1975), Heanly&Goldstein(1976) και Van Rijckevorsel&de Leeuw(1978).

- **Η προσέγγιση της ανάλυσης των κυριών συνιστωσών**

Ο Horst(1935) και ο Guttman(1941) διερεύνησε την προσέγγιση των κυρίων συνιστωσών. Η προσέγγιση αυτή διατυπώθηκε από τον Burt (1950, 1953), ενώ στη συνέχεια αναπτύχθηκε εκτενώς από τους: Benzecri (1973, 1977a, 1977b), Lebart(1975), Lebart, Morineau&Tabart(1977), Cazes, Baumerder, Bonnefous&Pages(1977), Greenacre(1984) και Lebart, Moorineau&Warwick(1984). Η προσέγγιση που προτάθηκε από τους Van Rijckevorsel

και de Leeuw διαχειρίζεται ελλιπή δεδομένα (*missing data*), ενώ σε εκείνη των Nishisato&Inukai(1972) πραγματοποιούνται άλλες ενδιαφέρουσες γενικεύσεις.

- **Η προσέγγιση της γενικευμένης κανονικής ανάλυσης.**

Η προσέγγιση της γενικευμένης κανονικής ανάλυσης έχει μελετηθεί από τους McKeon(1966), Masson(1974), Saporta(1975), Bouroche, Saporta&Tenenhaus(1975), Tenenhaus(1977), Saporta(1980) και Leclerc(1980).

Μία πολύ λεπτομερής περιγραφή των δύο πρώτων προσεγγίσεων υπάρχει στο Nishisato(1980). Η ιστορία της ανάλυσης αντιστοιχιών περιγράφεται από τον Benzecri(1977a). Μία σύνθεση των τεσσάρων προσεγγίσεων μπορεί να βρεθεί στον Leclerc(1980).

6.2 Το σύστημα Gifi

Το σύστημα Gifi είναι μια συλλογή από τεχνικές πολυμεταβλητής ανάλυσης για κατηγορικά δεδομένα. Κυρίαρχα θέματα του συστήματος είναι η ιδέα του optimal scaling των κατηγορικών δεδομένων και η υλοποίησή τους με τη βοήθεια αλγορίθμων εναλλασσόμενων ελαχίστων τετραγώνων (*ALS-alternating least squares*). Το σημείο εκκίνησης του συστήματος είναι η ανάλυση ομοιογένειας, μια ειδική μορφή optimal scaling. Η χρήση διαφόρων τύπων περιορισμών επιτρέπει την ανάλυση ομοιογένειας να μπαίνει σε καλούπι και να μετατρέπεται σε άλλους τύπους μη γραμμικών πολυμεταβλητών τεχνικών. Αυτές οι τεχνικές έχουν χρησιμοποιηθεί εκτενώς σε περιπτώσεις ανάλυσης δεδομένων.

Στο σύστημα Gifi, η συνάρτηση ζημίας γενικεύεται και τίθενται διάφοροι περιορισμοί στις ποσοτικοποιήσεις των κατηγοριών (σκορ), έτσι ώστε να ενσωματωθούν σε αυτό και άλλες δημοφιλείς πολυμεταβλητές τεχνικές, π.χ. μη γραμμική ανάλυση κυριών συνιστωσών, ενώ κυριαρχεί η γραφική αναπαράσταση των δεδομένων και η φυσική τους ερμηνεία. Ωστόσο, ένα βασικό σημείο του συστήματος αυτού είναι ότι οι μέθοδοι δεν παρουσιάζονται με όρους εκτίμησης παραμέτρων, βασιζόμενων σε μοντέλα, αλλά τίθεται ζήτημα βελτιστοποίησης μια συνάρτησης ζημίας. Οι μέθοδοι στο σύστημα Gifi πρέπει να θεωρηθούν και να χρησιμοποιηθούν ως περίπλοκες μορφές της επεξηγηματικής, περιγραφικής ανάλυσης δεδομένων.

Στο βιβλίο του A.Gifi, “Nonlinear Multivariate Analysis”, το κεφάλαιο 13 είναι καθολικά αφιερωμένο σε εφαρμογές της ανάλυσης ομοιογένειας, οι οποίες καλύπτουν το πεδίο

της εκπαίδευσης, της κοινωνιολογίας και της ψυχολογίας. Επίσης, οι Greenacre και Benzecri, στα βιβλία τους, παρέχουν μια μεγάλη ποικιλία από εφαρμογές σε πολλαπλή ανάλυση αντιστοιχιών, στα πεδία της γενετικής, της κοινωνικής ψυχολογίας, των κλινικών ερευνών, της εκπαίδευσης, της εγκληματολογίας, της γλωσσολογίας, της οικολογίας, της παλαιοντολογίας και της μετεωρολογίας. Άλλες εφαρμογές αυτών των τεχνικών περιλαμβάνουν το μάρκετινγκ, τη ζωολογία, τις περιβαλλοντικές μελέτες, τη φαρμακευτική και την τεχνολογία τροφίμων. Ωστόσο, το σύστημα Gifi επεκτείνεται και πέραν της ανάλυσης ομοιογένειας και των γενικεύσεών της και νέες τεχνικές έχουν αναπτυχθεί για path μοντέλα, για μοντέλα χρονοσειρών, για γραμμικά δυναμικά συστήματα κ.ο.κ. Τέλος, πρέπει να αναφερθεί ότι το σύστημα Gifi είναι τμήμα ενός ακόμα αρκετά ενεργού ερευνητικού πεδίου.

6.3 Ανάλυση Ομοιογένειας

Στη συνέχεια, θα μελετήσουμε την ανάλυση ομοιογένειας και θα παραθέσουμε το αντίστοιχο παράδειγμα.

6.3.1 Εισαγωγή

Η ανάλυση ομοιογένειας είναι πολύ δημοφιλής τεχνική που στοχεύει κυρίως στην κατασκευή γραφικών απεικόνιση των κατηγορικών πολυμεταβλητών δεδομένων. Επίσης, παρουσιάζεται ως μια τεχνική μελέτης διμερών (*bipartite*) γραφημάτων, που θα αναφερθούμε στη συνέχεια.

Η ανάλυση ομοιογένειας είναι βασική τεχνική στο σύστημα Gifi, της μη γραμμικής πολυμεταβλητής ανάλυσης. Όπως προαναφέρθηκε, στόχος της είναι η αναπαράσταση της δομής των κατηγορικών πολυμεταβλητών δεδομένων και ένα συγκεκριμένο κριτήριο για αυτό το σκοπό είναι η βελτιστοποίηση της κλιμακοποίησης των δεδομένων με το να τίθενται σκορ στα αντικείμενα και κατηγορίες στις μεταβλητές. Αυτά τα σκορ στη συνέχεια χρησιμοποιούνται ώστε να πραγματοποιηθεί μια γεωμετρική αναπαράσταση των συναφειών των δεδομένων σε ένα ευκλείδειο χώρο μικρότερης διάστασης.

Ωστόσο, έχουν γίνει αρκετές προσπάθειες για μετατροπή της ανάλυσης αντιστοιχιών ενός πίνακα συνάφειας πολλαπλής εισόδου σε μια προσέγγιση βασιζόμενη σε μοντελοποίηση και έτσι έχουμε τα μοντέλα συνάφειας, τα μοντέλα συσχέτισης και τις επεκτάσεις τους για κατηγορικές μεταβλητές. Επίσης, αναπτύχθηκαν για αυτό το σκοπό και τα μοντέλα λανθανουσών μεταβλητών για την ανάλυση ενός πολυμεταβλητού πίνακα συνάφειας ή ενός συνόλου τέτοιων πινάκων συνάφειας. Τα μοντέλα αυτά ήδη μελετήθηκαν στα κεφάλαια 3, 4 και 5.

6.3.2 Αρχές ομοιογένειας

Στο σημείο αυτό, είναι απαραίτητο να δοθούν ορισμοί που αφορούν στην έννοια της ομοιογένειας, ώστε να συμβάλλουν στην καλύτερη κατανόηση της μεθόδου. Στο σύνολό τους χαρακτηρίζονται ως αρχές ομοιογένειας και είναι οι εξής:

- Μια κλίμακα που αποτελείται από αριθμητικές μεταβλητές είναι ομοιογενής (*homogeneous)*, αν όλες οι μεταβλητές στην κλίμακα είναι γραμμικά συσχετισμένες.
- Μια κλίμακα αποτελούμενη από αριθμητικές μεταβλητές είναι ομοιογενοποιήσιμη (*homogenizable),* αν όλες οι μεταβλητές στην κλίμακα είναι μετασχηματισμένες με τέτοιο τρόπο ώστε η κλίμακα που προκύπτει να είναι ομοιογενής.
- Μια κλίμακα που αποτελείται από ονοματικές, διατάξιμες ή αριθμητικές μεταβλητές είναι ομοιογενοποιήσιμη (*homogenizable),* αν όλες οι μεταβλητές στην κλίμακα είναι μετασχηματισμένες ή ποσοτικοποιημένες με τέτοιο τρόπο ώστε η κλίμακα που προκύπτει να είναι ομοιογενής.
- Η ομοιογένεια ενός συνόλου από (κεντροποιημένες) μεταβλητές μετράται με το να υπολογιστεί το άθροισμα των τετραγώνων «μέσα» στα αντικείμενα (*within objects*) και το άθροισμα των τετραγώνων «μεταξύ» των αντικειμένων (*between objects)*. Τέλεια ομοιογένεια αντιστοιχεί σε μηδενική διασπορά «μέσα» στα αντικείμενα *(variation within objects)*. Ένα μέτρο ομοιογένειας είναι ο λόγος του αθροίσματος των τετραγώνων «μέσα» στα αντικείμενα προς το συνολικό άθροισμα τετραγώνων.
- Η ανάλυση ομοιογένειας μετασχηματίζει αριθμητικές μεταβλητές (π.χ. αναθέτει αριθμητικές τιμές σε καθεμιά από τις κατηγορίες της μεταβλητής) ή ποσοτικοποιεί διατάξιμες ή ονοματικές μεταβλητές (π.χ. αναθέτει αριθμητικές τιμές σε καθεμιά από τις κατηγορίες της μεταβλητής) με τέτοιο τρόπο ώστε η ομοιογένεια να μεγιστοποιείται.

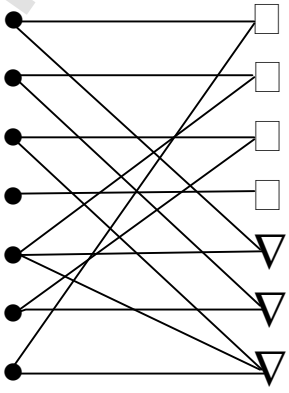
6.3.3 Γεωμετρική εισαγωγή στην Ανάλυση Ομοιογένειας/Διμερές Γράφημα

Η Ανάλυση Ομοιογένειας έχει παρουσιαστεί με πολλούς διαφορετικούς τρόπους και όπως προαναφέρθηκε, ο καλύτερος τρόπος να την αντιμετωπίσουμε είναι από τη σκοπιά της δημιουργίας γραφημάτων, αφού περίπλοκες δομές πολυμεταβλητών δεδομένων μπορούν να κατασταθούν πιο οικείες με το να απεικονιστούν οι κύριες ιδιότητές τους γραφικά.

Ας υποτεθεί ότι έχουν συλλεχθεί δεδομένα από *N* αντικείμενα (άτομα, προϊόντα, περιοχές, κλπ.) σε *J* κατηγορικές μεταβλητές με *kj* κατηγορίες. Βάσει των μεταβλητών αυτών, τοποθετούνται τα αντικείμενα σε ένα πεπερασμένο σύνολο από κατηγορίες *(profiles*). Οι κατηγορίες της κάθε μεταβλητής έχουν ένα συγκεκριμένο επίπεδο μέτρησης (*measurement*

level). Το επίπεδο μέτρησης μπορεί να είναι αριθμητικό (οι μεταβλητές μετρώνται σε μη συμπίπτοντα διαστήματα), διατάξιμο (παίζει ρόλο η διάταξη των κατηγοριών), ή ονοματικό (μόνο οι κλάσεις που σχηματίζονται από τα αντικείμενα παίζουν ρόλο). Επίσης, το ενδιαφέρον επικεντρώνεται στην αναπαράσταση των δεδομένων σε έναν χώρο μικρότερης διάστασης, έστω *p*-διάστατο χώρο(*p*<*J*), πράγμα που σημαίνει ότι επιθυμούμε να κατασκευάσουμε *p* κλίμακες (*scales)*, που θα λάμβαναν υπόψη τους περιορισμούς που τίθενται από το επίπεδο μέτρησης των μεταβλητών.

Όλη η απαραίτητη πληροφορία που περιγράφηκε στην παραπάνω δομή, θα μπορούσε να απεικονιστεί με ένα διμερές γράφημα. Έστω ότι έχουμε ένα πρώτο σύνολο *N* κορυφών που αντιστοιχεί στα αντικείμενα και ένα δεύτερο σύνολο από $\sum_{j\in J} k_j$ κορυφές που αντιστοιχούν στις k_j κατηγορίες των *J* μεταβλητών. Κάθε αντικείμενο συνδέεται με μια ακμή με τις κατηγορίες της μεταβλητής στις οποίες ανήκει. Συνεπώς, το σύνολο των $N\sum_{j\in J} k_j$ ακμών παρέχει πληροφορίες σχετικά σε ποιες κατηγορίες ανήκει ένα αντικείμενο και αντίστροφα, ποια αντικείμενα ανήκουν σε μια συγκεκριμένη κατηγορία. Άρα, οι *N* κορυφές που αντιστοιχούν στα αντικείμενα έχουν βαθμό *J*, ενώ οι $\sum_{j\in J} k_j$ κορυφές που αντιστοιχούν στις κατηγορίες έχουν διάφορους βαθμούς που ισούνται με τον αριθμό των αντικειμένων που ανήκει στις κατηγορίες αυτές. Συνεπώς, η Ανάλυση Ομοιογένειας μπορεί να αντιμετωπιστεί και ως μια τεχνική παραγωγής πληροφοριακών σχημάτων με τη μορφή διμερών γραφημάτων (*bipartite graphs)*. Το γράφημα 6.1 αποτελεί ένα παράδειγμα διμερούς γραφήματος, όπως περιγράφηκε παραπάνω.



Γράφημα 6.1

Διμερές γράφημα.

Ωστόσο, εκτός από πολύ μικρά σε μέγεθος σύνολα δεδομένων, όσον αφορά σε αντικείμενα ή μεταβλητές, τέτοιου είδους αναπαράσταση δεν είναι και ιδιαιτέρως χρήσιμη από πλευράς πληροφόρησης της δομής και των σχέσεων μεταξύ των δεδομένων. Στην ανάλυση ομοιογένειας, θα προσπαθήσουμε να εντοπίσουμε ένα χώρο μικρότερης διάστασης στον οποίο τα αντικείμενα και οι κατηγορίες να τοποθετούνται με τέτοιο τρόπο ώστε όσο το δυνατό περισσότερη πληροφορία να διατηρείται από τα αρχικά δεδομένα και να επιτυγχάνεται η αναπαράστασής τους ως σημεία του ευκλείδειου χώρου ($\Re^p$). Η επιλογή της μικρότερης διάστασης οφείλεται στο γεγονός ότι το γράφημα που θα προκύψει θα μπορεί να σχεδιαστεί εύκολα και η επιλογή του ευκλείδειου χώρου είναι εξαιτίας των σημαντικότατων ιδιοτήτων του (προβολές, τριγωνική ανισότητα, κλπ.), αλλά και από την οικειότητα των μελετητών με την ευκλείδεια γεωμετρία.

Σύμφωνα με τα παραπάνω, ο στόχος μας γίνεται τώρα να κατασκευάσουμε ένα γράφημα, το οποίο να ελαχιστοποιεί το συνολικό τετραγωνικό μήκος των ακμών που ενώνει τα αντικείμενα με τις κατηγορίες των μεταβλητών. Αυτό το κριτήριο επιλέγεται διότι τελικά οδηγεί σε ένα πρόβλημα ιδιοτιμών, το οποίο σχετίζεται με τον καλύτερο τρόπο με διάφορες κλασικές μεθόδους πολυμεταβλητής ανάλυσης.

Τέλος, το πρόβλημα σχεδιασμού γραφημάτων που είναι εύκολο να κατανοηθούν και να παρουσιαστούν, έχει τραβήξει την προσοχή του τομέα των υπολογιστικών επιστημών και έχει προσεγγιστεί από διάφορες απόψεις, στις οποίες ορίζεται ένα συγκεκριμένο σύνολο από κριτήρια και στη συνέχεια υιοθετείται ένας αλγόριθμος, όπου θα αναφερθούμε στη συνέχεια.

6.3.4 Μαθηματική θεμελίωση της Ανάλυσης Ομοιογένειας

Θα προχωρήσουμε στην ακριβή μαθηματική θεμελίωση της μεθόδου της ανάλυσης ομοιογένειας, η οποία περιγράφηκε στην παράγραφο 6.3.1. Στην περαιτέρω ανάλυση, χρησιμοποιούνται πίνακες-δείκτες (*indicator matrices*), ώστε να κωδικοποιηθούν οι J μεταβλητές. Έστω ότι $G_j, j \in J$ συμβολίζει τον $N \times k_j$ πίνακα-δείκτη που αντιστοιχεί στη μεταβλητή j και είναι ένας δυαδικός πίνακας με στοιχεία $g_{it} = 1, i = 1,...,N, t = 1,...,k_j$, αν το αντικείμενο i ανήκει στην κατηγορία t, και $g_{it} = 0$, αν ανήκει σε κάποια άλλη κατηγορία.

Σύμφωνα με τις αρχές ομοιογένειας, επιθυμούμε να ποσοτικοποιήσουμε τις μεταβλητές ώστε να πετύχουμε μέγιστη ομοιογένεια. Έστω ότι Y_j συμβολίζει τον $k_j \times p$ πίνακα που περιλαμβάνει τις ποσοτικοποιήσεις της μεταβλητής $j \in J$ και έστω ο $N \times p$ πίνακας X που περιλαμβάνει τις προκύπτουσες p βέλτιστες τιμές κλίμακας (*optimal scales*). Τα στοιχεία του

πίνακα X είναι γνωστά ως σκορ των αντικειμένων (*object scores)*. Στην πραγματικότητα, δεν είναι δυνατό να βρεθεί μια τέλεια λύση, πράγμα που σημαίνει να καθοριστούν οι πίνακες Y_j και X που παρουσιάζουν τέλεια ομοιογένεια. Συνεπώς, επιθυμούμε να ελαχιστοποιήσουμε τις αποκλίσεις από την τέλεια ομοιογένεια, όπως μετράται από την συνάρτηση ζημίας Gifi:

$$s\left(\mathrm{X};Y_1,...,Y_J\right)=J^{-1}\sum_{j=1}^{J}SSQ(X-G_jY_j) \tag{6.1}$$

όπου $SSQ(H)=tr(H'H)$ δηλώνει την Frobenius νόρμα του πίνακα H (π.χ. το άθροισμα τετραγώνων των στοιχείων του πίνακα H). Με στόχο να αποφευχθεί η τετριμμένη λύση που αντιστοιχεί στο {X=0 και Y_j=0, για κάθε $j\in J$ }, θέτουμε τους εξής περιορισμούς:

$$X'X=NI_p \tag{6.2}$$

$$u'X=0 \tag{6.3}$$

όπου u είναι ένα απλό διάνυσμα με κατάλληλες διαστάσεις. Η λύση στο πρόβλημα της ελαχιστοποίησης βρίσκεται με τη χρήση του αλγορίθμου των εναλλασσόμενων ελαχίστων τετραγώνων (*ALS-Alternating Least Squares*).

Σύμφωνα με τον αλγόριθμο αυτό, στο πρώτο βήμα, η σχέση (6.1) ελαχιστοποιείται ως προς Y_j για καθορισμένο X. Το αποτέλεσμα δίνεται από τη σχέση:

$$\hat{Y}_j=D_j^{-1}G_j'X,\ j\in J \tag{6.4}$$

όπου $D_j=G_j'G_j$ είναι ο $k_j\times k_j$ διαγώνιος πίνακας που περιέχει τα μονομεταβλητά περιθώρια αθροίσματα της μεταβλητής j. Στο δεύτερο βήμα του αλγορίθμου, η σχέση (6.1) ελαχιστοποιείται ως προς X για καθορισμένα Y_j, και το αποτέλεσμα δίνεται από τον τύπο:

$$\hat{X}=J^{-1}\sum_{j=1}^{J}G_jY_j \tag{6.5}$$

Στο τρίτο βήμα του αλγορίθμου, ο πίνακας X είναι στήλη κεντροποιημένη και μετά ορθογωνιοποιημένη, έτσι ώστε οι περιορισμοί κανονικοποίησης (6.2) και (6.3) να ικανοποιούνται. Αυτά τα τρία βήματα επαναλαμβάνονται μέχρι ο αλγόριθμος να συγκλίνει σε ένα ολικό ελάχιστο και τελικά δίνει την επιθυμητή λύση στο πρόβλημα που περιγράφεται με τη βοήθεια του τύπου (6.1). Αυτή η λύση είναι επίσης γνωστή στη βιβλιογραφία ως λύση HOMALS (*HOMogeneity analysis by means of Alternating Least Squares*).

Τέλος, πρέπει να αναφερθεί ότι η σχέση (6.4) εκφράζει την λεγόμενη αρχή του πρώτου κεντροειδούς (*first centroid principle*) και πρόκειται για μια ποσοτικοποίηση κατηγορίας που τοποθετείται στο κεντροειδές των σκορ των αντικειμένων που ανήκουν σε αυτό, ενώ η σχέση (6.5) δηλώνει ότι το σκορ ενός αντικειμένου είναι ο μέσος όρος των ποσοτικοποιήσεων των

κατηγοριών στις οποίες ανήκει. Κατά συνέπεια, αυτή η λύση φέρνει εις πέρας το στόχο της δημιουργίας ενός γραφήματος με τα αντικείμενα κοντά στις κατηγορίες στις οποίες συντάσσονται και οι κατηγορίες κοντά στα αντικείμενα στα οποία ανήκουν. Επίσης, πρέπει να σημειωθεί ότι η χρήση των πινάκων-δεικτών καθιστά την παραπάνω διαδικασία ALS ισοδύναμη με την μέθοδο reciprocal averaging που προαναφέρθηκε στην παράγραφο 6.1.

6.3.5 Βασικές ιδιότητες της λύσης HOMALS

Παραθέτουμε μερικές βασικές ιδιότητες της λύσης HOMALS, όπως διατυπώθηκαν σε δημοσίευση των G.Michailidis και J.de Leeuw (1998):

(1) Οι ποσοτικοποιήσεις των κατηγοριών και τα σκορ των αντικειμένων αναπαρίστανται ως σημεία του ίδιου χώρου.

(2) Ένα σημείο της κατηγορίας είναι το κεντροειδές των αντικειμένων που ανήκουν στην κατηγορία αυτή. Αυτό είναι άμεση συνέπεια της σχέσης (6.4).

(3) Αντικείμενα με τον ίδιο αποκριτικό πρότυπο (πανομοιότυπα προφίλ) αποδέχονται πανομοιότυπα σκορ αντικειμένων (προκύπτει από την σχέση (6.5)). Γενικά, η απόσταση μεταξύ δύο σημείων των αντικειμένων σχετίζεται με την «ομοιότητα» μεταξύ των προφίλ τους.

(4) Μια μεταβλητή διαχωρίζει καλύτερα στην επέκταση του ότι τα σημεία της κατηγορίας της είναι τοποθετημένα μακριά το ένα από το άλλο (προκύπτει από την σχέση (6.7)).

(5) Αν μια κατηγορία αντιστοιχεί σε ένα μόνο αντικείμενο, τότε το σημείο του επιπέδου και το σημείο αυτής της κατηγορίας θα συμπίπτουν.

(6) Σημεία της κατηγορίας με μικρές περιθώριες συχνότητες θα τοποθετούνται μακριά από την προέλευσή τους στον από κοινού χώρο, ενώ κατηγορίες με μεγάλες περιθώριες συχνότητες θα τοποθετούνται κοντά στη θέση από όπου προέρχονται (προκύπτει από την σχέση (6.7)).

(7) Αντικείμενα με «μοναδικό» προφίλ θα τοποθετούνται μακριά από την προέλευσή τους στον από κοινού χώρο, ενώ αντικείμενα με προφίλ παρόμοιο με το «μέσο» προφίλ, θα τοποθετούνται κοντά στην προέλευσή τους (άμεση συνέπεια από την προηγούμενη ιδιότητα).

(8) Οι ποσοτικοποιήσεις της κατηγορίας κάθε μεταβλητής $j \in J$ έχει σταθμισμένο άθροισμα όλων των κατηγοριών που ισούται με μηδέν. Αυτό προκύπτει από την κανονικοποίηση των σκορ των αντικειμένων, αφού $u' D_j Y_j = u' D_j D_j^{-1} G_j' X = u' G_j' X = u' X = 0$.

(9) Η λύση HOMALS είναι εμφωλευμένη (*nested*). Αυτό σημαίνει ότι αν κάποιος απαιτεί μια p_1-διάστατη HOMALS λύση και μετά μια δεύτερη ($p_2>p_1$)-διάστατη λύση, τότε οι πρώτες p_1 διαστάσεις της τελευταίας λύσης ταυτίζονται με την p_1-διάστατη λύση.

(10) Οι λύσεις από τις επόμενες διαστάσεις είναι διατεταγμένες. Αυτό σημαίνει ότι η πρώτη διάσταση έχει την απόλυτα μέγιστη ιδιοτιμή. Η δεύτερη διάσταση έχει την μέγιστη ιδιοτιμή αντικείμενη στον περιορισμό ότι το X(.,2) δεν συσχετίζεται με το X(., 1) κ.ο.κ.

(11) Τα σκορ των αντικειμένων δεν είναι συσχετισμένα με επόμενες διαστάσεις (προκύπτει από την σχέση(6.3)). Ωστόσο, οι ποσοτικοποιήσεις των κατηγοριών δεν χρειάζεται απαραίτητα να είναι μη συσχετισμένες. Στην πραγματικότητα, τα πρότυπα των συσχετίσεων πρέπει να είναι μάλλον μη προβλέψιμα.

(12) Η λύση είναι αμετάβλητη σχετικά με περιστροφές των σκορ των αντικειμένων στον *p*-διάστατο χώρο και στις ποσοτικοποιήσεις των κατηγοριών. Για να το διαπιστώσουμε, ας υποθέσουμε ότι επιλέγουμε μια διαφορετική βάση για τον χώρο στήλη των σκορ των αντικειμένων X, έστω X'=X×R, όπου R είναι ο πίνακας περιστροφής (*rotation matrix*) που ικανοποιεί τη σχέση R'R=RR'=I_p. Τότε, έχουμε από τη σχέση (6.4) ότι $Y_j^{'} = D_j^{-1} G_j^{'} X' = \hat{Y}_j R$, πράγμα που σημαίνει ότι οποιαδήποτε περιστροφή των σκορ των αντικειμένων και των ποσοτικοποιήσεων των κατηγοριών αντιστοιχεί σε λύση του προβλήματος που δίνεται από τη σχέση (6.1).

(13) *Κανονικοποίηση*: ο πίνακας των σκορ των αντικειμένων X κεντροποιείται με το να αφαιρέσουμε από κάθε στοιχείο του τη μέση τιμή της αντίστοιχης στήλης, για παράδειγμα αν θέσουμε W=X-u(u'X/N). Ο πίνακας W είναι ορθοκανονικοποιημένος είτε από την τροποποιημένη Gram-Schmidt διαδικασία ή από τη διαδικασία QR-factorization. Η τροποποιημένη Gram-Schmidt διαδικασία στη λύση HOMALS έχει μέτριες υπολογιστικές απαιτήσεις της, αλλά πρέπει να χρησιμοποιείται μόνο όταν τα διανύσματα που είναι να ορθογωνιοποιηθούν είναι τέλεια ανεξάρτητα, αλλιώς η QR-factorization πρέπει να είναι η προτιμούμενη μέθοδος. Τελικά, θέτουμε $X = \sqrt{N} W$ και είναι εύκολο να διαπιστώσουμε ότι $X'X = NI_p$.

6.4 Σύγκριση της Ανάλυσης Ομοιογένειας με άλλες τεχνικές ανάλυσης δεδομένων

Στη συνέχεια, θα συγκρίνουμε την ανάλυση ομοιογένειας με άλλες τεχνικές ανάλυσης δεδομένων, ώστε να εντοπίσουμε ομοιότητες και διαφορές με αυτήν.

6.4.1 Σύγκριση με την Ανάλυση Αντιστοιχιών

Στην ανάλυση αντιστοιχιών (c*orrespondence analysis*), θεωρείται ένας πίνακας συχνοτήτων δύο κατηγορικών μεταβλητών και γίνεται προσπάθεια ποσοτικοποίησης των στηλών (ή των γραμμών) του πίνακα με τέτοιο τρόπο ώστε το άθροισμα τετραγώνων μεταξύ των στηλών (ή μεταξύ των γραμμών) να μεγιστοποιείται. Στην παραπάνω περιγραφή της ανάλυσης αντιστοιχιών, οι γραμμές και οι στήλες αντιμετωπίζονται συμμετρικά. Ωστόσο, στην ανάλυση ομοιογένειας δεν υπάρχει αυτή η συμμετρία. Οι γραμμές αντιμετωπίζονται ως αντικείμενα και οι στήλες ως μεταβλητές και ο στόχος είναι να αποδοθούν σκορ στα αντικείμενα τέτοια ώστε το άθροισμα τετραγώνων μεταξύ των γραμμών να μεγιστοποιείται. Με στόχο τον χειρισμό περισσότερων μεταβλητών, η ιδέα ενός πίνακα Burt μπορεί να χρησιμοποιηθεί και η ανάλυση ομοιογένειας μπορεί να θεωρηθεί ως ισοδύναμη με την ανάλυση αντιστοιχιών στον πίνακα Burt, πράγμα που θα μελετήσουμε στη συνέχεια.

Στην ανάλυση αντιστοιχιών δίνεται έμφαση στη γεωμετρική απεικόνιση των δεδομένων και αυτό είναι κυρίαρχο χαρακτηριστικό στη γαλλική βιβλιογραφία. Η ανάλυση ενός πίνακα διπλής εισόδου καλείται ανάλυση αντιστοιχιών (*"analyse des correspondences"*) και η ανάλυση ενός συνόλου από πίνακες-δείκτες, που είναι ισοδύναμη με την ανάλυση ομοιογένειας, καλείται πολλαπλή ανάλυση αντιστοιχιών (*"analyse des correspondences multiple"*).

6.4.2 Σύγκριση με τη μέθοδο Multidimensional Scaling

Στη μέθοδο multidimensional scaling (MDS) μελετώνται οι ανομοιότητες (*dissimilarities*) μεταξύ των αντικειμένων, οι οποίες αντιστοιχίζονται σε ευκλείδειες αποστάσεις μεταξύ σημείων σε χώρο μικρότερης διάστασης. Στην κλασική περίπτωση, αυτές οι ανομοιότητες αποδίδουν τη φυσική απόσταση μεταξύ των αντικειμένων.

Στην ανάλυση ομοιογένειας, επίσης μελετώνται οι ανομοιότητες μεταξύ των αντικειμένων που βασίζονται στην διάταξη που αποδίδεται στα δεδομένα (η πληροφορία αυτή περιέχεται στους πίνακες-δείκτες G_j) και από τα δεδομένα συμπεραίνουμε σε ποιες κατηγορίες ανήκουν τα αντικείμενα (η πληροφορία αυτή στηρίζεται στις περιθώριες συχνότητες και περιέχεται στους πίνακες D_j) Οι ανομοιότητες των αντικειμένων που αναπαρίστανται από τις X^2-αποστάσεις βασίζονται στις παραπάνω πληροφορίες και αυτές οι αποστάσεις δίνουν μια τοποθέτηση των αντικειμένων στον $\sum_{j\in J}(k_j-1)$-διάστατο χώρο. Στην ανάλυση ομοιογένειας, αυτή η διάταξη προσεγγίζεται από ένα χώρο μικρότερης διάστασης έτσι ώστε τα αντικείμενα να διαχωρίζονται με τον καλύτερο δυνατό τρόπο. Η προσέγγιση επιτυγχάνεται μέσω των

ευκλείδειων αποστάσεων και έτσι ένας τρόπος να αναλογιστεί κανείς την ανάλυση ομοιογένειας είναι ότι προσεγγίζει τις X^2-αποστάσεις που χαρακτηρίζουν τις ανομοιότητες στα δεδομένα με ευκλείδειες αποστάσεις. Από τη σκοπιά αυτή, η ανάλυση ομοιογένειας μοιάζει πολύ με τη μέθοδο multidimensional scaling.

Τέλος, η κύρια διαφορά των δύο μεθόδων έγκειται στο εξής: η ανάλυση ομοιογένειας στηρίζεται σε συνθήκες κανονικοποίησης και στην μετρική ερμηνεία των δεδομένων, ενώ η μέθοδος multidimensional scaling είναι μια μη μετρική ανάλυση όπου παίζει ρόλο μόνο η διάταξη των ανομοιοτήτων μεταξύ των αντικειμένων.

6.4.3 Σύγκριση με την Ανάλυση Συστάδων

Όπως έχουμε ήδη διαπιστώσει, η Ανάλυση Ομοιογένειας παρέχει μια προσέγγιση των X^2-αποστάσεων των γραμμών του (***superindicator***) πίνακα G με μικρότερης διάστασης ευκλείδειες αποστάσεις. Οι X^2-αποστάσεις είναι ένα μέτρο των ανομοιοτήτων μεταξύ των αντικειμένων, που στηρίζεται στη διάταξη που αποδίδεται στα δεδομένα. Τα δεδομένα υποδεικνύουν ποιες κατηγορίες καθορίζουν τα αντικείμενα και επίσης πόσα αντικείμενα ανήκουν σε κάθε κατηγορία. Αυτά τα δύο «κομμάτια» πληροφορίας περιέχονται στους πίνακες G και D (πίνακας των μονομεταβλητών περιθωριών για όλες τις μεταβλητές). Η ταξινόμηση προκύπτει από την ερμηνεία της τοποθέτησης των σημείων των αντικειμένων στον *p*-διάστατο χώρο. Δηλαδή, αναζητούμε να καθορίσουμε συστάδες των σκορ των αντικειμένων και να τα χαρακτηρίσουμε. Με αυτή την έννοια, η Ανάλυση Ομοιογένειας μοιάζει με μία τεχνική ανάλυσης συστάδων (c***luster analysis***).

6.4.4 Σύγκριση με τη Διαχωριστική Ανάλυση και την Ανάλυση Διασποράς

Η Ανάλυση Ομοιογένειας μπορεί να αντιμετωπιστεί και με τη βοήθεια της διαχωριστικής ανάλυσης (***discriminant analysis***) και ανάλυσης διασποράς (***analysis of variance***). Ας υποθέσουμε ότι ο πίνακας των σκορ των αντικειμένων X είναι γνωστός. Κάθε κατηγορική μεταβλητή $j \in J$ ορίζει ένα διαμερισμό αυτών των σκορς. Αυτό σημαίνει ότι καθίσταται δυνατή η ανάλυση της ολικής διακύμανσης T του πίνακα X σε μια συνιστώσα B «μεταξύ» (***between objects***) των αντικειμένων και μια συνιστώσα W «μέσα» (***within objects***) στα αντικείμενα. Εδώ, ο στόχος είναι η κλιμακοποίηση των αντικειμένων, δηλαδή η εύρεση του βέλτιστου πίνακα X, με τέτοιο τρόπο ώστε η διάσταση του πίνακα να είναι όσο το δυνατό πιο μικρή, ενώ συγχρόνως ο πίνακας T να παραμείνει ως πίνακας-σταθερά, π.χ. ως μοναδιαίος πίνακας.

Συνεπώς, η ανάλυση ομοιογένειας μεγιστοποιεί τη διασπορά μεταξύ των αντικειμένων, διατηρώντας την ολική διασπορά σταθερή. Επίσης, στην ανάλυση ομοιογένειας ο αριθμός των μεταβλητών είναι μεγαλύτερος της μονάδας και ο πίνακας X είναι άγνωστος, ενώ στη διαχωριστική ανάλυση, έχουμε μια κατηγορική μεταβλητή και ο πίνακας X πρέπει να είναι της μορφής UV, όπου U είναι γνωστός πίνακας και ο V είναι ο πίνακας των βαρών και είναι άγνωστος.

6.5 Πολλαπλή Ανάλυση Αντιστοιχιών

Θα μελετήσουμε την ανάλυση ομοιογένειας από διαφορετική σκοπιά και για αυτό το λόγο θα την αποκαλούμε εναλλακτικά πολλαπλή ανάλυση αντιστοιχιών (*Multiple Correspondence Analysis-MCA*).

6.5.1 Ορισμός της Πολλαπλής Ανάλυσης Αντιστοιχιών

Στην περαιτέρω ανάλυση μας, θα χρησιμοποιήσουμε για την περιγραφή της μεθόδου την ονομασία πολλαπλή ανάλυση αντιστοιχιών, αντί ανάλυση ομοιογένειας, αν και όπως υποστηρίζεται (βλ.Gifi, 1981) είναι τελικά η ίδια μέθοδος. Επιλέγουμε αυτήν την ονομασία εξαιτίας της διαφορετικής σκοπιάς που θα αντιμετωπίσουμε τη μέθοδο αυτή τώρα.

Η κλασική μέθοδος εφαρμογής πολλαπλής ανάλυσης αντιστοιχιών είναι η εφαρμογή απλής ανάλυσης αντιστοιχιών στον πίνακα δείκτη, έστω Z. Έστω ότι έχουμε Q κατηγορικές μεταβλητές, η q μεταβλητή έχει J_q κατηγορίες και ο συνολικός αριθμός των κατηγοριών είναι $J=\sum_q J_q$. Αν έχουμε n περιπτώσεις (υποκείμενα), τότε ο πίνακας δείκτης $Z=\{z_{ij}\}$ είναι ένας nxJ πίνακας με στοιχεία 1 ή 0 και οι γραμμές του αντιστοιχούν στα υποκείμενα και οι στήλες στις μεταβλητές-δείκτριες: $z_{ij}=1$, αν το υποκείμενο i ανήκει στην κατηγορία j, αλλιώς $z_{ij}=0$. Κάθε γραμμή του πίνακα Z έχει άθροισμα Q και το ολικό άθροισμα των στοιχείων του πίνακα Z είναι nJ. Στην ανάλυση του πίνακα Z, η ολική διακύμανση (*total variation*) θα μπορούσε να είναι η X^2 στατιστική συνάρτηση υπολογισμένη για τον πίνακα Z, όπως θα ήταν και σε πίνακα συνάφειας, η οποία απλοποιείται στον τύπο $X_Z^2=n(J-Q)$.

Έστω $Z=[Z_1 \ldots Z_Q]$, έτσι ώστε ο $Z_q(n\times J_q)$ να αναφέρεται στην q κατηγορική μεταβλητή με J_q κατηγορίες. Τότε, ο πίνακας $B=Z^TZ$ ονομάζεται πίνακας Burt (1950), και είναι ο υπερπίνακας όλων των πινάκων συνάφειας $Z_q^TZ_q$ μεταξύ των ζευγών των μεταβλητών, συμπεριλαμβανομένων και των διαγώνιων πινάκων $Z_q^TZ_q$ των περιθωριακών συχνοτήτων. Μπορεί να αποδειχθεί (Greenacre, 1984) ότι οι βέλτιστες παράμετροι στηλών στην ανάλυση αντιστοιχιών του πίνακα-δείκτη Z ταυτίζονται με αυτές που προκύπτουν από την ανάλυση του

πίνακα B, όπου είναι παράμετροι είτε των γραμμών είτε των στηλών του, μιας και ο πίνακας B είναι συμμετρικός. Επίσης, οι κύριες αδράνειες του πίνακα B είναι τα τετράγωνα των αντίστοιχων αδρανειών του πίνακα Z. Αυτό σημαίνει ότι μπορεί να ανακτήσει η λύση για τις στήλες, π.χ κατηγορίες των μεταβλητών, στην πολλαπλή ανάλυση αντιστοιχιών του Z με το να αναλυθεί ο πίνακας B. Αυτό αποδεικνύει ότι η πολλαπλή ανάλυση αντιστοιχιών αναλύει μόνο πληροφορία διπλής εισόδου, αν και ο πίνακας Z περιέχει όλη την πληροφορία πολλαπλής εισόδου των δεδομένων (de Leeuw, 1984).

Έχοντας αποδειχθεί η παραπάνω ισοδυναμία, θεωρούμε την πολλαπλή ανάλυση αντιστοιχιών ως μια προσέγγιση σταθμισμένων ελαχίστων τετραγώνων του πίνακα Burt. Μια στατιστική συνάρτηση X^2 μπορεί να υπολογιστεί για τον πίνακα B σαν να πρόκειται για πίνακα συνάφειας και απλοποιείται στον τύπο:

$$X_B^2 = \sum\sum_{q \neq s} X_{qs}^2 + n(J - Q)$$

όπου X_{qs}^2 είναι η X^2 στατιστική συνάρτηση για τον μη διαγώνιο υποπίνακα. $N_{qs}= Z_q{}^T Z_s$ και το $\sum\sum_{q \neq s}$ δηλώνει το διπλό άθροισμα ως προς q που δεν είναι ίσα με s. Σε αυτό το μέτρο ολικής διασποράς, εμφανίζεται να μην υπάρχει αιτιολόγηση για την προσαρμογή των υποπινάκων στην διαγώνιο του πίνακα B που συνεισφέρει τον όρο n(J-Q). Στην πραγματικότητα, η παρουσία αυτού του όρου τεχνητά διογκώνει την ολική διακύμανση σε βαθμό όπου τα ποσοστά για τα οποία υπολογίζεται από τον μεγαλύτερο κύριο άξονα μπορεί να είναι πολύ μικρά, ειδικά όταν η ποσότητα J-Q είναι μεγάλη. Ένα πιο φυσικό μέτρο της ολικής διασποράς είναι το $\sum\sum_{q \neq s} X_{qs}^2$. Αυτό προτείνει μια εναλλακτική γενίκευση της ανάλυσης αντιστοιχιών που προσαρμόζει μόνο τους μη διαγώνιους πίνακες συνάφειας, ανάλογη με την ανάλυση παραγόντων όπου οι τιμές στην διαγώνιο του πίνακα διακύμανσης ή συσχέτισης δεν παρουσιάζουν άμεσο ενδιαφέρον.

6.5.2 Η Ανάλυση Ομοιογένειας (Πολλαπλή Ανάλυση Αντιστοιχιών) ως ένα πρόβλημα Ιδιοτιμών και Singular Value Decomposition

Στην ανάλυση ομοιογένειας, το τετραγωνικό μήκος των ακμών που συνδέουν τα αντικείμενα και τις ποσοτικοποιήσεις των κατηγοριών αποτελεί κριτήριο γιατί καθιστά το πρόβλημα ελαχιστοποίησης ως ένα πρόβλημα ιδιοτιμών.

Αν αντικαταστήσουμε τη βέλτιστη λύση $\hat{Y}_j = D_j^{-1} G_j' X, j \in J$, για δοθέν X στη συνάρτηση ζημίας (6.1) έχουμε:

$$s(\mathrm{X};\bullet) = J^{-1}\sum_{j=1}^{J} tr(X - G_j D_j^{-1} G_j^{'} X)'(X - G_j D_j^{-1} G_j^{'} X) =$$

$$= J^{-1}\sum_{j=1}^{J} tr(X'X - X'G_j D_j^{-1} G_j^{'} X) \qquad (6.6)$$

όπου η τελεία σημαίνει ότι έχει αντικατασταθεί το στοιχείο ως προς το οποίο η συνάρτηση ζημίας έχει ελαχιστοποιηθεί.

Αν θέσουμε $P_j = G_j D_j^{-1} G_j^{'}$, τότε ο P_j εκφράζει την ορθογώνια προβολή στον υπόχωρο που ορίζεται από τις στήλες του πίνακα δείκτη G_j. Έστω $\bar{P} = J^{-1}\sum_{j=1}^{J} P_j$ ο μέσος των J προβολών. Τότε, η εξίσωση (6.6) μπορεί να γραφεί ως εξής:

$$s(\mathrm{X};\bullet) = J^{-1}\sum_{j=1}^{J} tr(X - P_j X)'(X - P_j X) =$$

$$= J^{-1}\sum_{j=1}^{J} tr(X'X - X'P_j X) \qquad (6.7)$$

Η βέλτιστη λύση του X αντιστοιχεί στην μεγαλύτερη από τα p ιδιοδιανύσματα του πίνακα και η ελάχιστη τιμή της συνάρτησης ζημίας είναι:

$$s(\mathrm{X};\bullet) = N(p - \sum_{s=1}^{p} l_s), \qquad (6.8)$$

όπου λ_s, $s=1,\ldots, p$ είναι οι p μεγαλύτερες ιδιοτιμές του πίνακα $\bar{P}$. Συνεπώς, η ελάχιστη τιμή της συνάρτησης ζημίας της ανάλυσης ομοιογένειας είναι μια συνάρτηση των p μεγαλύτερων ιδιοτιμών της μέσης προβολής.

Το πλεονέκτημα της χρήσης του αλγορίθμου ALS για την ελαχιστοποίηση της συνάρτησης ζημίας όπως δίνεται από τον τύπο (6.8) είναι ότι υπολογίζει τις p μεγαλύτερες διαστάσεις της λύσης και έτσι αυξάνεται η υπολογιστική αποδοτικότητα και μειώνονται οι απαιτήσεις σε υπολογιστική μνήμη.

Μια άλλη προοπτική είναι τα σκορ του αντικειμένου X μπορούν να προκύψουν ως αριστερά μοναδιαία διανύσματα του πίνακα $J^{-1/2}(I - uu'/N)GD^{-1/2}$, όπου $G = [G_1|...|G_J]$ είναι ο super-indicator πίνακας, με αποκλίσεις από τους μέσους των στηλών και σταθμισμένος ως προς τις περιθώριες συχνότητες.

Η ολοκληρωμένη Singular Value Decomposition (*SVD*) λύση έχει $l = \sum_{j=1}^{J} k_j - J$ διαστάσεις. Αφού τα σκορ των αντικειμένων προσδιοριστούν, οι ποσοτικοποιήσεις των κατηγοριών υπολογίζονται με χρήση της σχέσης (6.1). Το πλεονέκτημα του να υιοθετηθεί ο αλγόριθμος ALS είναι ότι αυτός φαίνεται μόνο στις πρώτες $p<<l$ διαστάσεις της λύσης SVD, που σημαίνει πως αυξάνει η υπολογιστική αποδοτικότητα του και μειώνονται οι υπολογιστικές απαιτήσεις μνήμης.

6.6 Γεωμετρική ερμηνεία της πολλαπλής ανάλυσης αντιστοιχιών

Η ερμηνεία της πολλαπλής ανάλυσης αντιστοιχιών στηρίζεται στην αναπαράσταση σημείων σε έναν μικρότερης διάστασης χώρο, συνήθως διδιάστατο ή τρισδιάστατο. Οι αποστάσεις μεταξύ των σημείων και η σύγκριση αυτών έχουν νόημα μόνο για σημεία από το ίδιο σύνολο (γραμμές με γραμμές, στήλες με στήλες). Ειδικότερα, αν δυο σημεία γραμμών είναι κοντά το ένα στο άλλο τείνουν να ανήκουν στα ίδια επίπεδα της εκάστοτε μεταβλητής.

Στη γεωμετρική ερμηνεία της απόστασης μεταξύ δυο μεταβλητών πρέπει να διακρίνουμε δυο περιπτώσεις:

i. Η τοποθέτηση κοντά στο χώρο των σημείων που αντιστοιχούν σε επίπεδα διαφορετικών μεταβλητών σημαίνει ότι τα επίπεδα αυτά τείνουν να εμφανίζονται μαζί στις παρατηρήσεις.
ii. Επειδή τα επίπεδα της ίδιας μεταβλητής δεν μπορούν να αντιστοιχούν σε διαφορετικά σημεία, απαιτείται μια διαφορετική ερμηνεία σε αυτήν την περίπτωση. Εδώ, η τοποθέτηση κοντά στο χώρο των σημείων που αντιστοιχούν στα επίπεδα σημαίνει ότι οι ομάδες των παρατηρήσεων που σχετίζονται με αυτά τα επίπεδα είναι μεταξύ τους παρόμοιες.

Επίσης, η κύρια αδράνεια είναι ο σταθμικός μέσος των τετραγώνων (X^2) των αποστάσεων από το κεντροειδές στις προβολές των προφίλ των γραμμών στον κύριο άξονα. Αποτελεί ένα απόλυτο μέτρο της διασποράς των προφίλ των γραμμών στην κατεύθυνση του άξονα αυτού. Η μέγιστη τιμή της είναι 1, όταν όλα τα προβεβλημένα προφίλ γραμμών συμπίπτουν με τα προβεβλημένα σημεία των στηλών. Η σημαντικότητα ενός κύριου άξονα κρίνεται με δυο διαφορετικούς τρόπους. Πρώτον, ο άξονας θεωρείται «επιτυχής», αν τα δεδομένα ανακτούν μια διαταξιμότητα που να έχει νόημα. Δεύτερον, αν τα δεδομένα δεν προέρχονται από πολυωνυμική δειγματοληψία, η μηδενική υπόθεση της τυχαίας διασποράς τους κατά μήκος του πρώτου άξονα μπορεί να ελεγχθεί με τη χρήση ασυμπτωτικών αποτελεσμάτων (βλ.O'Neill, 1978, 1980 και Greenacre, 1984).

6.7 Παράδειγμα

Η πολλαπλή ανάλυση αντιστοιχιών είναι ουσιαστικά η εφαρμογή του αλγορίθμου της απλής ανάλυσης αντιστοιχιών σε πολυμεταβλητά κατηγορικά δεδομένα που έχουν κωδικοποιηθεί με τη μορφή ενός πίνακα δείκτη ή ενός πίνακα Burt. Οι Blasius&Greenacre(1994), αλλά και οι Necadic&Greenacre περίεγραψαν αναλυτικά τέτοιους αλγορίθμους για ανάλυση αντιστοιχιών και πολλαπλή ανάλυση αντιστοιχιών.

Τα δεδομένα για το παράδειγμα αυτό περιλαμβάνονται στο βιβλίο του J.A Hartigan, Clustering Algorithms, Wiley, New York(1975), μελετήθηκαν και στη δημοσίευση των G.Michailides&J.de Leeuw, The Gifi System of Descripitive Multivariate Analysis, Statistical Science, 1998 και παρατίθενται στον Πίνακα 6.1. Αφορούν στην οδοντοφυΐα 66 θηλαστικών, των οποίων τα δόντια τοποθετούνται σε 8 κατηγορίες: άνω και κάτω κοφτήρες, άνω και κάτω κυνόδοντες, άνω και κάτω προγόμφιοι, άνω και κάτω τραπεζίτες. Μια περιγραφή αυτών, καθώς και η κωδικοποίησή τους, δίνεται στη συνέχεια:

- TI: άνω κοφτήρες, όπου (1)0 , (2)1 (3)2, (4)3 ή περισσότερους κοφτήρες.
- BI: κάτω κοφτήρες, όπου (1)0 , (2)1 (3)2, (4)3, (5)4 κοφτήρες.
- TC: άνω κυνόδοντες, όπου (1)0 , (2)1 κυνόδοντες.
- BC: κάτω κυνόδοντες, όπου (1)0 , (2)1 κυνόδοντες.
- TP: άνω προγόμφιοι, όπου (1)0 , (2)1 (3)2, (4)3 , (5)4 προγόμφιοι.
- BP: κάτω προγόμφιοι, όπου (1)0 , (2)1 (3)2, (4)3 , (5)4 προγόμφιοι.
- TM: άνω τραπεζίτες, όπου (1)0, 1 ή 2 , (2)περισσότερους από 2 τραπεζίτες.
- BM: κάτω τραπεζίτες, όπου (1)0, 1 ή 2 , (2)περισσότερους από 2 τραπεζίτες.

Θηλαστικό	**TI**	**BI**	**TC**	**BC**	**TP**	**BP**	**TM**	**BM**
Δίδελφυ(μαρσιποφόρο ζώο Αμερικής)	4	5	2	2	4	4	2	2
Ασπάλακας με φουντωτή ουρά	4	4	2	2	5	5	2	2
Κοινός ασπάλακας	4	3	2	1	4	4	2	2
Ασπάλακας με αστεροειδή μύτη	4	4	2	2	5	5	2	2
Καφέ νυχτερίδα	3	4	2	2	4	4	2	2
Νυχτερίδα με ασημένιο τρίχωμα	3	4	2	2	3	4	2	2
Πυγμαία νυχτερίδα	3	4	2	2	3	3	2	2
Ποντικονυχτερίδα	3	4	2	2	2	3	2	2

Κόκκινη νυχτερίδα	2	4	2	2	3	3	2	2
Υπόλευκη νυχτερίδα	2	4	2	2	3	3	2	2
Νυχτερίδα με στρογγυλή μύτη	3	4	2	2	3	4	2	2
Μικρό φολιδωτό ζώο Νοτίου Αμερικής(armadillo)	1	1	1	1	1	1	2	2
Pika	3	2	1	1	3	3	2	2
Λαγός με άσπρα πόδια	3	2	1	1	4	3	2	2
Κάστορας	2	2	1	1	3	2	2	2
Αγριοποντικός	2	2	1	1	3	2	2	2
Χοίρος	2	2	1	1	3	2	2	2
Είδος υλακτούντος τρωκτικού	2	2	1	1	3	2	2	2
Σκίουρος εδάφους	2	2	1	1	3	2	2	2
Είδος σκιούρου(chipmunk)	2	2	1	1	3	2	2	2
Γκρι σκίουρος	2	2	1	1	2	2	2	2
Αλεποσκίουρος	2	2	1	1	2	2	2	2
Γεωμύς τσέπης (είδος σκιούρου)	2	2	1	1	2	2	2	2
Καγκουροαρουραίος	2	2	1	1	2	2	2	2
Αρουραίος-φορτηγόν	2	2	1	1	1	1	2	2
Ποντίκι πεδιάδας	2	2	1	1	1	1	2	2
Μοσχοαρουραίος	2	2	1	1	1	1	2	2
Μαύρος αρουραίος	2	2	1	1	1	1	2	2
Ποντίκι σπιτιού	2	2	1	1	1	1	2	2
Σκαντζόχοιρος	2	2	1	1	1	2	2	2
Χοίρος Γουινέας	2	2	1	1	1	2	2	2
Κογιότ	2	4	2	2	5	5	5	2
Λύκος	4	4	2	2	5	5	1	2
Αλεπού	4	4	2	2	5	5	1	2
Αρκούδα	4	4	2	2	5	5	1	2
Μοσχογαλής (είδος γάτας)	4	4	2	2	5	5	1	1
Ρακούν	4	4	2	2	5	5	2	1
Κουνάβι	4	4	2	2	5	5	1	1
Ψαράς (fisher)(είδος	4	4	2	2	5	5	1	1

κουναβιού)								
Νυφίτσα	4	4	2	2	4	4	1	1
Είδος νυφίτσας (mink)	4	4	2	2	4	4	1	1
Είδος κουναβιού (ferrer)	4	4	2	2	4	4	1	1
Σαρκοφάγο του Βορρά	4	4	2	2	5	5	1	1
Είδος κουναβιού (badger)	4	4	2	2	4	4	1	1
Είδος κουναβιού Αμερικής (skunk)	4	4	2	2	4	4	1	1
Βύδρα ποταμού	4	4	2	2	5	4	1	1
Βύδρα θάλασσας	4	3	2	2	4	4	1	1
Τζάγκουαρ	4	4	2	2	4	3	1	1
Αιλουροπάρδαλη	4	4	2	2	4	3	1	1
Ορεινό λιοντάρι	4	4	2	2	4	3	1	1
Λυγξ	4	4	2	2	4	3	1	1
Φώκια	4	3	2	2	5	5	1	1
Θαλάσσιο λιοντάρι	4	3	2	2	5	5	1	1
Θαλάσσιος ίππος	2	1	2	2	4	4	1	1
Γκρί φώκια	4	3	2	2	4	4	1	1
Θαλάσσιος ελέφαντας	3	2	2	2	5	5	1	1
Χοίρος Νοτίου Αμερικής	3	4	2	2	4	4	2	2
Μεγάλο ελάφι	1	5	2	1	4	4	2	2
Ελάφι	1	5	1	1	4	4	2	2
Αμερικάνικο ελάφι	1	5	1	1	4	4	2	2
Ρέννος (είδος ελαφιού)	1	5	2	1	4	4	2	2
Αντιλόπη	1	5	1	1	4	4	2	2
Βίσονας	1	5	1	1	4	4	2	2
Κατσίκα (βουνού)	1	5	1	1	4	4	2	2
Μοσχο-βούς	1	5	1	1	4	4	2	2
Πρόβατο (βουνού)	1	5	1	1	4	4	2	2

Πίνακας 6.1

Δεδομένα οδοντοφυΐας θηλαστικών.

Οι στήλες του πίνακα των δεδομένων είναι Q=8 για τις 8 κατηγορίες δοντιών. Καθεμία από τις q κατηγορίες ορίζουν έναν παράγοντα με J_q επίπεδα, π.χ. ο παράγοντας q=1, που αντιστοιχεί στη μεταβλητή ΤΙ έχει J_1=4 επίπεδα. Η κλασική μέθοδος εφαρμογής πολλαπλής ανάλυσης αντιστοιχιών, όπως προαναφέρθηκε είναι η εφαρμογή απλής ανάλυσης αντιστοιχιών στον πίνακα δείκτη, έστω Z. Ο πίνακας δείκτης Z={z_{ij}} αντιστοιχεί σε μια δυαδική κωδικοποίηση των παραγόντων, δηλαδή αντί να χρησιμοποιηθεί ένας παράγοντας με J_q επίπεδα, χρησιμοποιούνται J_q στήλες με δυαδικές τιμές, οι οποίες καλούνται και βουβές μεταβλητές (***dummy variables***). Οι γραμμές του πίνακα-δείκτη είναι 66, όσο και το πλήθος των υποκειμένων, εδώ τα θηλαστικά. Στη συνέχεια, θα εφαρμόσουμε απλή ανάλυση αντιστοιχιών στηριζόμενη στην Singular Value Decomposition (*SVD*).

Ο αριθμός των μη μηδενικών ιδιοτιμών του πίνακα-δείκτη, που βασίζεται στις Q=8 μεταβλητές με συνολικά J=27 επίπεδα (J=$\sum_q J_q$), είναι J-Q=27-8=19. Ο Πίνακας 6.2 περιλαμβάνει μερικές από τις 19 κύριες αδράνειες (*principal inertias*), οι οποίες είναι τα τετράγωνα των ιδιοτιμών, και το αντίστοιχο ποσοστό (%) της αδράνειας που εξηγούν.

k	**1**	**2**	**3**	**4**	**…**	**17**	**18**	**19**
λ_k	0.730	0.380	0.275	0.218	…	0.004	0.002	0
Ποσοστό(%) της αδράνειας που εξηγείται	30.80	16.00	11.60	9.20	…	0.20	0.10	0

Πίνακας 6.2

Μερικές από τις 19 κύριες αδράνειες (λ_k) και το αντίστοιχο ποσοστό(%) της αδράνειας που εξηγούν.

Μερικές τυποποιημένες και κύριες συντεταγμένες των στηλών (b_{jk}, g_{jk} αντίστοιχα) για τις δύο πρώτες διαστάσεις, k=1, 2, περιλαμβάνονται στον Πίνακα 6.3, ενώ οι κύριες συντεταγμένες των γραμμών (f_{i1}, f_{i2} αντίστοιχα) για τις δύο πρώτες διαστάσεις, k=1, 2, περιλαμβάνονται στον Πίνακα 6.4.

	TI1	**TI2**	**TI3**	**TI4**	**BI1**	**BI2**	**BI3**	**…**
b_{j1}	0.692	1.203	-0.073	-1.212	0.389	1.348	-1.192	
b_{j2}	3.250	-1.170	-0.176	-0.244	0.966	-1.207	0.143	
g_{j1}	0.591	1.028	-0.063	-1.036	0.332	1.152	-1.019	
g_{j2}	2.002	-0.721	-0.108	-0.150	0.595	-0.744	0.088	
	BC1	**BC2**	**TP1**	**TP2**	**TP3**	**TP4**	**TP5**	**…**
b_{j1}	1.153	-0.961	1.573	1.326	0.771	-0.354	-1.252	
b_{j2}	0.371	-0.309	-0.620	-1.408	-0.943	1.593	-1.102	
g_{j1}	0.986	-0.821	1.344	1.133	0.659	-0.302	-1.070	
g_{j2}	0.228	-0.190	-0.382	-0.868	-0.581	0.982	-0.679	
	BP5	**TM1**	**TM2**	**BM1**	**BM2**			
b_{j1}	-1.243	-1.284	0.687	-1.289	0.602			
b_{j2}	-1.159	-0.325	0.174	- 0.247	0.115			
g_{j1}	-1.062	-1.097	0.587	-1.102	0.514			
g_{j2}	-0.714	-0.200	0.107	-0.152	0.071			

Πίνακας 6.3

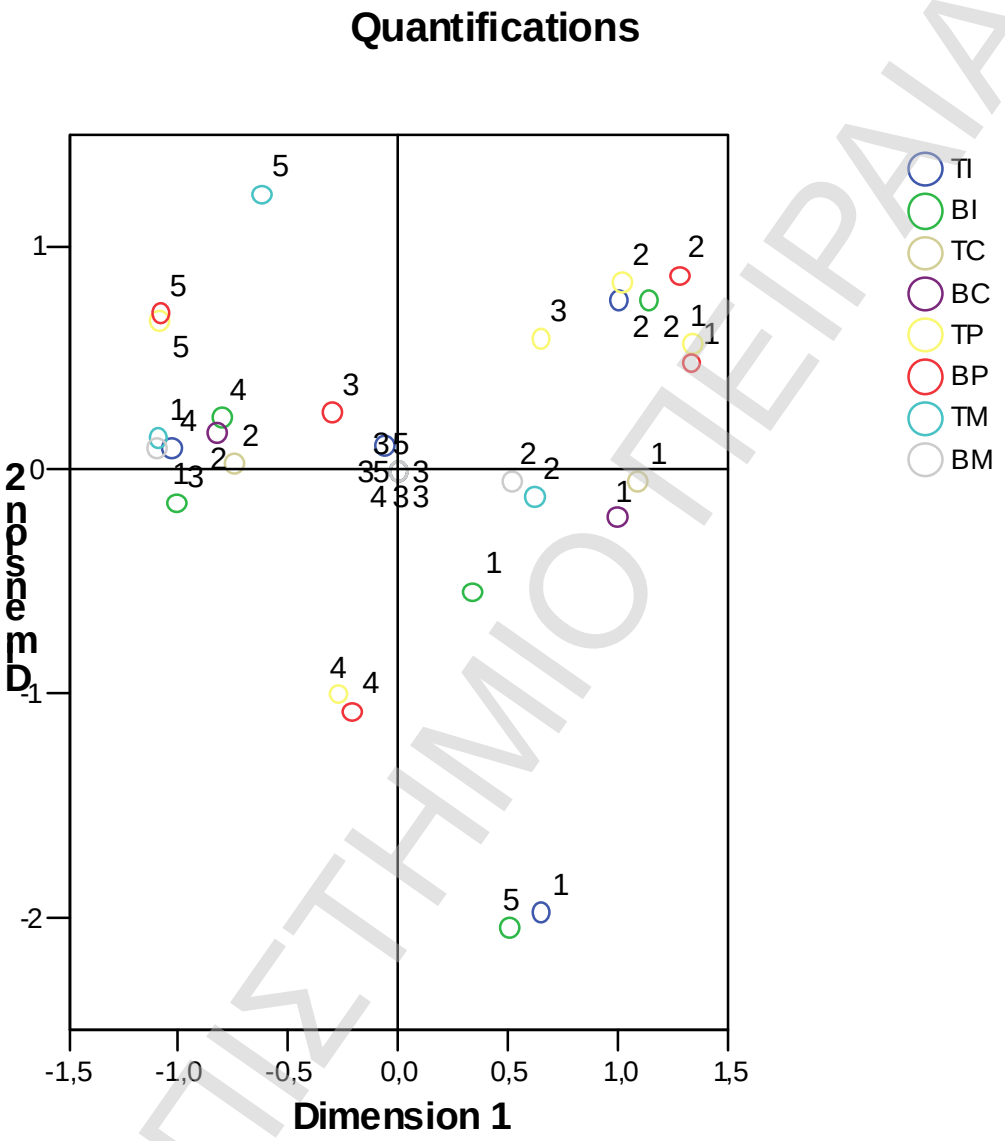
Τυποποιημένες και κύριες συντεταγμένες των στηλών (b_{jk}, g_{jk} αντίστοιχα) για τις δύο πρώτες διαστάσεις.

i	**1**	**2**	**3**	**4**	**5**	**…**	**66**
f_{i1}	-0.230	-0.644	-0.180	-0.644	-0.269	**…**	0.540
f_{i2}	0.803	-0.365	0.488	-0.365	0.342	**…**	1.319

Πίνακας 6.4

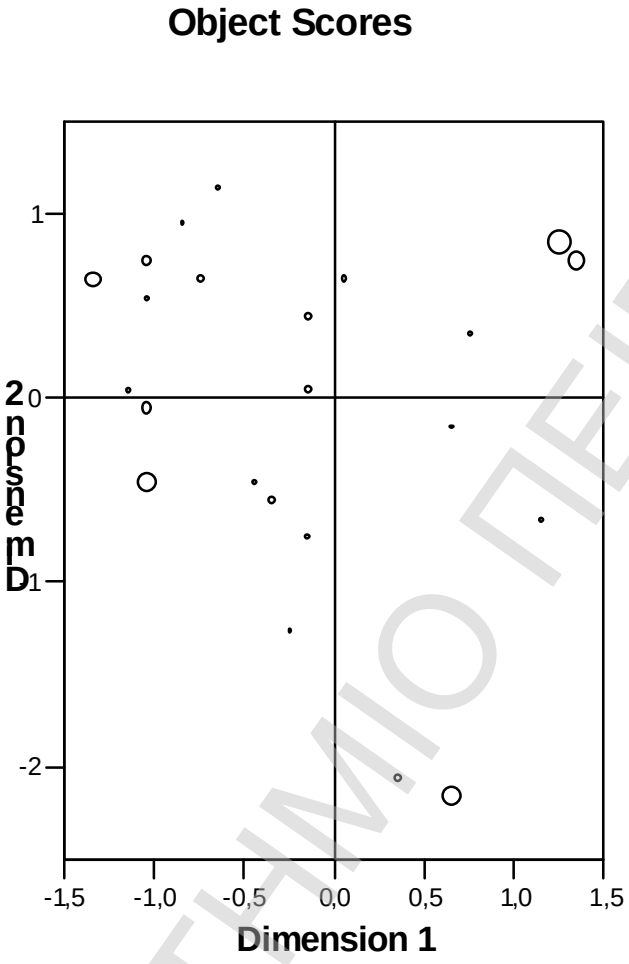
Τυποποιημένες και κύριες συντεταγμένες των γραμμών (f_{i1}, f_{i2} αντίστοιχα) για τις δύο πρώτες διαστάσεις.

Τα γραφήματα 6.1 και 6.2 αποτελούν τη γραφική απεικόνιση των σκορ των στηλών και των σκορ των γραμμών αντίστοιχα για τις δυο πρώτες διαστάσεις:



Γράφημα 6.1

Γραφική αναπαράσταση των σκορ των στηλών για τις δυο πρώτες διαστάσεις.



Γράφημα 6.2

Γραφική αναπαράσταση των σκορ των γραμμών για τις δυο πρώτες διαστάσεις.

6.7.1 Σχολιασμός των γραφημάτων

Τα χαρακτηριστικά των δοντιών χρησιμοποιούνται για την ταξινόμηση των θηλαστικών. Ένας λόγος που συμβαίνει αυτό είναι ότι τα δόντια των θηλαστικών αποτελούν συνηθισμένα λείψανα, μαζί με τα κόκαλα των σαγονιών τους και τα κρανία τους, αφού είναι αρκετά ανθεκτικά στα χημικά και φυσικά φαινόμενα. Εξαιτίας της αφθονίας των λειψάνων των δοντιών στα απολιθώματα των θηλαστικών, έχει δοθεί έμφαση στη μελέτη της οδοντοφυΐα

τους όσον αφορά στην ερμηνεία της φυλογένειας και των σχέσεων των θηλαστικών. Επίσης, τα χαρακτηριστικά των δοντιών παρέχουν πληροφορία για τις διατροφικές συνήθειες των θηλαστικών.

Στην πολλαπλή ανάλυση αντιστοιχιών, το κύριο ζητούμενο είναι να αποδοθεί μια ξεκάθαρη εικόνα για τα δεδομένα και των σχέσεων μεταξύ αυτών και να αποκαλυφθούν μερικά ενδιαφέροντα χαρακτηριστικά του συνόλου των δεδομένων αυτών.

Από το γράφημα 6.1 μπορούμε να διαπιστώσουμε ότι σχηματίζονται τέσσερις ομάδες. Στην πάνω αριστερή γωνία είναι συγκεντρωμένες οι κατηγορίες BM1, TM1, TC2, BC2, TI4, BI3, BI4, TP5 και BP5. Συνεπώς, τα αντικείμενα που τοποθετούνται στην περιοχή αυτή σχετίζονται με αυτές τις κατηγορίες, πράγμα που σημαίνει παρουσία των κυνοδόντων και των τραπεζιτών και με έναν μεγάλο αριθμό από κοφτήρες και προγόμφιους. Τέτοιοι συνδυασμοί δοντιών χαρακτηρίζουν τα σαρκοβόρα ζώα, τα οποία χρησιμοποιούν τους κοφτήρες για να πνίξουν τη λεία τους, τους κυνόδοντες για να την αρπάξουν και να την κρατήσουν και τους τραπεζίτες και τους προγόμφιους για να αιχμαλωτίσουν την τροφή τους. Στην πάνω δεξιά γωνία, βρίσκονται οι κατηγορίες BP1, BP2, TP1, TP2, TP3, BI2, TI2, ενώ κοντά στο κέντρο υπάρχουν οι κατηγορίες TC1, BC1, TM2, BM2, BI1, TI3, BP3, BP4, TP4. Ωστόσο, η τελευταία ομάδα μπορεί να διαχωριστεί σε δυο επιμέρους ομάδες: οι κατηγορίες BP4 και TP4 είναι κοντά γιατί τα θηλαστικά με τρεις άνω προγομφίους έχουν συνήθως και τρεις κάτω προγομφίους. Ομοίως, οι κατηγορίες TC1, BC1, TM2, BM2 είναι κοντά, γιατί τα θηλαστικά που δεν έχουν άνω και κάτω κυνόδοντες, έχουν τρεις ή περισσότερους άνω και κάτω τραπεζίτες. Τέλος, κάτω και ελαφρώς δεξιά στο γράφημα 6.1, υπάρχουν οι κατηγορίες TI1, BI5.

Στη συνέχεια, αν μελετήσουμε το γράφημα 6.2, παρατηρούμε ότι τα αντικείμενα τοποθετούνται περιφερειακά σε αυτό, άμεση συνέπεια της αρχής του πρώτου κεντροειδούς. Επίσης, τα αντικείμενα σχηματίζουν τέσσερα ξεχωριστά νέφη, τα οποία τοποθετούνται στην άνω δεξιά, άνω αριστερά, κάτω αριστερά και στη μέση του γραφήματος.

6.8 Ανάλυση με τη βοήθεια του πίνακα Burt

Ο πίνακας Burt, έστω B, προκύπτει απευθείας από τον πίνακα-δείκτη Z ως εξής $B=Z^TZ$. Ένα τμήμα του πίνακα B για το παράδειγμα 6.7 παρουσιάζεται στον Πίνακα 6.5.

	TI 1	TI 2	TI 3	TI 4	BI 1	BI 2	BP 3	BP 4	BP 5	TM 1	TM 2	BM 1	BM 2
TI 1	10	0	0	0	1	0	0	9	0	0	10	0	10
TI 2	0	21	0	0	1	17	2	1	1	1	20	1	20
TI 3	0	0	9	0	0	3	4	4	1	1	8	1	8
TI 4	0	0	0	26	0	0	4	10	12	21	5	19	7
BI 1	1	1	0	0	2	0	0	1	0	1	1	1	1
BI 2	0	17	3	0	0	20	2	0	1	1	19	1	19
BP 3	0	2	4	4	0	2	10	0	0	4	6	4	6
BP 4	9	1	4	10	1	0	0	24	0	9	15	9	15
BP 5	0	1	1	12	0	1	0	0	14	10	4	8	6
TM 1	0	1	1	21	1	1	4	9	10	23	0	20	3
TM 2	10	20	8	5	1	19	6	15	4	0	43	1	42
BM 1	0	1	1	19	1	1	4	9	8	20	1	21	0
BM 2	10	20	8	7	1	19	6	15	6	3	42	0	45

Πίνακας 6.5

Τμήμα του πίνακα Burt.

6.8.1 Ιδιότητες της ανάλυσης με τη βοήθεια του πίνακα Burt

Οι υπολογισμοί της πολλαπλής ανάλυσης αντιστοιχιών είναι πάλι εφαρμογή της απλής ανάλυσης αντιστοιχιών στον πίνακα Β. Ωστόσο, παραθέτουμε μερικές ιδιότητες της ανάλυσης αυτής και της σχέσης με την ανάλυση αντιστοιχιών βάσει του πίνακα Z.

- Αφού ο πίνακας Β είναι συμμετρικός, η λύση για τις γραμμές και για τις στήλες είναι πανομοιότυπη.
- Η ανάλυση του πίνακα Β δίνει μια λύση για τις κατηγορίες απόκρισης (και είναι ότι ήταν προηγουμένως οι στήλες του Z).
- Οι τυποποιημένες συντεταγμένες των γραμμών (ή στηλών) του πίνακα Β ταυτίζονται με τις τυποποιημένες συντεταγμένες των στηλών του πίνακα Z.
- Οι κύριες αδράνειες του πίνακα Β είναι τα τετράγωνα των αντίστοιχων του πίνακα Z.
- Αφού ο πίνακας των τυποποιημένων καταλοίπων της ανάλυσης του πίνακα Β είναι θετικά ορισμένος και συμμετρικός, οι τιμές της ανάλυσης του πίνακα Β είναι επίσης ιδιοτιμές.

6.8.2 Αλγόριθμος υλοποίησης της πολλαπλής ανάλυσης αντιστοιχιών βάσει του πίνακα Burt-Εφαρμογή στο παράδειγμα 6.7

Στη συνέχεια, παραθέτουμε τον αλγόριθμο για την υλοποίηση της πολλαπλής ανάλυσης αντιστοιχιών σε μορφή βημάτων, με βάση τον πίνακα Β:

1. Διαιρούμε τον πίνακα Β με το σύνολο των στοιχείων του $n=\sum_{i,j} b_{ij}$ και έχουμε τον πίνακα $P=\{p_{ij}\}=c_{ij}/n$ και υπολογίζουμε τα αθροίσματα γραμμών r_i (*masses*), που ισούνται με τα αθροίσματα των στηλών.
2. Πραγματοποιούμε μια ανάλυση ιδιοτιμών-ιδιοδιανυσμάτων στον πίνακα Α των τυποποιημένων καταλοίπων (παρόμοια με την SVD) και προκύπτει ο πίνακας $S=\{s_{ij}\}=\frac{p_{ij}-r_i r_j}{\sqrt{r_i r_j}}$. Η ανάλυση δίνει τα ιδιοδιανύσματα $U=\{u_{ik}\}$ και τις ιδιοτιμές λ_k από τη λύση $S=V\Lambda V^T$. Οι ιδιοτιμές είναι ισοδύναμες με τις αντίστοιχες ιδιοτιμές του πίνακα 6.2 και είναι οι κύριες αδράνειες του πίνακα Z. Αν απαιτούνται οι κύριες αδράνειες του πίνακα Β, πρέπει να υψωθούν στο τετράγωνο. Ο Πίνακας 6.6 περιλαμβάνει τις κύριες αδράνειες που στηρίζονται στον πίνακα Β.

k	**1**	**2**	**3**	**4**	**…**	**17**	**18**	**19**
λ_k^2	0.534	0.144	0.076	0.048	…	0	0	0
Ποσοστό(%) της αδράνειας που εξηγείται	60.40	16.30	8.60	5.40	…	0	0	0

Πίνακας 6.6

Μερικές από τις 19 κύριες αδράνειες (λ_k) και το αντίστοιχο ποσοστό(%) της αδράνειας που εξηγούν βάση του πίνακα Β.

3. Η τυποποιημένη συντεταγμένη της i γραμμής (ή στήλης) για την k διάσταση δίνεται από τον τύπο: $\alpha_{ik}=u_{ik}/\sqrt{r_i}$.
4. Οι αντίστοιχες κύριες συντεταγμένες δίνονται από τον τύπο: $f_{ik}=a_{ik}*\lambda_k$.

Ο Πίνακας 6.7 περιλαμβάνει μερικά ιδιοδιανύσματα για τις δυο πρώτες διαστάσεις (u_{i1}, u_{i2}), τις ποσότητες r_i (*masses*) για τις κατηγορίες γραμμών και στηλών και τις τυποποιημένες και κύριες συντεταγμένες των γραμμών και των στηλών ($\alpha_{i1,}$ α_{i2}, f_{i1}, f_{i2}) για τις δύο πρώτες διαστάσεις.

	TI1	**TI2**	**TI3**	**TI4**	**BI1**	**BI2**	**BI3**	**…**
u_{i1}	-0.095	-0.240	0.010	0.269	-0.024	-0.262	0.116	…
u_{i2}	0.447	-0.233	-0.023	-0.054	0.059	-0.235	0.014	…
r_i	0.019	0.040	0.017	0.049	0.004	0.038	0.009	…
α_{i1}	-0.692	-1.203	0.073	1.212	-0.389	-1.348	1.192	…
α_{i2}	3.250	-1.170	-0.176	-0.244	0.966	-1.207	0.143	…
f_{i1}	-0.505	-0.879	0.054	0.886	-0.284	-0.985	0.871	...
f_{i2}	1.234	-0.444	-0.067	-0.093	0.367	-0.458	0.054	...

Πίνακας 6.7

Μερικά ιδιοδιανύσματα (u_{i1}, u_{i2}), ποσότητες r_i και τυποποιημένες και κύριες συντεταγμένες γραμμών και στηλών ($\alpha_{i1,}$ α_{i2}, f_{i1}, f_{i2}) για τις δύο πρώτες διαστάσεις.

6.9 Διόρθωση στις ιδιοτιμές της πολλαπλής ανάλυσης αντιστοιχιών

Η πολλαπλή ανάλυση αντιστοιχιών κωδικοποιεί τα δεδομένα δημιουργώντας δυαδικές στήλες για κάθε μεταβλητή με τον περιορισμό ότι μια και μόνο στήλη λαμβάνει την τιμή 1. Αυτή η κωδικοποίηση δημιουργεί τεχνητές επιπρόσθετες διαστάσεις επειδή μια κατηγορική μεταβλητή κωδικοποιείται με πολλές στήλες. Συνεπώς, το ποσοστό της αδράνειας που εξηγείται από την πρώτη διάσταση είναι σοβαρά υποεκτιμημένο. Στην πραγματικότητα, αποδεικνύεται ότι όλοι οι παράγοντες με ιδιοτιμή μικρότερη ή ίση του 1/Q απλά κωδικοποιούν αυτές τις επιπλέον διαστάσεις.

Δυο τύποι διορθώσεων έχουν προταθεί, ο ένας οφείλεται στον Benzecri (1979) και ο άλλος στον Greenacre(1993). Αυτές οι διορθώσεις λαμβάνουν υπόψη τους ότι οι ιδιοτιμές που είναι μικρότερες από 1/Q κωδικοποιούνται για τις επιπλέον διαστάσεις και η πολλαπλή ανάλυση αντιστοιχιών είναι ισοδύναμη με την ανάλυση ενός πίνακα Burt, του οποίου οι ιδιοτιμές είναι ίσες με τα τετράγωνα των ιδιοτιμών που προκύπτουν από την ανάλυση του πίνακα-δείκτη Z. Συγκεκριμένα, οι προσαρμοσμένες αδράνειες l_k^{adj} δίνονται από τον τύπο:

$$l_k^{adj} = \begin{cases} \left(\frac{Q}{Q-1}\right)^2 \left(l_k - \frac{1}{Q}\right)^2, & l_k > \frac{1}{Q} \\ 0, & l_k \le \frac{1}{Q} \end{cases} \quad (6.8)$$

Χρησιμοποιώντας τον τύπο (6.8), προκύπτει μια καλύτερη προσέγγιση της αδράνειας που αποκτάται από κάθε ιδιοτιμή.

Το ποσοστό της αδράνειας που εξηγείται, υπολογίζεται με το διαιρείται κάθε ιδιοτιμή με το άθροισμα των ιδιοτιμών και μπορεί και τώρα να χρησιμοποιηθεί αυτή η προσέγγιση. Ωστόσο, μια καλύτερη εκτίμηση του ποσοστού της αδράνειας προτάθηκε από τον Greenacre(1993), σχετίζεται με το μέσο όρο της αδράνειας των μη διαγώνιων στοιχείων των υποπινάκων του πίνακα Burt και δίνεται από τον τύπο:

$$\frac{Q}{Q-1}\left(\sum_k l_k^2 - \frac{J-Q}{Q^2}\right)^2. \quad (6.9)$$

Μερικές από τις προσαρμοσμένες αδράνειες (l_k^{adj}) και το αντίστοιχο ποσοστό(%) της αδράνειας που εξηγούν για το παράδειγμά μας περιέχονται στον Πίνακα 6.8.

k	**1**	**2**	**3**	**4**	**5**	**…**
l_k^{adj}	0.02041	0.01987	0.01912	0.01779	0.01683	…
Ποσοστό(%) της αδράνειας που εξηγείται	5.20	5.00	4.90	4.50	4.30	…

Πίνακας 6.8

Μερικές από τις προσαρμοσμένες αδράνειες (l_k^{adj}) και το αντίστοιχο ποσοστό(%) της αδράνειας που εξηγούν.

6.10 Από κοινού Ανάλυση Αντιστοιχιών

Επειδή η ερμηνεία στην πολλαπλή ανάλυση αντιστοιχιών είναι πιο «λεπτή» από ότι στην απλή ανάλυση αντιστοιχιών, διάφορες προσεγγίσεις έχουν προταθεί ώστε να επιτευχθεί η απλότητα στην ερμηνεία της ανάλυσης αντιστοιχιών με πίνακες-δείκτες. Μια προσέγγιση είναι να χρησιμοποιηθεί άλλη μετρική συνάρτηση από την X^2 και η πιο «δελεαστική» εναλλακτική είναι η απόσταση Hellinger. Μια άλλη προσέγγιση είναι η από κοινού ανάλυση αντιστοιχιών (*Joint Correspondence Analysis-JCA*), όπου προσαρμόζει μόνο τα μη διαγώνια στοιχεί του πίνακα Burt και μπορεί να αναλυθεί ως ένα παραγοντικό μοντέλο.

6.11 Σύντομη ιστορική αναφορά στην απλή ανάλυση αντιστοιχιών-Σχέση της ανάλυσης αντιστοιχιών με την πολλαπλή ανάλυση αντιστοιχιών και τα λογαριθμογραμμικά μοντέλα

Η ανάλυση αντιστοιχιών (*Correspondence Analysis-CA*) έχει μια αρκετά μακρά ιστορία ως μέθοδος ανάλυσης κατηγορικών δεδομένων και χρονολογείται από το 1935. Από τότε η CA έχει ανακαλυφθεί ξανά και ξανά πολλές φορές. Η ανάλυση αντιστοιχιών διακρίνεται σε απλή ανάλυση αντιστοιχιών, όπου είναι ανάλυση αντιστοιχιών για πίνακες συνάφειας και σε πολλαπλή ανάλυση αντιστοιχιών, όπου πρόκειται για ανάλυση αντιστοιχιών για πίνακες-δείκτες. Σήμερα, η ανάλυση αντιστοιχιών είναι γνωστή με διάφορα ονόματα όπως reciprocal averaging, dual (ή optimal scaling), canonical correlation analysis και ταυτόχρονη γραμμική παλινδρόμηση. Ο Greenacre(1984) ασχολήθηκε με τις διαφορές στην προσέγγιση μεταξύ της απλής ανάλυσης αντιστοιχιών και τις μεθόδους που προαναφέρθηκαν.

Ο όρος ανάλυση αντιστοιχιών προέρχεται από τη Γαλλία που είναι αρκετά δημοφιλής. Αυτό οφείλεται κυρίως στον Benzecri και τους συναδέλφους του. Ανέπτυξαν την ανάλυση αντιστοιχιών τη δεκαετία του 1960, και η δουλειά τους κορυφώνεται σε συγκεκριμένες αναφορές, όπως Benzecri και άλλοι(1973) και στη σειρά Pratique de l΄Analyse des Donnees (Benzecri κλπ., 1980, 1986). Επιπλέον, από το 1976 υπάρχει ένα ειδικό περιοδικό αφιερωμένο αποκλειστικά στην ανάλυση αντιστοιχιών και τις εφαρμογές της, το οποίο ονομάζεται *Les Cahiers de l' Analyse des Donnees*.

Στην ανάλυση αντιστοιχιών δίνεται ιδιαίτερη έμφαση στις γεωμετρικές αναπαραστάσεις και αυτός ίσως είναι ένας λόγος που αυτή η προσέγγιση έγινε τόσο δημοφιλής τουλάχιστον στην Γαλλία. Ένας λόγος για την καθυστέρηση της ανάπτυξης της μεθόδου αυτής στην υπόλοιπη Ευρώπη ίσως είναι το πρόβλημα της γλώσσας. Ωστόσο, πρόσφατες δημοσιεύσεις στα αγγλικά συνέβαλαν αρκετά στο να ξεπεραστεί αυτή η κατάσταση. Ένας άλλος λόγος ίσως είναι ότι η ανάλυση αντιστοιχιών παρουσιάζεται πολύ συχνά με αναφορές σε άλλες μεθόδους στατιστικής αντιμετώπισης των κατηγορικών δεδομένων που έχουν αποδείξει την χρησιμότητα και ευελιξία τους.

Μια μεγάλη διαφορά μεταξύ της ανάλυσης αντιστοιχιών και των άλλων τεχνικών για κατηγορικά δεδομένα βρίσκεται στην χρήση των μοντέλων. Στην λογαριθμογραμμική ανάλυση (*loglinear analysis-LLA*), για παράδειγμα, πρώτα θεωρείται η κατανομή υπό την οποία συλλέχθηκαν τα δεδομένα και κατόπιν θεωρείται ένα μοντέλο για τα δεδομένα και γίνονται εκτιμήσεις των παραμέτρων κάτω από την υπόθεση ότι το μοντέλο είναι αληθές, και τελικά αυτές οι εκτιμήσεις συγκρίνονται με τις παρατηρούμενες συχνότητες για να αξιολογηθεί το μοντέλο. Με αυτόν τον τρόπο είναι δυνατό να γίνουν αναφορές όσον αφορά στον πληθυσμό στον οποίο βασίζονται τα δεδομένα του δείγματος. Στην ανάλυση

αντιστοιχιών, δεν θεωρείται καμία κατανομή και κανένα μοντέλο, αλλά μια ανάλυση των δεδομένων γίνεται με στόχο να μελετηθεί η δομή των δεδομένων και η γεωμετρική τους ερμηνεία.

Τα λογαριθμογραμμικά μοντέλα μπορούν να παίξουν τον ακόλουθο ρόλο: πρώτον, μερικά λογαριθμογραμμικά μοντέλα μπορεί να έχουν κεντρική θέση στην ανάλυση και η ανάλυση αντιστοιχιών μπορεί να χρησιμοποιηθεί στην ανάλυση των λογαριθμικών μοντέλων συμπληρωματικά με την ανάλυση των καταλοίπων του μοντέλου. Η ανάλυση αντιστοιχιών μπορεί να είναι χρήσιμη μόνο όταν η απομάκρυνση από το μοντέλο που έχει υποτεθεί είναι σημαντική και αν υπάρχει νόημα να μελετηθεί η δομή των καταλοίπων. Δεύτερον, η χρήση των λογαριθμογραμμικών μοντέλων εξαφανίζει την αλληλεπίδραση στις λύσεις της ανάλυσης αντιστοιχιών. Αυτό σημαίνει ότι δεν υπάρχει ενδιαφέρον για τις παραμέτρους του μοντέλου, αλλά στα κατάλοιπα. Τρίτον, η ανάλυση αντιστοιχιών μπορεί να χρησιμοποιηθεί ως επεξηγηματικό εργαλείο ώστε να προσδιοριστεί το οικονομικότερο και το πιο εύκολο στην ερμηνεία λογαριθμικό μοντέλο. Τα αποτελέσματα του Goodman υποδεικνύουν με ποιο τρόπο σχετίζεται η κλασική ανάλυση αντιστοιχιών με τις προσεγγίσεις που βασίζονται σε μοντέλα, με το να παρουσιάσει τα μοντέλα συνάφειας RC και τα μοντέλα συσχέτισης. Σχέσεις μεταξύ της γενικευμένης ανάλυση αντιστοιχιών και των μοντέλων δεν είναι ακόμη γνωστή και είναι ένα θέμα που αξίζει να μελετηθεί περαιτέρω.

6.12 Εναλλακτικοί τρόποι αντιμετώπισης της ανάλυσης ομοιογένειας (ή πολλαπλής ανάλυσης αντιστοιχιών)

Στη συνέχεια, παρατίθενται τρεις τρόποι όπου θα μπορούσε να αντιμετωπιστεί εναλλακτικά η ανάλυση ομοιογένειας ή πολλαπλή ανάλυση αντιστοιχιών:

1. Η βασική προϋπόθεση είναι τα πολυμεταβλητά δεδομένα να είναι πιο προσβάσιμα με το να εκτεθούν οι βασικές τους ιδιότητες και γενικότερα η δομή τους με τη βοήθεια γραφημάτων. Για αυτό το λόγο, η ανάλυση ομοιογένειας αποδίδει σκορ στα αντικείμενα, ώστε να τα απεικονίσει με σημεία σε μικρότερης διάστασης χώρο με τέτοιο τρόπο ώστε τα αντικείμενα με παρόμοια προφίλ να είναι κοντά και αντικείμενα με διαφορετικά προφίλ να είναι σχετικά μακριά.
2. Μια δεύτερη πιθανότητα να παρουσιαστεί η ανάλυση ομοιογένειας είναι μέσω της γραμμικοποίησης των παλινδρομήσεων. Η λύση HOMALS μπορεί να στηριχθεί στο γεγονός ότι τα σκορ των αντικειμένων είναι ανάλογα των μέσων των κατηγοριών και αντίστροφα.

3. Μια τρίτη πιθανή ερμηνεία της λύσης HOMALS είναι με όρους ανάλυσης κυριών συνιστωσών του ποσοτικοποιημένου πίνακα δεδομένων. Μπορεί να αποδειχθεί ότι το άθροισμα των τετραγώνων των συσχετίσεων των βέλτιστων ποσοτικοποιημένων μεταβλητών $G_jY(\bullet, s)$ και του διανύσματος $X(\bullet, s)$, μεγιστοποιείται.

6.13 Το σύστημα κωδικοποίησης του πίνακα δείκτη-Πλεονεκτήματα και μειονεκτήματα

Το σύστημα κωδικοποίησης που χρησιμοποιήθηκε έως στην παρούσα εργασία είναι το επονομαζόμενο σύστημα crisp κωδικοποίησης του πίνακα δείκτη. Τα κύρια πλεονεκτήματα του είναι τα εξής:

1. είναι απλό και υπολογιστικά αποδοτικό.
2. επιτρέπει τους μη γραμμικούς μετασχηματισμούς των μεταβλητών.
3. είναι πολύ συμπαγές ακόμα και όταν κωδικοποιούνται δεδομένα με θόρυβο.
4. ο αριθμός των παραμέτρων (κατηγοριών) ανά μεταβλητή είναι γενικά μικρός.

Τα βασικά μειονεκτήματα του συστήματος αυτού είναι τα εξής:

i. σε πολλές περιπτώσεις ανάλυσης δεδομένων, ο καθορισμός των κατηγοριών γίνεται αυθαίρετα και

ii. όταν κωδικοποιούνται διαστηματικά δεδομένα υπάρχει αβεβαιότητα σχετικά με την τοποθέτηση των τιμών κοντά στα όρια της κάθε κατηγορίας.

Εναλλακτικά, έχει προταθεί στη βιβλιογραφία το σύστημα fuzzy κωδικοποίησης, το οποίο αποτελεί μια γενίκευση της αυστηρής λογικής κωδικοποίησης του πίνακα-δείκτη.

6.14 Συμπέρασμα-Μειονεκτήματα της πολλαπλής ανάλυσης αντιστοιχιών

Η γενίκευση της πολλαπλής ανάλυσης αντιστοιχιών για πολυμεταβλητά κατηγορικά δεδομένα μέσω της ανάλυσης του πίνακα δείκτη Z ή του πίνακα Burt B έχει αρκετά μειονεκτήματα και θεωρείται σε αρκετά σημεία ανεπαρκής. Το βασικό μειονέκτημα είναι η υποτίμηση της μεταβλητότητας που εξηγείται από τον κύριο άξονα. Αυτό έχει αντίκτυπο στην ερμηνεία της ανάλυσης αντιστοιχιών, όπου η αδράνεια διασπάται σε επιμέρους συνιστώσες. Το πρόβλημα αυτό, ωστόσο, ξεπερνιέται με την προσέγγιση των σταθμισμένων ελαχίστων τετραγώνων στην πολλαπλή ανάλυση αντιστοιχιών(βλ. Greenacre, 1988).

Αν και οι διαφορετικοί ορισμοί της ανάλυσης αντιστοιχιών ταυτίζονται τελικά μεταξύ τους σε πίνακες διπλής εισόδου, ωστόσο αυτό δεν συμβαίνει πάντοτε αν έχουμε παραπάνω από δύο κατηγορικές μεταβλητές. Η μέθοδος που είναι σήμερα γνωστή ως πολλαπλή ανάλυση αντιστοιχιών είναι η κατάλληλη γενίκευση του ορισμού του dual scaling της ανάλυσης

αντιστοιχιών, όπου βέβαια δεν υπάρχει εκεί η ιδέα της ολικής διακύμανσης (*total variation*). Και αυτό το πρόβλημα επιλύεται με την εφαρμογή των σταθμισμένων ελαχίστων τετραγώνων σε όλους τους πίνακες συνάφειας, όπου το άθροισμα των τετραγώνων θεωρείται ως ολική διακύμανση. Η απλή ανάλυση αντιστοιχιών αποτελεί ειδική περίπτωση της μεθόδου αυτής, πράγμα που δεν συμβαίνει όμως με την πολλαπλή ανάλυση αντιστοιχιών.

Η γεωμετρική ερμηνεία που συνήθως σχετίζεται με την ανάλυση αντιστοιχιών (Greenacre&Hastie, 1987) δεν γενικεύεται εύκολα με την ανάλυση του πίνακα Burt ή του τροποποιημένου πίνακα Burt. Επίσης, οι αποστάσεις μεταξύ των ποσοτικοποιήσεων των γραμμών ενός πίνακα δείκτη δεν έχουν πάντοτε ικανοποιητική διαισθητική ερμηνεία.

Μια ικανοποιητική γενίκευση της γεωμετρικής ερμηνείας της ανάλυσης αντιστοιχιών για πολυμεταβλητά δεδομένα θεωρείται ότι ακόμη λείπει και απαιτεί περαιτέρω μελέτη. Οι προσπάθειες που έχουν γίνει για αυτό το σκοπό έχουν αρκετά πλεονεκτήματα, αλλά καμία δεν συγκεντρώνει όλες αυτές τις ιδιότητες που οφείλει να έχει μια και μόνο μέθοδος. Οι διαφορές στις γενικεύσεις αυτές πρέπει να μελετηθούν ώστε να κατανοηθεί καλύτερα και πληρέστερα η ανάλυση αντιστοιχιών των πολυμεταβλητών κατηγορικών δεδομένων.

6.15 Πολλαπλή ανάλυση αντιστοιχιών με τη χρήση υπολογιστικών προγραμμάτων

a. με τη βοήθεια της R

Η υλοποίηση της πολλαπλής ανάλυσης αντιστοιχιών μπορεί να γίνει βήμα προς βήμα όπως στο παράδειγμα της παραγράφου 6.7. Ωστόσο, μια εναλλακτική λύση είναι η απευθείας χρήση εντολών που παρέχει το πακέτο mjca της R. Οι εντολές πραγματοποιούν απλή ανάλυση αντιστοιχιών, αλλά και πολλαπλή ανάλυση αντιστοιχιών και από κοινού ανάλυση αντιστοιχιών.

Η εντολή *ca( )* πραγματοποιεί απλή ανάλυση αντιστοιχιών βασιζόμενη στη Singular Value Decomposition.

Η εντολή *mjca( )* πραγματοποιεί και πολλαπλή και από κοινού ανάλυση αντιστοιχιών.

b. με τη βοήθεια του SPSS

Το υπολογιστικό πρόγραμμα SPSS παρέχει επίσης απευθείας χρήση εντολών για υλοποίηση απλής αλλά και πολλαπλής ανάλυσης αντιστοιχιών. Για πολλαπλή ανάλυση αντιστοιχιών, αρκεί να επιλεγούν τα εξής βήματα από το μενού:

Analyse®Data Reduction®Optimal Scaling.

Το SPSS παρέχει μέσω της διαδικασίας αυτής τις ιδιοτιμές από τη Singular Value Decomposition, τα σκορ, τις ποσοτικοποιήσεις των κατηγοριών, μέτρα διαχωριστικότητας, γραφήματα για τα σκορ και για τις ποσοτικοποιήσεις των κατηγοριών και γραφήματα για μέτρα διαχωριστικότητας.

ΠΑΡΑΡΤΗΜΑ Α

Παρατίθενται οι εντολές στην γλώσσα προγραμματισμού R, οι οποίες χρησιμοποιήθηκαν για την εφαρμογή του παραδείγματος της παραγράφου 2.2 για τη δημιουργία μωσαϊκών και γραφικών αναπαραστάσεων συνάφειας.

```
sex<-c("M","M","M","M","M","F","F","F","F","F")
as.factor(sex)
sex1<-rep(sex,6)
age<-c("10-20","25-35","40-50","55-65","70-90")
as.factor(age)
age1<-rep(age,12)
method<-c("poison","gas","hang","drown","gun","jump")
as.factor(method)
method1<-rep(method,each=10)
counts<-
c(1160,2823,2465,1531,938,921,1672,2224,2283,1548,335,883,625,201,45,40,113,91,45,29,1
524,2751,3936,3581,2948,212,575,1481,2014,1355,67,213,247,207,212,30,139,354,679,501,5
12,852,875,477,229,25,64,52,29,3,189,366,244,273,268,131,276,327,388,383)
suicide<-tapply(counts,list(method1, age1, sex1),c)
counts1<-array(counts1,dim=c(10,6))
suicide<-dimnames(counts1)<-list(age0=c("10-20","25-35","40-50","55-65","70-
90"),method0=c("poison","gas","hang","drown","gun","jump"))
mosaiplot(suicide)
```

ΠΑΡΑΡΤΗΜΑ Β

Παρατίθενται οι εντολές στην γλώσσα προγραμματισμού R, οι οποίες χρησιμοποιήθηκαν για την εφαρμογή του παραδείγματος της παραγράφου 6.7 και υπάρχουν στη δημοσίευση των Necadic&Greenacre, *Computation of Multiple Correspondence Analysis, with code in R.*

#Πίνακας 6.1 (Δημιουργία του πίνακα των δεδομένων(*mammals*) και του πίνακα δείκτη (*Z*))

```
TI<-c(4,4,4,4,3,3,3,3,2,2,3,1,3,3,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,4,4,4,4,4,4,4,4,4,4,
4,4,4,4,4,4,4,4,4,4,4,4,2,4,3,3,1,1,1,1,1,1,1,1,1)
BI<-
c(5,4,3,4,4,4,4,4,4,4,4,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,4,4,4,4,4,4,4,4,4,4,4,4,4,3,4,4,4
,4,3,3,1,3,2,4,5,5,5,5,5,5,5,5,5)
TC<-
c(2,2,2,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2
,2,2,2,2,2,2,2,2,2,1,2,2,1,1,1,1,1)
BC<-
c(2,2,1,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2
,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1)
TP<-
c(4,5,4,5,4,3,3,2,3,3,3,1,3,4,3,3,3,3,3,3,2,2,2,2,1,1,1,1,1,2,2,5,5,5,5,5,5,5,5,4,4,4,5,4,4,5,4,4,4,4
,4,5,5,4,4,5,4,4,4,4,4,4,4,4,4,4)
BP<-
c(4,5,4,5,4,4,3,3,3,3,4,1,3,3,2,2,2,2,2,2,2,2,2,2,1,1,1,1,1,2,2,5,5,5,5,5,5,5,5,4,4,4,5,4,4,4,4,3,3,3
,3,5,5,4,4,5,4,4,4,4,4,4,4,4,4,4)
TM<-
c(2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,1,2,1,1,1,1,1,1,1,1,1,1,1
,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2)
BM<-
c(2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1
,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2)
TI<-as.factor(TI)
BI<-as.factor(BI)
```

```
TC<-as.factor(TC)
BC<-as.factor(BC)
TP<-as.factor(TP)
BP<-as.factor(BP)
TM<-as.factor(TM)
BM<-as.factor(BM)
mammals<-data.frame(TI,BI,TC,BC,TP,BP,TM,BM)

lev.n<-unlist(lapply(mammals,nlevels))
n<-cumsum(lev.n)
J.t<-sum(lev.n)
Q.t<-dim(mammals)[2]
Z<-matrix(0, nrow=I, ncol=J.t)
newmammals<-lapply(mammals,as.numeric)
offset<-(c(0,n[-length(n)]))
for(i in 1:Q.t)
+ Z[1:I+(I*(offset[i]+newmammals[[i]] - 1))]<-1
 fn<-rep(names(mammals),unlist(lapply(mammals,nlevels)))
ln<-c(1,2,3,4,1,2,3,4,5,1,2,1,2,1,2,3,4,5,1,2,3,4,5,1,2,1,2)
dimnames(Z)[[2]]<-paste(fn,ln, sep=" ")
dimnames(Z)[[1]]<-as.character(1:I)
```

#Πίνακας 6.2(Οι κύριες αδράνειες του πίνακα Z και το αντίστοιχο ποσοστό(%) της αδράνειας που εξηγούν)

```
P<-Z/sum(Z)
cm<-apply(P,2, sum)
rm<-apply(P, 1, sum)
eP<-rm%*%t(cm)
S<-(P-eP)/sqrt(eP)
dec<-svd(S)
lam<-dec$d[1:(J.t-Q.t)]^2
expl<-100*(lam/sum(lam))
rbind(round(lam[c(1:19,(J.t-Q.t))], 3), round(expl[c(1:19,(J.t-Q.t))], 1))
```

 # Πίνακας 6.3 (Τυποποιημένες και κύριες συντεταγμένες των στηλών (b_{jk}, g_{jk} αντίστοιχα))

```
b.s1<-dec$v[,1]/sqrt(cm)
b.s2<-dec$v[,2]/sqrt(cm)
g.s1<-b.s1*sqrt(lam[1])
g.s2<-b.s2*sqrt(lam[2])
round(rbind(b.s1,b.s2,g.s1,g.s2),3)
```

#Πίνακας 6.4 (Τυποποιημένες και κύριες συντεταγμένες των γραμμών (f_{i1}, f_{i2} αντίστοιχα))

```
f.s1<-dec$u[,1]*sqrt(lam[1])/sqrt(rm)
f.s2<-dec$u[,2]*sqrt(lam[2])/sqrt(rm)
a.s1<-f.s1/sqrt(lam[1])
a.s2<-f.s2/sqrt(lam[2])
round(rbind(f.s1,f.s2,a.s1,a.s2), 3)
```

 # Πίνακας 6.5(Δημιουργία του πίνακα Burt από τον πίνακα Z)

```
B<-t(Z)%*%Z
```

#Πίνακας 6.6(Οι κύριες αδράνειες (λ_k) και το αντίστοιχο ποσοστό(%) της αδράνειας που εξηγούν, βάση του πίνακα Burt)

```
P.2<-B/sum(B)
eP.2<-cm.2%*%t(cm.2)
S.2<-(P.2-eP.2)/sqrt(eP.2)
dec.2<-eigen(S.2)
delt.2<dec.2$values[1:(J.t-Q.t)]
expl.2<-100*(delt.2/sum(delt.2))
lam.2<delt.2^2
expl.2b<-100*(lam.2/sum(lam.2))
rbind(round(lam.2,3),round(expl.2b,1))
```

#Πίνακας 6.7(Τα ιδιοδιανύσματα (u_{i1}, u_{i2}), ποσότητες r_i και τυποποιημένες και κύριες συντεταγμένες γραμμών και στηλών (α_{i1}, α_{i2}, f_{i1}, f_{i2}) για τις δύο πρώτες διαστάσεις του πίνακα Burt)

```
u.s1<-dec.2$vectors[,1]
u.s2<-dec.2$vectors[,2]
```

```
a2.s1<-u.s1/sqrt(cm.2)
a2.s2<-u.s2/sqrt(cm.2)
f2.s1<-a2.s1*sqrt(lam.2[1])
f2.s2<-a2.s2*sqrt(lam.2[2])
round(rbind(u.s1,u.s2, cm, a2.s1,a2.s2,f2.s1,f2.s2),3)
```

#Πίνακας 6.8(Μερικές από τις προσαρμοσμένες αδράνειες (l_k^{adj}) και το αντίστοιχο ποσοστό(%) της αδράνειας που εξηγούν, βάση του πίνακα Burt)

```
lam.adj<-(Q.t/(Q.t-1))^2*(delt.2[delt.2<=1/Q.t]-1/Q.t)^2
lam.adj
total.adj<-(Q.t/(Q.t-1))*((sum(delt.2^2)-((J.t-Q.t)/Q.t^2)))^2
rbind(round(lam.adj,5),100*round(lam.adj/total.adj,3))
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Κατέρη, Μ., (2005), Βιοστατιστική και Στατιστικές Μέθοδοι στην Επιδημιολογία, Σημειώσεις Πανεπιστημίου Πειραιώς.

Κατέρη, Μ., (2005), Ανάλυση Διακριτών Δεδομένων, Σημειώσεις Πανεπιστημίου Πειραιώς.

Ξένη

Agresti, A., (2002), *Categorical Data Analysis,* Wiley Interscience, New York.

Agresti, A., and, Kezouh, A., (1983), *Association Models for Multi-Dimensional Cross-Classifications of Ordinal Variables,* Communication in Statistics, Series A, **12**, 1261-1276.

Anderson, C.J., (1996), The analysis of Three-way Contingency Tables by Three-Mode Association Models, *Psychometrica*, **61**, 465-483.

Anderson, C.J., (2002), Analysis of Multivariate Frequency Data by Graphical Models and Generalizations of the Multidimensional Row-Column Association Model, *Psychometrica,* **7**, 446-467.

Anderson, C.J., and, Vermunt, J.K., (2000), Log-multiplicative Association Models as Latent Variable Models Nominal for/and Ordinal data, *Sociological Methodology*, **30**, 81-121.

Becker, M., (1989), Models for the Analysis of Association in Multivariate Contingency Tables, *Journal of the American Statistical Association,* **84**, 1014-1019.

Boik, R.J., (1996), An Efficient Algorithm for Joint Correspondence Analysis, *Psychometrika*, **61**, 255-269.

Choulakian, V., (1988), Exploratory Analysis of Contingency Tables by Loglinear Formulation and Generalizations of Correspondence Analysis, *Psychometrica*, **53**, 235-250.

Clogg, C, (1982), Some Models for the Analysis of Association in Multivariate Cross-Classifications Having Ordered Categories, *Journal of the American Statistical Association*, **77**, 803-815.

Clogg, C, and, Goodman, L., (1984), Latent Structure Analysis of a Set of Multidimensional Contingency Tables, *Journal of the American Statistical Association,* **79**, 762-771.

Colombi, R., Generilized Hierarchical Models for Multiple Contingency Tables.

Cuadras, C., and, Cuadras, D., (2006), A Parametric Approach to Correspondence Analysis, *Linear Algebra and its Applications*, **417**, 64-74.

Friendly, M., (1994), Mosaic Displays for Multi-Way Contingency Tables, *Journal of the American Statistical Association,* **89**, 190-200.

Friendly, M., Visualizing Categorical Data: Data, Stories, and Pictures, SUGI25, Paper.

Genest, C., and, Green, P., (1980), A Graphical Display of Association in Two-way Contingency Tables, *The Statistician,* **36**, 371-380.

Gilula, Z, and, Haberman, S., (1988), The Analysis of Multivariate Contingency Tables by Restricted Canonical and Restricted Association Models, *Journal of the American Statistical Association,* **83**, 760-771.

Gilula, Z., and, Ritov, Y., (1990), Inferential Ordinal Correspondence Analysis: Motivation, Derivation and Limitations, *International Statistical Review,* **58**, 99-108.

Goodman, L., (1970), The Multivariate Analysis of Qualitative Data: Interactions among Multiple Classifications, *Journal of the American Statistical Association,* **65**, 226-256.

Goodman, L., (1971), Partioning of Chi-square, Analysis of Marginal Contingency Tables, and Estimation of Expected Frequencies in Multidimensional Contingency Tables, *Journal of the American Statistical Association,* **66**, 339-344.

Goodman, L., (1979), Simple Models for the Analysis of Association in Cross-Classifications having Ordered Categories, *Journal of the American Statistical Association,* **74** , 537-552.

Goodman, L., (1985), The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories: Association Models, Correlation Models, and Assymetry Models for Contingency Tables with or without Missing entries, *The Annals of Statistics,* **13**, 10-69.

Goodman, L., (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., (1991), Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data, *Journal of the American Statistical Association,* **86**, 1085-1111.

Green, M., (1988), Generalizations of the Goodman Association Model for the Analysis of Multi-dimensional Contingency Tables.

Green, M., (1989), A Modelling Approach to Multiple Correspondence Analysis, COMPSTAT88, PhysicaVerlang, 317-322.

Greenacre, M., (1988), Correspondence Analysis of Multivariate Categorical Data by Weighted Least-Squares, *Biometrika,* **75**, 457-467.

Greenacre, M., and Hastie, T., (1987), The Geometric Interpretation of Correspondence Analysis, *Journal of the American Statistical Association,* **82**, 437-447.

Heijden, P., Falguerolles, A., and, de Leeuw, J., (1989), A Compined Approach to Contingency Table Analysis Using Correspondence Analysis and Log-Linear Analysis, *Applied Statistics*, **38**, 249-292.

Heiser, W., and, Meulman, J., (1983), Analysing Rectangular Tables by Joint and Constrained Multidimensional Scaling, *Journal of Econometrics,* **22**, 139-167.

Hoffman, H., (2002), Constructing and Reading Mosaiplots, *Computational Statistics&Data Analysis,* **43**, 565-580.

Hoffman, D., and, de Leeuw, J., (1992), Interpreting Multiple Correspondence Analysis as a Multidimensional Scaling Method, *Marketing Letters 3:3,* 259-272.

Hwang, H., and, Takane, Y., (2002), Generalized Constrained Multiple Correspondence Analysis, *Psychometrika,* **67**, 211-224.

Israels, A.Z., Bethlehem, J.G., van Driel, J., Jansen, M.E., Pannekoek, J., de Ree S.J.M, Sikkel, D., (1982), Multivariate Analysis Methods for Discrete Variables, *Metron*.

Jordan, M.I, (2004), Graphical Models, *Statistical Science,* **19**, 140-155.

Lauritzen, St., and, Sheehan, N., (2003), Graphical Models for Genetic Analyses, *Statistical Science*, **18**, 489-514.

Liu, I., and, Agresti, A., The Analysis of Ordered Categorical Data: An Overview and a Survey of recent Developments, (2005), *Sociedad de Estadistica e Investigation Operativa Test,* **14**, 1-73.

Meyer, D., Zeileis, A., and, Hornik, K., *Visualizing Independence Using Extended Association Plots*, Proceedings of the 3rd International Workshop in Distributed Statistical Computing(DSC 2003), March, 20-22, Vienna, Austria.

Michailidis, G., and, de Leeuw, J., (1998), The Gifi System of Descriptive Multivariate Analysis, *Statistical Science*, **13**, 307-336.

Michailidis, G., and, de Leeuw, J., (2005), Homogeneity Analysis Using Absolute Deviations, *Computational Statistics&Data Analysis,* **48**, 587-603.

Michailidis, G., and, de Leeuw, J., Constrained Homogeneity Analysis with Applications to Hierarchical Data, UCLA Statistics Series#207.

Necadic, O., and, Greenacre, M., Computation of Multiple Correspondence Analysis, with code in R, *BBVA Foundation*, Madrid.

Panadiotakos, D., and, Pitsavos, C., (2004), Interpretation of Epidemiological Data using Multiple Correspondence Analysis and Log-Linear Models, *Journal of Data Science,* **2**, 75-86.

Schriever, B.F., (1983), Scaling of Ordered Dependent Categorical with Correspondence Analysis, *International Statistical Review,* **51**, 225-238.

Siciliano, R, and, Mooijaart, A., (1997), *Three-factor Association Models for Three-way Contingency Tables,* Computational Statistics&Data Analysis, **24**, 337-356.

Simonoff, J.S., (2003), *Analysing Categorical Data,* Springer.

Tenenhaus, M., and, Young, F., (1985), An analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data, *Psychometrika,* **50**, 91-119.

Theus, M., (1997), Visualisation of Categorical Data, Advances in Statistical Software 6, Softstat'97 eds.Frank Faulbaum&Wolfgang Bardilla, Luciuc&Lucius, Stuttgart, 47-55.

Wong, R.S., (2001), Multidimensional Association Models, *Sociological Methods&Research,* **30**, 197-240.